Introduction to Modern Algebra

David Joyce
Clark University

Version 0.0.6, 3 Oct 2008 [1]

I dedicate this book to my friend and colleague Arthur Chou. Arthur encouraged me to write this book. I'm sorry that he did not live to see it finished.

Contents

Chapter 1

Introduction

1.1 Structures in Modern Algebra

Fields, rings, and groups. We'll be looking at several kinds of algebraic structures this semester, the major kinds being fields, rings, and groups, but also minor variants of these structures.

We'll start by examining the definitions and looking at some examples. For the time being, we won't prove anything; that will come later when we look at each structure in depth.

1.1.1 Operations on sets

We're familiar with many operations on the real numbers $\mathbf{R}$—addition, subtraction, multiplication, division, negation, reciprocation, powers, roots, etc.

Addition, subtraction, and multiplication are examples of binary operations, that is, functions $\mathbf{R} \times \mathbf{R} \to \mathbf{R}$ which take two real numbers as their arguments and return another real number. Division is almost a binary operation, but since division by 0 is not defined, it's only a partially defined operation. Most of our operations will be defined everywhere, but some won't be.

Negation is a unary operation, that is, a function $\mathbf{R} \to \mathbf{R}$ which takes one real number as an argument and returns a real number. Reciprocation is a partial unary operation since the reciprocal of zero is not defined.

The operations we'll consider are all binary or unary. Ternary operations can certainly be defined, but useful ones are rare.

Some of these operations satisfy familiar identities. For example, addition and multiplication are both commutative; they satisfy the identities

$$x + y = y + x \qquad \text{and} \qquad xy = yx.$$

A binary operation is said to be *commutative* when the order that the two arguments are applied doesn't matter, that is, interchanging them, or commuting one across the other, doesn't change the result. Subtraction and division, however, are not commutative.

Addition and multiplication are also associative binary operations

$$(x + y) + z = x + (y + z) \qquad \text{and} \qquad (xy)z = x(yz).$$

A binary operation is said to be *associative* when the parentheses can be associated with either the first pair or the second pair when the operation is applied to three arguments and the result is the same. Neither subtraction nor division are associative.

Both addition and multiplication also have identity elements

$$0 + x = x = x + 0 \qquad \text{and} \qquad 1x = x = x1.$$

An *identity element*, also called a *neutral element*, for a binary operation is an element in the set that doesn't change the value of other elements when combined with them under the operation. So, 0 is the identity element for addition, and 1 is the identity element for multiplication. Subtraction and division don't have identity elements.

Also, there are additive inverses and multiplicative inverses (for nonzero) elements. That is to say, given any x there is another element, namely $-x$, such that $x + (-x) = 0$, and given any nonzero x there is another element, namely $\frac{1}{x}$ such that $x(\frac{1}{x}) = 1$. Thus, a binary operation that has an identity element is said to have *inverses* if for each element there is an inverse element such that when combined by the operation they yield the identity element for the operation. Addition has inverses, and multiplication has inverses of nonzero elements. (Well, they do on the right, since $x - 0 = x$ and $x/1 = x$, but not on the left, since usually $0 - x \neq x$ and $1/x \neq x$.)

Finally, there is a particular relation between the operations of addition and multiplication, that of distributivity.

$$x(y + z) = xy + xz \qquad \text{and} \qquad (y + z)x = yx + zx.$$

Multiplication distributes over addition, that is, when multiplying a sum by x we can distribute the x over the terms of the sum.

Exercise 1.1. On properties of operations.

(a). Is the binary operation $x * y = \dfrac{x + y}{xy}$ for positive x and y a commutative operation? Is it associative?

(b). Is it true that $(w - x) - (y - z) = (w - y) - (x - z)$ is an identity for real numbers? Can you say why or why not?

(c). Although multiplication in $\mathbf{R}$ distributes over addition, addition doesn't distrubute over multiplication. Give an example where it doesn't.

Algebraic structures. We'll define fields, rings, and groups as three kinds of algebraic structures. An algebraic structure will have an underlying set, binary operations, unary operations, and constants, that have some of the properties mentioned above like commutativity, associativity, identity elements, inverse elements, and distributivity. Different kinds of structures will have different operations and properties.

The algebraic structures are abstractions of familiar ones like those on the real numbers $\mathbf{R}$, but for each kind of structure there will be more than one example, as we'll see.

1.1.2 Fields

Informally, a field is a set equipped with four operations—addition, subtraction, multiplication, and division that have the usual properties. (They don't have to have the other operations that **R** has, like powers, roots, logs, and the myriad functions like $\sin x$.)

Definition 1.1 (Field)**.** A *field* is a set equipped with two binary operations, one called *addition* and the other called *multiplication*, denoted in the usual manner, which are both commutative and associative, both have identity elements (the additive identity denoted 0 and the multiplicative identity denoted 1), addition has inverse elements (the inverse of x denoted $-x$), multiplication has inverses of nonzero elements (the inverse of x denoted $\frac{1}{x}$ or x^{-1}), multiplication distributes over addition, and $0 \neq 1$.

Of course, one example of a field in the field of real numbers **R**. What are some others?

Example 1.2 (The field of rational numbers, **Q**)**.** Another example is the field of rational numbers. A rational number is the quotient of two integers a/b where the denominator is not 0. The set of all rational numbers is denoted **Q**. We're familiar with the fact that the sum, difference, product, and quotient (when the denominator is not zero) of rational numbers is another rational number, so **Q** has all the operations it needs to be a field, and since it's part of the field of the real numbers **R**, its operations have the the properties necessary to be a field. We say that **Q** is a *subfield* of **R** and that **R** is an *extension* of **Q**. But **Q** is not all of **R** since there are irrational numbers like $\sqrt{2}$.

Example 1.3 (The field of complex numbers, **C**)**.** Yet another example is the field of complex numbers **C**. A complex number is a number of the form $a + bi$ where a and b are real numbers and $i^2 = -1$. The field of real numbers **R** is a subfield of **C**. We'll review complex numbers before we use them. See my *Dave's Short Course on Complex Numbers* at `http://www.clarku.edu/∼djoyce/complex/`

As we progress through this course, we'll look at many other fields. Some will only have a finite number of elements. (They won't be subfields of **Q**.) Some will have **Q** as a subfield but be subfields themselves of **R** or **C**. Some will be even larger.

Exercise 1.2. On fields. None of the following are fields. In each case, the operations of addition and multiplication are the usual ones.

(a). The integers **Z** is not a field. Why not?

(b). The positive real numbers $\{x \in \mathbf{R} \mid x > 0\}$ do not form a field. Why not?

(c). The set of real numbers between -10 and 10, that is,

$$(-10, 10) = \{x \in \mathbf{R} \mid -10 < x < 10\}$$

is not a field. Why not?

1.1.3 Rings

Rings will have the three operations of addition, subtraction, and multiplication, but don't need division. Most of our rings will have commutative multiplication, but some won't, so we won't require that multiplication be commutative in our definition. All the rings we'll look at have a multiplicative identity, 1, so we'll include that in the definition.

Definition 1.4 (Ring)**.** A *ring* is a set equipped with two binary operations, one called *addition* and the other called *multiplication*, denoted in the usual manner, which are both associative, addition is commutative, both have identity elements (the additive identity denoted 0 and the multiplicative identity denoted 1), addition has inverse elements (the inverse of x denoted $-x$), and multiplication distributes over addition. If, furthermore, multiplication is commutative, then the ring is called a *commutative ring.*

Of course, all fields are automatically rings, but what are some other rings?

Example 1.5 (The ring of integers, $\mathbf{Z}$)**.** The ring of integers $\mathbf{Z}$ includes all integers (whole numbers)—positive, negative, or 0. Addition, subtraction, and multiplication satisfy the requirements for a ring, indeed, a commutative ring. But there are no multiplicative inverses for any elements except 1 and -1. For instance, $1/2$ is not an integer. We'll find that although the ring of integers looks like it has less structure than a field, this very lack of structure allows us to discover more about integers. We'll be able to talk about prime numbers, for example.

Example 1.6 (Polynomial rings)**.** A whole family of examples are the rings of polynomials. Let R be any commutative ring (perhaps a field), and let $R[x]$ include all polynomials with coefficients in R. We know how to add, subtract, and multiply polynomials, and these operations have the properties required to make $R[x]$ a commutative ring. We have, for instance, the ring of polynomials with real coefficients $\mathbf{R}[x]$, the ring with integral coefficients $\mathbf{Z}[x]$, etc.

Example 1.7 (Matrix rings)**.** How about an example ring that's not commutative? The ring of $n \times n$ matrices with entries in a commutative ring R gives such an example, this ring being denoted $M_n(R)$. It won't be commutative when $n \geq 2$. An example of a matrix ring is the ring of 2×2 matrices with real entries, $M_2(\mathbf{R})$. Addition and subtraction is computed coordinatewise. The additive identity, 0, the matrix ring is the matrix with all 0 entries. Matrix multiplication is not coordinatewise, but it is associative, and multiplication does distribute over addition. The multiplicative identity for this matrix ring is what's usually called the identity matrix, denoted I. It has 1's down the main diagonal and 0's elsewhere.

Sylvester, in 1850, called rectangular arrangements of numbers *matrices*, and Cayley wrote much about them in his papers of 1855–1858.

Example 1.8 (Integers modulo n)**.** An important family of rings is the ring of integers modulo n. We'll study this in more detail later, but here's an incomplete overview. If we think of two integers a and b as being the same if $a \equiv b \pmod{n}$, that is if n divides $b-a$, then there are only n integers modulo n. One way of doing that is to represent integers modulo n by the integers from 0 through $n-1$. Thus, we'll say, for instance, that 5 plus 3 equals 1 modulo 7, by which we mean $5+3 \equiv 1 \pmod{7}$. Thus, we can turn congruence modulo n, which is an equivalence relation on $\mathbf{Z}$ into equality on an n-element set. That n-element set is denoted $\mathbf{Z}/n\mathbf{Z}$, read $\mathbf{Z}$ mod $n\mathbf{Z}$, or more simply as $\mathbf{Z}_n$, read $\mathbf{Z}$ sub n. So, we can take the elements of $\mathbf{Z}_n$ to be the integers from 0 through $n-1$, where we understand that addition, subtraction, and multiplication are done modulo n. And it turns out that this is a ring, as we'll see when we study $\mathbf{Z}_n$ in detail.

Incidentally, when n is a prime number p, then $\mathbf{Z}_p$ is not just a ring, but a field, as we'll see later.

Exercise 1.3. On rings. None of the following are rings. In each case, the operations of addition and multiplication are the usual ones.

(a). The set of nonzero integers, $\{x \in \mathbf{Z} \,|\, x \neq 0\}$ is not a ring. Why not?

(b). The set of even integers $\{2x \,|\, x \in \mathbf{Z}\}$ is not a ring. Why not?

(c). The set of odd degree polynomials with real coefficients

$$\{f(x) \in \mathbf{R}[x] \,|\, \text{the degree of } f \text{ is odd}\}$$

is not a ring. Why not? (How about the set of even degree polynomials?)

Exercise 1.4. On noncommutative rings. Are the following rings? (The operations are the usual matrix operations.) Explain in a sentence or two, but a proof is not necessary.

(a). The set of all matrices with real coefficients (any size).

(b). The set of all 2×2 matrices with real entries of the form

$$\begin{bmatrix} a & b \\ 0 & d \end{bmatrix}.$$

(c). The set of all 2×2 matrices with real entries of the form

$$\begin{bmatrix} a & b \\ -b & a \end{bmatrix}.$$

1.1.4 Groups

Unlike fields and rings which have two primary binary operations, groups only have one binary operation.

Definition 1.9 (Group)**.** A *group* is a set equipped with a binary operation that is associative, has an identity element, and has inverse elements. If, furthermore, multiplication is commutative, then the group is called a *commutative group* or an *Abelian group.* Abelian groups can be denoted either additively or multiplicatively, but nonabelian groups are usually denoted multiplicatively. We'll use the term *order of the group* to indicate how many elements a group G has and denote this order by $|G|$.

Example 1.10 (The underlying additive group of a ring)**.** Of course, if you have a field or ring, and just consider addition (and forget about multiplication) you've got an Abelian group. Sometimes this is called the *underlying additive group* of the field or ring. We'll use the same notation for the underlying additive group as we do for the ring. Thus, $\mathbf{Z}$ could mean either the ring of integers or the Abelian group of integers under addition, depending on the context.

Example 1.11 (The group of units in a ring)**.** In order to use the multiplication for a group operation, we'll have to only include the units, also called invertible elements. A *unit* or *invertible element* of a ring R is an element $x \in R$ such that there exists another element $y \in R$ so that $xy = yx = 1$. The subset of units is denoted

$$R^* = \{x \in R \,|\, \exists y \in R, xy = 1\}.$$

You can easily show that the units form a group under multiplication, called the *multiplicative group of units* of R. When R is a field, then R^* is all of R except 0, but for rings there will be other elements than 0 that aren't invertible. The group R^* will be Abelian when the ring R is commutative, but usually it will be nonabelian when R is not commutative.

Example 1.12 (A general linear group, $GL_2(\mathbf{R})$)**.** As a particular example of a multiplicative group of units, take the invertible elements of the matrix ring $M_2(\mathbf{R})$. The invertible 2×2 matrices are those matrices

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

whose determinants $ad - bc$ are nonzero. The group of invertible $n \times n$ matrices, $M_n(R)^*$, is the *general linear group* with coefficients in the ring R, denoted $GL_n(R)$. Note that $GL_n(R)$ is a nonabelian group for $n \geq 2$. The real general linear group $GL_2(\mathbf{R})$ can be interpreted as the group of linear transformations of the plane $\mathbf{R}^2$ that leave the origin fixed.

Exercise 1.5. Find two matrices in $GL_2(\mathbf{Z})$ that don't commute thereby proving $GL_2(\mathbf{Z})$ is a nonabelian group.

There are many examples of finite nonabelian groups found in geometry. We'll look at the group of symmetries of an equilateral triangle.

Example 1.13 (The dihedral group D_3)**.** Consider an equilateral triangle. Place a coordinate system on the plane of the triangle so that its center is at $(0, 0)$, one vertex, A, at $(1, 0)$, and the other two, B and C, at $(-\frac{1}{2}, \pm\frac{1}{2}\sqrt{3})$. This triangle has six symmetries. A symmetry is a transformation of the plane that preserves distance (that is, an *isometry*) that maps the triangle back to itself. Three of these symmetries are rotations by 0°, 120°, and 240°

$$1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \rho = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2}\sqrt{3} \\ \frac{1}{2}\sqrt{3} & -\frac{1}{2} \end{bmatrix} \qquad \rho^2 = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2}\sqrt{3} \\ -\frac{1}{2}\sqrt{3} & -\frac{1}{2} \end{bmatrix}$$

The identity transformation, 1, fixes A, B, and C; the rotation ρ by 120° maps A to B, B to C, and C to A; and the rotation ρ^2 by 240° maps A to C, B to A, and C to B. There are also three reflections.

$$\varphi = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \qquad \rho\varphi = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2}\sqrt{3} \\ -\frac{1}{2}\sqrt{3} & \frac{1}{2} \end{bmatrix} \qquad \rho^2\varphi = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2}\sqrt{3} \\ -\frac{1}{2}\sqrt{3} & \frac{1}{2} \end{bmatrix}$$

The refection φ fixes A, and interchanges B and C; the reflection $\rho\varphi$ fixes C and interchanges A and B; and the reflection $\rho^2\varphi$ fixes B and interchanges A and C This is a particular nonabelian group that has 6 elements. It is a subgroup of $GL_2(\mathbf{R})$ mentioned above.

Example 1.14 (Euler's circle group)**.** The unit circle, $S^1 = \{x + yi \in \mathbf{C} \,|\, x^2 + y^2 = 1\}$, is a group under multiplication. The product of two complex numbers on this unit circle is another. You can directly verify that or you can show it by trigonometry. If $x + yi$ is on the unit circle, then we can identify x with $\cos\theta$ and y with $\sin\theta$ where θ is, as usual, the angle between the positive x-axis and the ray from 0 to $x + yi$. Then the product of two complex numbers on the unit circle corresponds to adding their angles together. The addition formulas for cosines and sines give this correspondence.

Exercise 1.6. Compute the product of $\cos\theta + i\sin\theta$ times $\cos\varphi + i\sin\varphi$. If $x + iy = (\cos\theta + i\sin\theta)(\cos\varphi + i\sin\varphi)$, then what is x, the real part of the product, in terms of θ? What is y, the imaginary part?

Comment 1.15. Although the sphere

$$S^2 = \{(x, y, z) \in \mathbf{R}^3 \,|\, x^2 + y^2 + z^2 = 1\}$$

has no group structure, the 3-sphere in 4-space does. The 3-sphere is

$$S^3 = \{(x, y, z, w) \in \mathbf{R}^4 \,|\, x^2 + y^2 + z^2 + w^2 = 1\}.$$

We don't have time or space to discuss that group structure here. (The 2-sphere S^2, in fact, spheres in all dimensions, does have quandle structures, whatever a quandle might be.)

1.1.5 Other algebraic structures besides fields, rings, and groups

There are an unlimited number of other algebraic structures. Some are similar to those listed above.

For instance, there are division rings (also called skew fields) that have all the properties of fields except multiplication doesn't have to be commutative. The primary example is the quaternions $\mathbf{H}$. I hope we have half an hour sometime during the semester to introduce the quaternions.

There are a number of structures that are just commutative rings that have nice properties, and we'll look at some of them including integral domains, unique factorization domains, principal ideal domains, and Euclidean domains.

Sometimes rings that don't have a multiplicative identity are studied, but for us, we'll always have 1.

You've already studied vector spaces over the real numbers. Most of the things that you've studied about vector spaces over $\mathbf{R}$ also hold for vector spaces over other fields.

The analogous structure for vector spaces when a field is replaced by a ring is called a *module* over the ring. We won't study modules over a ring, but when we look at ideals in a ring, they are, in fact, examples of modules over the ring.

If I have time, I'll tell you about quandles.

1.2 Isomorphisms, homomorphisms, etc.

Frequently we'll be looking at two algebraic structures A and B of the same kind, for instance, two groups or two rings or two fields, and we'll want to compare them. For instance, we might think they're really the same thing, but they have different names for their elements. That leads to the concept of isomorphism $f : A \cong B$, and we'll talk about that first. Other times we'll know they're not the same thing, but there is a relation between them, and that will lead to the next concept, homomorphism, $f : A \to B$. We'll then look as some special homomorphisms such as monomorphisms. When we have a homomorphism $f : A \to A$, we'll call it an endomorphism, and when an isomorphism $f : A \cong A$, we'll call it an automorphism. We'll take each of these variants in turn.

Injections, surjections, and bijections. These are words that describe certain functions $f : A \to B$ from one set to another. An *injection*, also called a *one-to-one function* is a function that maps distinct elements to distinct elements, that is, if $x \neq y$, then $f(x) \neq f(y)$. Equivalently, if $f(x) = f(y)$ then $x = y$. If A is a subset of B, then there is a natural injection $\iota : A \to B$, called the *inclusion function*, defined by $\iota(x) = x$.

A *surjection*, also called an *onto function* is one that includes all of B in its image, that is, if $y \in B$, then there is an $x \in A$ such that $f(x) = y$.

A *bijection*, also called a *one-to-one correspondence*, is a function that is simultaneously injective and bijective. Another way to describe a bijection is to say that there is an inverse function $g : B \to A$ so that the composition $g \circ f : A \to A$ is the identity function on A while $f \circ g : B \to B$ is the identity function on B. The usual notation for the function inverse to f is f^{-1}. In this situation f and g are inverse to each other, that is, if g is f^{-1}, then f is g^{-1}. Thus, $(f^{-1})^{-1} = f$.

We'll use the following theorem about finite sets when we consider homomorphisms between finite algebraic structures.

Theorem 1.16. Suppose that $f : A \to B$ is a function between two finite sets of the same cardinality. Then the following three conditions are equivalent: (1) f is a bijection, (2) f is an injection, and (3) f is a surjection.

1.2.1 Isomorphisms

We'll say two algebraic structures A and B are isomorphic if they have exactly the same structure, but their elements may be different. For instance, let A be the ring $\mathbf{R}[x]$ of polynomials in the variable x with real coefficients while B is the ring $\mathbf{R}[y]$ of polynomials in y. They're both just polynomials in one variable, it's just that the choice of variable is different in the two rings. We need to make this concept more precise.

Definition 1.17 (Ring isomorphism)**.** Two rings A and B are *isomorphic* if there is a bijection $f : A \to B$ which preserves addition and multiplication, that is, for all x and y in A,

$$f(x + y) = f(x) + f(y) \text{ and } f(xy) = f(x)f(y).$$

The correspondence f is called a *ring isomorphism*.

After we introduce homomorphisms, we'll have another way to describe isomorphisms.

You can prove various properties of ring isomorphism from this definition.

Exercise 1.7. Since the structure of rings is defined in terms of addition and multiplication, if f preserves them, it will preserve structure defined in terms of them. Verify that f preserves 0, 1, negation, and subtraction.

Exercise 1.8. Prove that if f is a ring isomorphism, then so is the inverse correspondence $f^{-1} : B \to A$.

Exercise 1.9. Prove that if $f : A \to B$ and $g : B \to C$ are both ring isomorphisms, then so is their composition $(g \circ f) : A \to C$.

Since a field is a special kind of ring, and its structure is defined in terms of addition and multiplication, we don't need a special definition for a field isomorphism. A field isomorphism is just a ring isomorphism between fields.

Exercise 1.10. Prove that if a ring is isomorphic to a field, then that ring is a field.

We do need a different definition for a group isomorphism since a group is defined in terms of just one binary operation instead of two.

Definition 1.18 (Group isomorphism)**.** Two groups A and B are isomorphic if there is a bijection $f : A \to B$ which preserves the binary operation. If both are written additively, that means for all x and y in A,

$$f(x+y) = f(x) + f(y);$$

if multiplicative notation is used in both, then $f(xy) = f(x)f(y)$; if additive in A but multiplicative in B, then $f(x+y) = f(x)f(y)$; and if multiplicative in A and additive in B, then $f(xy) = f(x) + f(y)$. The correspondence f is called a *group isomorphism*.

Usually A and B will use the same notation, both additive or both multiplicative, but not always.

Exercise 1.11. Suppose that both A and B are written multiplicatively and that $f : A \to B$ is a group isomorphism. Prove that $f(1) = 1$ and $f(x^{-1}) = f(x)^{-1}$ for all $x \in A$.

Example 1.19. Let $A = \mathbf{Z}$ be the group of integers under addition. Let B be the integral powers of 2, so $B = \{\ldots, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, \ldots\}$ with multiplication as the operation in B. Then an isomorphism $f : A \to B$ is defined by $f(n) = 2^n$. There's actually another isomorphism $g : A \to B$, too, defined by $g(n) = 2^{-n}$.

1.2.2 Homomorphisms

Whereas isomorphisms are bijections that preserve the algebraic structure, homomorphisms are simply functions that preserve the algebraic structure. Since the word homomorphism is so long, alternate words are often used like morphism and map, especially in spoken mathematics. I'll probably just say map, but continue to use homomorphism in these notes.

Definition 1.20 (Homomorphism)**.** A *ring homomorphism* $f : A \to B$ between rings is a function that preserves addition, multiplication, and 1.

A *group homomorphism* $f : A \to B$ between groups preserves the binary operation (addition or multiplication depending on the notation used for the group).

Comment 1.21. It's a peculiarity of rings that preserving addition and multiplication doesn't imply that 1 is also preserved, so that condition has to be required as well. We'll see plenty of examples of homomorphisms in the course, and I'll give more examples in the next section on monomorphisms. Of course, isomorphisms are special cases of homomorphisms.

Example 1.22 (A ring homomorphism)**.** Let $\mathbf{Z}[x]$ be the ring of polynomials with integral coefficients. Evaluating a polynomial $f(x)$ at a particular number, like 3, to give $f(3)$, is a ring homomorphism $\varphi : \mathbf{Z}[x] \to \mathbf{Z}$. It preserves addition since $\varphi(f(x)+g(x)) = f(3)+g(3) = \varphi(f(x)) + \varphi(g(x))$, and you can check that it preserves multiplication and 1.

Example 1.23 (A group homomorphism)**.** Let A be the integers under addition, and let $B = \{1, -1\}$ with multiplication as the binary operation. Then $f : A \to B$ defined by $f(n) = (-1)^n$ is a group homomorphism.

You can prove several properties of homomorphisms from the definition, but for the time being I'll just mention two because they'll lead to the concept of *category* which we'll talk about some other time.

1. The composition of two homomorphisms (of the same kind) is another homomorphism.

2. The identity function $1_A : A \to A$, which maps every element to itself, is a homomorphism, indeed, it's an isomorphism.

When we have a homomorphism $f : A \to B$, we'll call A the *domain* of f and we'll call B the *codomain* of f.

A more natural way to characterize isomorphism is in terms of homomorphisms. Two rings A and B are isomorphic if and only if there if there are two ring homomorphisms $f : A \to B$ and $g : B \to A$ such that $g \circ f$ is the identity on A and $f \circ g$ is the identity on B.

1.2.3 Monomorphisms and epimorphisms

Two common kinds of homomorphisms are monomorphisms and epimorphisms, often called monos and epis for short. When a homomorphism $f : A \to B$ is an injective function, it's called a *monomorphism*; and when it a surjective function, it's an *epimorphism* (but, in the category of rings, we'll see there are more epimorphisms than just the surjective ring homomorphisms). You might wonder why we need these words when we've got more than enough words already to describe injective (one-to-one) and surjective (onto) as well as others not mentioned here. The main reason is that they're special kinds of injections or surjections—they preserve the algebraic structure. Another is that, although for group homomorphisms monos and epis have these particular correspondences to injective and surjective, there are other categories in which they don't.

Note that every isomorphism is simultaneously a monomorphism and and epimorphism. The converse holds for groups, but not for rings.

Example 1.24 (Inclusion)**.** Inclusions are monomorphisms. When one ring (or group) A is a subring (or subgroup) of another B, then the inclusion function $\iota : A \to B$, which maps an element to itself, is a monomorphism. That's an important example of a monomorphism, but there are others.

Example 1.25. For example, let A and B both be the additive group of integers $\mathbf{Z}$, and let $f(n) = 2n$. This f is a monomorphism, but it's not an inclusion (which in this case would be the identity map since A and B are the same).

Comment 1.26. Note that if $f : A \to B$ is a ring homomorphism where A is a field and $0 \neq 1$ in B, then f is always an injection, and so it's a monomorphism. You can prove this statement in two stages. First, show that if $f(x) = 0$ then $x = 0$. Second, show that if $f(x) = f(y)$, then $x = y$.

Thus, every field homomorphism is a monomorphism.

Example 1.27 (A group epimorphism)**.** We'll see plenty of epimorphisms when we talk the integers modulo n, but for the time being, here's one example of a group epimorphism. Let A be the additive group of integers $\mathbf{Z}$ and let B be the two element group $\{1, -1\}$ under multiplication. Then $f : A \to B$ defined by $f(n) = (-1)^n$ is a group epimorphism.

1.2.4 Endomorphisms and automorphisms

An endomorphism is just a homomorphism $f : A \to A$ where the domain and codomain are the same, and an automorphism is just an isomorphism $f : A \to A$. These are important because we always have the identity automorphism $1_A : A \to A$ to compare f to, so we have more information when the domain and codomain are the same.

Example 1.28 (A field automorphism)**.** Let $\mathbf{C}$ be the complex field. Let $\phi : \mathbf{C} \to \mathbf{C}$ be complex conjugation, usually denoted by putting a bar above the complex number

$$\varphi(x + yi) = \overline{x + yi} = x - yi.$$

This is clearly a bijection since it is its own inverse, $\overline{\overline{x + yi}} = x+yi$. Also it preserves addition, multiplication, and 1, so it's a ring isomorphism.

$$\begin{array}{rcl} \overline{(x_1 + y_1 i) + (x_2 + y_2 i)} & = & \overline{x_1 + y_1 i} + \overline{x_2 + y_2 i} \\ \overline{(x_1 + y_1 i)(x_2 + y_2 i)} & = & \overline{x_1 + y_1 i}\; \overline{x_2 + y_2 i} \\ \overline{1} & = & 1 \end{array}$$

In fact, it's a field automorphism of $\mathbf{C}$.

The existence of this automorphism says that we can't distinguish between i and $-i$ in the sense that any true statement about the complex numbers remains true when all occurrences of i are replaced by $-i$.

Example 1.29 (A group endomorphisms and an automorphism)**.** There are many group endomorphisms $f : \mathbf{Z} \to \mathbf{Z}$ from the additive group of integers to itself. Fix any integer n and let $f(x) = nx$. This is a group homomorphism since $f(x + y) = n(x + y) = nx + ny = f(x) + f(y)$. For $n \neq 0$ it is also a monomorphism. For $n = -1$, it's a bijection, so it's an automorphism. That says if we only consider addition, we can't distinguish between positive and negative numbers.

But negation is not a ring automorphism on the ring of integers because $-(xy)$ does not equal $(-x)(-y)$. Thus, with the use of multiplication, we can distinguish between positive and negative numbers.

1.3 A little number theory

> In science nothing capable of proof ought to be accepted without proof. Though this demand seems so reasonable, yet I cannot regard it as having been met even in the most recent methods of laying the foundations for the simplest science; viz., that part of logic which deals with the theory of numbers.
>
> Dedekind, 1888

This course is not meant to be a course in number theory, but we will need a little bit of it. We'll quickly review mathematical induction on the natural numbers **N**, divisibility, prime numbers, greatest common divisors, and the Euclidean algorithm.

1.3.1 Mathematical induction on the natural numbers N

Richard Dedekind (1831–1916) published in 1888 a paper entitled *Was sind und was sollen die Zahlen?* variously translated as *What are numbers and what should they be?* or *The Nature of Meaning of Numbers.* In that work he developed basic set theory and characterized the natural numbers as a simply infinite set. You can find Joyce's notes on this paper on the web at

`http://www.aleph0.edu/∼djoyce/numbers/dedekind.html`

Definition 1.30. (Dedekind) A set **N** is said to be *simply infinite* when there exists a one-to-one function $\mathbf{N} \xrightarrow{'} \mathbf{N}$ called the *successor function*, such that there is an element, called the *initial element* and denoted 1, that is not the successor of any element, and if a subset S of **N** contains 1 and is closed under the successor function, then $S = \mathbf{N}$.

Such a simply infinite set **N** may be called the *natural numbers*. It is characterized by an element 1 and a transformation $\mathbf{N} \xrightarrow{'} \mathbf{N}$ satisfying the following conditions:

1. $\forall n, m, n \neq m$ implies $n' \neq m'$.
2. $\forall n, 1 \neq n'$.
3. If $S \subseteq \mathbf{N}$, $1 \in S$, and ($\forall n, n \in S$ implies $n' \in S$), then $S = \mathbf{N}$.

The Dedekind axioms, also called the Peano axioms, are this last characterization involving 1, the successor function, and the three conditions.

The last axiom is called mathematical induction. If you want to show a subset S of **N** is all of N, first show that $1 \in S$. Then show for each natural number n that $n \in S$ implies $n + 1$ in S. Finally conclude that $S = \mathbf{N}$.

A principle that is logically equivalent to mathematical induction is the well-ordering principle, also called the minimization principle. It says that each nonempty subset of **N** has a least element. To use it to prove a subset S of **N** is all of **N**, assume that it isn't, take the least element n in $\mathbf{N} - S$, and derive a contradiction, usually by showing there's a smaller element than n not in S.

1.3.2 Divisibility

We'll restrict our discussion now to **N**, the natural numbers, that is, the set of positive integers.

Recall that an integer m *divides* an integer n, written $m\big|n$, if there exists an integer k such that $mk = n$. A few basic properties of divisibility follow directly from this definition. Euclid uses some of these in Book VII of his *Elements*. You can find Joyce's translation of Euclid's *Elements* on the web at

`http://www.aleph0.edu/∼djoyce/java/elements/elements.html`

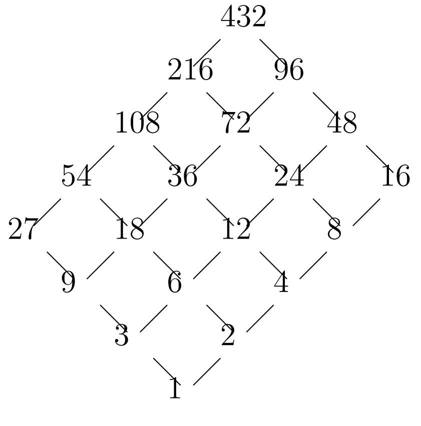
1. 1 divides every number. $1\big|n$.
2. Each number divides itself. $n\big|n$.
3. If one number m divides another number n, then m divides any multiple of n, that is, $m\big|n$ implies $m\big|kn$.
4. Divisibility is a transitive relation, that is, $m\big|n$ and $n\big|k$ imply $m\big|k$.
5. If one number divides two other numbers, then it divides both their sum and difference. $m\big|n$ and $m\big|k$ imply $m\big|(n+k)$ and $m\big|(n-k)$.
6. Cancellation law. One number divides another if and only if any multiple of that one number divides the same multiple of the other number. $m\big|n \iff kn\big|kn$.

The divisors of a number can be displayed graphically in what is called a Hasse diagram of the lattice of divisors. We'll look at a few of those in class.

Example 1.31. As an example, consider the number 432. Its prime factorization is $2^4 3^3$, so its divisors are of the form $2^m 3^n$ where $0 \le m \le 4$ and $0 \le n \le 3$. There are $5 \cdot 4 = 20$ of these divisors. They are

1	3	9	27
2	6	18	54
4	12	36	108
8	24	72	216
16	48	144	432

We can display these numbers and emphasize which ones divide which other ones if we put the large numbers at the top of the diagram, and connect the smaller divisors to the larger ones with lines.



Since divisibility is transitive, we don't have to include all possible connections. So long as there is a path of connections from a lower number to an upper one, then we can conclude the lower divides the upper. The resulting diagram is called a *Hasse diagram.*

Exercise 1.12. Draw Hasse diagrams for the divisors of 30, 32, and 60.

1.3.3 Prime numbers

An natural number greater than 1 is a *prime number* if its only divisors are 1 and itself, but if it has more divisors, it's called a *composite number.* We know intuitively that there are infinitely many primes, and that every number is a product of primes. Now let's prove those statements. We'll start by proving something that will help us prove these two statements. If a theorem is not particularly interesting, but is useful in proving an interesting statement, then it's often called a lemma. This one is found in Euclid's *Elements.*

Lemma 1.32. Every number greater than 1 has at least one prime divisor.

Proof. Let n be an integer greater than 1. We'll find a prime divisor of n. Let m be the smallest divisor of n greater than 1. (Note that we're using the minimization principle, also called the well-ordering principle, to conclude that such an m exists.) We'll show that m is prime thereby proving the lemma. We'll do that with a proof by contradiction, and that means that first we'll suppose that m is not prime, then derive a contradiction, and that will imply that m must be prime.

Suppose m is not prime, but composite. Them m is the product of two integers, j and k, each greater than 1. Now, $k\big|m$ and $m\big|n$, so $k\big|n$. But $k < m$. That gives us a divisor of n which is even smaller than m but still greater than 1. That contradicts the fact that m is the smallest divisor of n greater than 1. Thus, m is prime, and it's a divisor of n. Q.E.D.

Now we can prove one of the two statements.

Theorem 1.33 (Euclid)**.** Every number greater than 1 is either a prime or the product of primes.

Proof. This will be another proof by contradition that uses the well-ordering principle.

Suppose that the theorem is false. Then there is some composite number greater than 1 that that is not the product of primes. Let n be the smallest such. By our lemma, this n has some prime divisor, call it p. Then $m = n/p$ is a number smaller than n but larger than 1, so, by the minimality of n, m is either prime or the product of primes. In the first case, when m is prime, then $n = pm$ is the product of two primes. In the second case when m is a product of primes, then $n = pm$ is also a product of primes. In any case, n is the product of primes, a contradiction. Thus, the theorem is true. Q.E.D.

This last theorem is part of the so-called fundamental theorem of arithmetic that says every number greater than 1 can be uniquely factored as a product of primes. So far, we only have that every number is a product of primes, but we haven't seen the uniqueness. We'll prove that pretty soon.

Next, let's prove the other statement, that there are infinitely many primes. This is Euclid's proof.

Theorem 1.34 (Euclid)**.** There are infinitely many primes.

Proof. Again, this is a proof by contradiction.

Suppose that there are only finitely many primes, namely, $p_1, p_2, \ldots, p_k$. Let n be one more than the product of these primes,

$$n = p_1 p_2 \cdots p_k + 1.$$

By our lemma n has a prime factor, call it p. Since $p_1, p_2, \ldots, p_k$ are all the primes, therefore p must be one of them. Being one of them p divides their product $p_1 p_2 \cdots p_k$. But p also divides $n = p_1 p_2 \cdots p_k + 1$. Therefore, p divides the difference 1. But a prime p can't divide 1 since $p > 1$. From that contradiction, we conclude that there are infinitely many primes. Q.E.D.

1.3.4 The Euclidean algorithm

The Euclidean algorithm is an algorithm to compute the greatest common divisor of two natural numbers m and n.

Euclid defined the *greatest common divisor* of two natural numbers m and n, often denoted GCD(m, n) or more simply just (m, n), as the largest number d which is at the same time a divisor of m and a divisor of n.

There are two forms of the Euclidean algorithm. The first form, as Euclid stated it, repeatedly subtracts the smaller number from the larger replacing the larger by the difference, until the two numbers are reduced to the same number, and that's the greatest common divisor. (Note that the process has to stop by the well-ordering principle since at each step the larger number is reduced.)

The other form speeds up the process. Repeatedly divide the smaller number into the larger replacing the larger by the remainder. (This speeds up the process because if the smaller number is much smaller than the larger, you don't have to subtract it from the larger many times, just divide once and take the remainder which is the same as what you'd get if repeatedly subtracted it.)

This Euclidean algorithm works to produce the GCD, and the argument only depended on two properties of divisibility mentioned above, namely that if one number divides two other numbers, then it divides both their sum and difference.

Sometimes the GCD of two numbers turns out to be 1, and in that case we say the two numbers are *relatively prime.*

Theorem 1.35 (Euclidean algorithm)**.** Let d be the result of applying the Euclidean algorithm to m and n. Then d is the greatest common divisor GCD(m, n). Furthermore, the common divisors k of m and n are the divisors of GCD(m, n).

Proof. One step of the Euclidean algorithm replaces the pair (m, n) by $(m - n, n)$. It was mentioned above in the properties of divisibility that if one number divides two other numbers, then it divides both their sum and difference. Therefore, a number k divides both m and n if and only if k divides $m - n$ and n. Since the pair (m, n) have the same set of divisors as the pair $(m - n, n)$, therefore GCD$(m, n) =$ GCD$(m - n, n)$. Thus, at each step of the Euclidean algorithm the GCD remains invariant. Eventually, the two numbers are the same, but when that last step is reached, that number is the GCD. So, the end result of the Euclidean algorithm is $d = \text{GCD}(m, n)$.

The remarks above show that every divisor k of m and n also divides the result d of applying the Euclidean algorithm to m and n. Finally, if $k\big|d$, since $d\big|m$ and $d\big|n$, therefore $k\big|m$ and $k\big|n$. Q.E.D.

There's still more that we can get out of the algorithm. Let's use the division form for it. Let's suppose that $m > n$ to begin with. We divide n into m and get a quotient of q_1 and remainder of r_1, that is

$$m = q_1 n + r_1,$$

with r_1 between 1 and n. Then we work with n and r_1 instead of m and n. Divide r_1 into n to get q quotient of q_2 and a remainder of r_2, that is,

$$n = q_2 r_1 + r_2.$$

And we keep going until eventually we get a remainder of 0.

$$\begin{aligned} r_1 &= q_3 r_2 + r_3 \\ r_2 &= q_4 r_3 + r_4 \\ &\vdots \\ r_{s-3} &= q_{s-1} r_{s-2} + r_{s-1} \\ r_{s-2} &= q_s r_{s-1} + 0 \end{aligned}$$

We have

$$m > n > r_1 > r_2 > \cdots > r_{s-1}$$

and r_{s-1} is d, the GCD we're looking for.

We can use these equations to find d as a linear combination of the original numbers m and n as we did in an example last time. The first equation implies that r_1 is a linear combination of m and n. The next implies that r_2 is a linear combination of n and r_1, therefore a linear combination of m and n. Likewise the next shows r_3 is a linear combination of m and n, and so forth until we get to the next to the last equation which shows that r_{s-1}, which is the GCD of m and n is a linear combination of m and n. Thus, we've shown the following theorem.

Theorem 1.36 (Extended Euclidean algorithm)**.** The greatest common divisor $d = \text{GCD}(m, n)$ of m and n is a linear combination of m and n. That is, there exist integers a and b such that

$$d = am + bn.$$

Now that we have the major theorems on GCDs, there are a few more fairly elementary proprieties of GCDs that are straightforward to prove, such as these.

Theorem 1.37.
$(a, b + ka) = (a, b)$
$(ak, b,) = k(a, b)$
If $d = (a, b)$ then $(a/d, b/d) = 1$.

Greatest common divisors of more than two numbers The GCD of more than two numbers is defined the same way as for two numbers: the GCD of a set of numbers the largest number that divides them all. For example, $\gcd(14, 49, 91) = 7$. To find a GCD of three numbers, a, b, and c, first find $d = \gcd(a, b)$, then find $e = \gcd(d, c)$. Thus,

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

a statement that is easy to show.

Pairwise relatively prime numbers A set of numbers is said to be *pairwise relatively prime* if any two of them are relatively prime. For instance, 15, 22, and 49 are three pairwise relatively prime numbers. Thus, a, b, and c are pairwise relatively prime when

$$\gcd(a, b) = \gcd(a, c) = \gcd(b, c) = 1.$$

Note that $\gcd(a, b, c)$ can be 1 without a, b, and c being pairwise relatively prime. For instance, $\gcd(4, 3, 9) = 1$, but $\gcd(3, 9) = 3$.

Least common multiples The *least common multiple* of a set of positive integers is the smallest positive integer that they all divide. It is easy to show that the greatest common divisor of two integers times their least common multiple equals their product.

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

1.4 The fundamental theorem of arithmetic: the unique factorization theorem

We proved above that every natural number could be factored as a product of primes. But we want more than existence, we want uniqueness, that there is only one way that it can be factored as a product of primes.

The unique factorization theorem, a.k.a., the fundamental theorem of arithmetic Now, in order to make this general statement valid we have to extend a little bit what we mean by a product. For example, how do you write a prime number like 7 as a product of primes? It has to be written as the product 7 of only one prime. So we will have to accept a single number as being a product of one factor.

Even worse, what about 1? There are no primes that divide 1. One solution is to accept a product of no factors as being equal to 1. It's actually a reasonable solution to define the empty product to be 1, but until we find another need for an empty product, let's wait on that and restrict this unique factorization theorem to numbers greater than 1. So, here's the statement of the theorem we want to prove.

Theorem 1.38 (Unique factorization theorem)**.** Each integer n greater than 1 can be uniquely factored as a product of primes. That is, if n equals the product $p_1p_2 \cdots p_r$ of r primes, and it also equals the product $q_1q_2 \cdots q_s$ of s primes, then the number of factors in the two products is the same, that is $r = s$, and the two lists of primes $p_1, p_2, \ldots, p_r$ and $q_1, q_2, \ldots, q_s$ are the same apart from the order the listings.

We'll prove this by using a form of induction. The form that we'll use is this:

> In order to prove a statement $S(n)$ is true for all numbers, prove that $S(n)$ follows from the assumption that $S(k)$ is true for all $k < n$.

This principle of induction appears to be stronger than the one we've used before, but, in fact, it is equivalent to it. It's really the same as the minimization principle (i.e. well-ordering principle) applied to the negation of the statement. The advantage in using it is that a proof by contradiction is not needed making the proof more understandable.

Proof. We'll prove the unique factorization theorem in two cases. Case 1 will be where n is a prime number itself. Case 2 will be where n is composite.

Case 1: Suppose that n is a prime number. The only way that a prime number can be written as a product of primes is as itself; otherwise it would not be prime, but composite.

Case 2: Suppose that n is a composite number equal to both products of primes $p_1p_2\cdots p_r$ and $q_1q_2\cdots q_s$. Note that since n is composite, both r and s are at least 2; otherwise it would not be composite, but prime.

Now look at one of the primes, say p_1. It divides n, so it divides the product of the other primes $q_1q_2\cdots q_s$. We suspect that that implies it has to be one of those other primes. Let's put that off for a bit; that is, logically before we prove this theorem, we need to prove another theorem, listed next, that if a prime divides a product of primes, then it is one of those primes; but we'll actually do that next. Assuming we've done that, then we can conclude that p_1 is one of the q_i's. We can reorder the product $q_1q_2\cdots q_s$ to make it q_1 so that p_1 equals q_1. Now, since $p_1p_2\cdots p_r = q_1q_2\cdots q_s$ and the first first factors of the two products are equal, therefore $p_2\cdots p_r = q_2\cdots q_s$. Now, by our new induction principle, these are two prime factorizations of a number smaller than n, and hence are the same, except for their order. Therefore, they have the same number of factors, that is, $r = s$, and all the factors are the same except for their order. And the number n is that product times p_1, which equals q_1, therefore the original two products, $p_1p_2\cdots p_r$ and $q_1q_2\cdots q_s$, are the same except for order. Q.E.D.

Well, that finished the proof except we have to prove another theorem first, namely, the following one.

Theorem 1.39. If a prime divides a product of primes $q_1q_2\ldots q_s$, then it equals one of the primes $q_1, q_2, \ldots, q_s$.

We could do that, but we we'll prove a slightly stronger theorem, namely, the following one.

Theorem 1.40. If a prime divides a product of numbers $b_1b_2\ldots b_s$, then it divides one of the numbers $b_1, b_2, \ldots, b_s$.

Now the reason this theorem implies the previous theorem is because if a prime p divides a product of primes $q_1q_2\ldots q_s$, then it divides one of the primes $q_1, q_2, \ldots, q_s$, but the only way that one prime can divide another is if it equals the other.

Proof. A product of s numbers $b_1 b_2 \ldots b_s$ is actually a series of binary products. It's b_1 times $b_2 \ldots b_s$, and $b_2 \ldots b_s$ is b_2 times $b_3 \cdots b_s$, etc, where the last product is $b_{s-1} b_s$ is the product of b_{s-1} times b_s. That means that if we knew the following theorem, then, using ordinary induction, we could conclude this one. Q.E.D.

Theorem 1.41. If a prime divides a product of two numbers, then it divides one of the numbers.

Now, we could prove this theorem directly, but it turns out that there is a slightly stronger version that we can use in other places, so let's prove it, the one listed next, instead, and show this theorem follows from it.

Theorem 1.42. If n and a are relatively prime, and $n \big| ab$, then $n \big| b$.

Proof. That this theorem implies implies the previous one. Suppose that a prime p divides ab. If p doesn't divide a, then it's relatively prime to a, so by this theorem, it divides b. Therefore, either $p \big| a$ or $p \big| b$. Q.E.D.

Proof. Of this theorem. Suppose that $\textsc{gcd}(n, a) = 1$. Then, by the extended Euclidean algorithm, 1 is a linear combination of n and a, that is, there exist integers t and u such that

$$1 = tn + ua.$$

Multiply that equation by b to get

$$b = tnb + uab.$$

Now, if $n \big| ab$, then n divides the right hand side of the equation, but that equals the left hand side, so $n \big| b$. Q.E.D.

Comment 1.43. Typically in a mathematics book those theorems that come first logically are presented first. Here we started with our goal and discovered the theorems that were needed to prove the goal. (Actually, I made the list longer than it needed to be by strengthening a couple of them because the stronger versions are more useful, something you can only tell with hindsight.)

The advantage to presenting theorems in their logical order is that it is easier to follow the logic. The disadvantage is that the motivation for the preliminary theorems is not apparent until the final theorem, the interesting one, is reached.

Usually when we write the prime factorization of a number, we'll use exponents on those primes that are repeated. For instance, the number 40 had the prime factorization $2 \cdot 2 \cdot 2 \cdot 5$. An abbreviated form for this factorization is $2^3 \cdot 5$. We say that the prime 2 occurs with multiplicity 3, while the prime 5 occurs with multiplicity 1. The multiplicities are the exponents. So, in general, a number n has the prime factorization

$$n = p_1^{e_1} p_2^{e_1} \cdots p_k^{e_k}$$

where the primes $p_1, p_2, \ldots, p_k$ are all distinct, and their multiplicities are the exponents $e_1, e_2, \ldots, e_k$, respectively.

These exponents are called the orders of the primes in n. The *order* of p in n be the exponent of p in the prime factorization of n, denoted $\operatorname{ord}_p a$.

Immediate corollaries to the unique factorization theorem. A corollary is a theorem that logically follows very simply from a theorem. Sometimes it follows from part of the proof of a theorem rather than from the statement of the theorem. In any case, it should be easy to see why it's true. We can draw a couple of corollaries from the unique factorization theorem.

Corollary 1.44. The only primes that can divide a number n are the ones that appear in its prime factorization $p_1^{e_1} p_2^{e_1} \cdots p_k^{e_k}$.

Corollary 1.45. If the prime factorizations of m and n are $m = p_1^{e_1} p_2^{e_1} \cdots p_k^{e_k}$ and $n = p_1^{f_1} p_2^{f_1} \cdots p_k^{f_k}$ (where here some of the e_i's and f_i's equal 0 so we can use the same list of primes for both numbers), then their greatest common divisor $d = \text{GCD}(m, n)$ has the prime factorization $d = p_1^{g_1} p_2^{g_1} \cdots p_k^{g_k}$ where each exponent g_i is the minimum of the corresponding exponents e_i and f_i.

As an example of the last corollary, if $m = 1260 = 2^2 3^2 5^1 7^1$ and $n = 600 = 2^3 3^1 5^2$, then their GCD is $d = 2^2 3^1 5^1 = 60$.

Chapter 2

Fields

Informally, a field is a set equipped with four operations—addition, subtraction, multiplication, and division that have the usual properties.

2.1 Introduction to fields

A *field* is a set equipped with two binary operations, one called *addition* and the other called *multiplication*, denoted in the usual manner, which are both commutative and associative, both have identity elements (the additive identity denoted 0 and the multiplicative identity denoted 1), addition has inverse elements (the inverse of x denoted $-x$), multiplication has inverses of nonzero elements (the inverse of x denoted $\frac{1}{x}$), multiplication distributes over addition, and $0 \neq 1$.

2.1.1 Definition of fields

Here's a more complete definition.

Definition 2.1 (field)**.** A *field* F consists of

1. a set, also denoted F and called the *underlying set* of the field;
2. a binary operation $+ : F \times F \to F$ called *addition*, which maps an ordered pair $(x, y) \in F \times F$ to its *sum* denoted $x + y$;
3. another binary operation $\cdot : F \times F \to F$ called *multiplication*, which maps an ordered pair $(x, y) \in F \times F$ to its *product* denoted $x \cdot y$, or more simply just xy;

   such that

4. addition is commutative, that is, for all elements x and y, $x + y = y + x$;
5. multiplication is commutative, that is, for all elements x and y, $xy = yx$;
6. addition is associative, that is, for all elements x, y, and z, $(x + y) + z = x + (y + z)$;
7. multiplication is associative, that is, for all elements x, y, and z, $(xy)z = x(yz)$;

8. there is an additive identity, an element of F denoted 0, such that for all elements x, $0 + x = x$;

9. there is a multiplicative identity, an element of F denoted 1, such that for all elements x, $1x = x$;

10. there are additive inverses, that is, for each element x, there exists an element y such that $x + y = 0$; such a y is called the *negation* of x;

11. there are multiplicative inverses of nonzero elements, that is, for each nonzero element x, there exists an element y such that $xy = 1$; such a y is called a *reciprocal* of x;

12. multiplication distributes over addition, that is, for all elements x, y, and z, $x(y+z) = xy + xz$; and

13. $0 \neq 1$.

The conditions for a field are often call the *field axioms.*

Caveat: We're using the terminology and notation of arithmetic that we use for numbers, but the elements of our fields need not be numbers; often they will be, but sometimes they won't.

Note that we'll use the standard notational conventions on precedence for all fields so we don't have to fully parenthesize every expression. Multiplication and division have a higher precedence than addition and subtraction, so that, for example, $x - y/z$ means $x - (y/z)$, not $(x-y)/z$. Also, operations are executed from left to right, so that $x-y-z$ means $(x-y)-z$, not $x - (y - z)$. (An exception is that exponention is executed from right to left, so that x^{m^n} means $x^{(m^n)}$, not $(x^m)^n$.)

Commutativity and associativity of addition imply that terms can be added in any order, so of course we won't put parentheses when we're adding more than two terms together. Likewise for multiplication.

Although in parts 10 and 11 of the definition only the existence of an additive and multiplicative inverses is required, you can easily show uniqueness follows from the definition. Once that is done we can note that the additive inverse of x is called *the negation* of x and denoted $-x$, and the multiplicative inverse of x, when x is not 0, is called *the reciprocal* of x and denoted $1/x$ or x^{-1}.

2.1.2 Subtraction, division, multiples, and powers

With the help of negation, we can define subtraction as follows. The *difference* of two elements x and y is defined as $x-y = x+(-y)$. Likewise, with the help of reciprocation, we can define division. The *quotient* of two elements x and $y \neq 0$ is $x/y = xy^{-1}$. The expected properties of subtraction and division all follow from the definition of fields. For instance, multiplication distributes over subtraction, and division by z distributes over addition and subtraction.

Likewise, we can define integral multiples of elements in a field. First, we'll define nonnegative multiples inductively. For the base case, define $0x$ as 0. Then define $(n+1)x$ as $x + nx$ when n is a nonnegative integer. Thus nx is the sum of n x's. For instance, $3x = x + x + x$.

Then if $-n$ is a negative integer, we can define $-nx$ as $-(nx)$. The usual properties of multiples, like $(m+n)x = mx + nx$ will, of course, hold.

Furthermore, we can define integral powers of x. Define x^1 as x for a base case, and inductively for nonnegative n, define x^{n+1} as xx^n. Thus nx is the product of n x's. For instance, $x^3 = xxx$. Next, define x^0 as 1, so long as $x \neq 0$. (0^0 should remain undefined, but for some purposes, especially in algebra, it's useful to define 0^0 to be 1.) Finally, if $-n$ is positive and $x \neq 0$, define x^{-n} as $(x^n)^{-1}$. The usual properties of integral powers hold, like $x^{m+n} = x^m x^n$ and $(xy)^n = x^n y^n$.

2.1.3 Properties that follow from the axioms

There are numerous useful properties that are logical consequences of the axioms. Generally speaking, the list of axioms should be short, if not minimal, and any properties that can be proved should be proved. Here's a list of several things that can be proved from the axioms. We'll prove a few in class, you'll prove some as homework, and we'll leave the rest. (They make good questions for quizzes and tests.) In the following statements, unquantified statements are meant to be universal with the exception that whenever a variable appears in a denominator, that variable is not to be 0.

Exercise 2.1. Prove that 0 is unique. That is, there is only one element x of a field that has the property that for all y, $x + y = y$. Likewise, prove that 1 is unique.

Exercise 2.2. Prove that the inverses of the identity elements are themselves, that is, $-0 = 0$, and $1^{-1} = 1$.

Exercise 2.3. Prove that multiplication distributes over subtraction: $x(y-z) = xy - xz$.

Exercise 2.4. Prove that 0 times any element in a field is 0: $0x = 0$.

Exercise 2.5. Prove the following properties concerning multiplication by negatives: $(-1)x = -x$, $-(-x) = x$, $(-x)y = -(xy) = x(-y)$, and $(-x)(-y) = xy$.

Exercise 2.6. Prove the following properties concerning reciprocals: $(x^{-1})^{-1} = x$, and $(xy)^{-1} = x^{-1}y^{-1}$.

Exercise 2.7. Prove that $\dfrac{x}{y} = \dfrac{w}{z}$ if and only if $xz = yw$.

Exercise 2.8. Prove the following properties concerning division: $\dfrac{x}{y} \pm \dfrac{w}{z} = \dfrac{xz \pm yw}{yz}$, $\dfrac{x}{y}\dfrac{w}{z} = \dfrac{xw}{yz}$, and $\dfrac{x}{y}\Big/\dfrac{w}{z} = \dfrac{xz}{yw}$.

Exercise 2.9. Prove that if $xy = 0$, then either $x = 0$ or $y = 0$.

2.1.4 Subfields

Frequently we'll find one field contained in another field. For instance, the field of rational numbers **Q** is part of the field of real numbers **R**, and **R** is part of the field of complex numbers **C**. Here's the precise definition of subfield.

Definition 2.2 (subfield). A field E is a *subfield* of a field F if

1. the underlying set of E is a subset of the underlying set of F;
2. the addition operation $+_E$ on E is the restriction of the addition operation $+_F$ on F, that is, for all x and y in E, $x +_E y = x +_F y$; and
3. the multiplication operation $\cdot_E$ on E is the restriction of the multiplication operation $\cdot_F$ on F, that is, for all x and y in E, $x \cdot_E y = x \cdot_F y$.

When E is a subfield of F, we'll also say that F is an *extension* of E.

There is an alternate characterization of subfield. You can easily prove the theorem, but there are many steps.

Theorem 2.3. If a subset E of a field F has 0, 1, and is closed under addition, multiplication, negation, and reciprocation of nonzero elements, then E is a subfield of F.

The field of rational numbers Q. When we're trying to find the smallest example of a field, it looks like it will have to be **Q**. Later we'll see that it's not the smallest! But here's an argument (which must have a flaw in it) which says we need all the rational numbers to be in any field F.

To begin with, 0 and 1 have to be in F. But we also have to have $1+1$ in F and we'll denote that 2, of course. And we'll need $1+1+1 = 2+1$ which we'll denote 3. And so forth, so we've got 0 and all the positive integers in F. We also need negations of them, so all the negative integers are in F, too. But a rational number m/n is just an integer m divided by a positive integer n, so we'll have to have all rational numbers in F. That shows that **Q** is a subfield of F.

Thus, it looks like every field F includes the smallest field **Q**, the field of rational numbers.

There's one minor flaw in the argument above, but let's not pick it apart right now. Pretty soon we'll look at fields that don't contain **Q**.

2.1.5 Polynomials and fields of rational functions

We won't formally define polynomials. For now, we'll only look at polynomials in one variable x, but later we'll look at polynomials in two or more variables.

Informally a *polynomial* $f(x)$ with coefficients in a field F is an expression

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

where each coefficient $a_i \in F$.

It's convenient to denote a polynomial either by f or by $f(x)$. If the variable x is referred to somewhere nearby, then I'll use $f(x)$, otherwise I'll just use f. For instance, if I want to multiply two polynomials f and g together, I'll write fg, but if I want two multiply f by $x^2 - 3x + 2$, I'll write $f(x)\,(x^2 - 3x + 2)$.

The set of all polynomials is denoted $F[x]$. It has addition, subtraction, and multiplication, and satisfies the requirements of a ring. So $F[x]$ is the ring of polynomials with coefficients in F. But $F[x]$ doesn't have reciprocals, so it's not a field. Nonetheless, the field F is a subring of the ring $F[x]$; we can identify the constant polynomials as the elements of F.

A *rational function* with coefficients in F is a quotient of two polynomials $f(x)/g(x)$. Rational functions do form a field, the field $F(x)$ of rational functions with coefficients in F.

For example, one rational function in $\mathbf{Q}(x)$ is $\dfrac{5x^2 - 3x + 1/2}{x^3 + 27}$.

Note that the field F is a subfield of the $F(x)$. Again, we can identify the constant rational functions as the elements of F. For example, $\mathbf{Q}$ is a subfield of $\mathbf{Q}(x)$, and both $\mathbf{R}$ and $\mathbf{Q}(x)$ are subfields of $\mathbf{R}(x)$.

2.1.6 Vector spaces over arbitrary fields

When you studied vector spaces, you probably only studied vector spaces over the real numbers, although vector spaces over other fields might have been mentioned. In fact, vector spaces over an arbitrary field F have the same basic properties as vector spaces over $\mathbf{R}$.

The n-dimensional vector space F^n is defined the same way as $\mathbf{R}^n$ except the n-tuples have coordinates in F. Addition, scalar multiplication, inner products (also called dot products) are defined the same way for F^n as they are for $\mathbf{R}^n$. Likewise, cross products are defined for F^3 the same way as they are in $\mathbf{R}^3$.

Furthermore matrices in $M_{m\times n}(F)$ are defined the same way as matrices in $M_{m\times n}(\mathbf{R})$ except the entries are in F instead of $\mathbf{R}$, and the matrix operations are the same. You can use the same methods of elimination to solve a system of linear equations with coefficients in F or find the inverse of a matrix in $M_{n\times n}(F)$ if its determinant is nonzero. Determinants have the same properties. You can use methods of linear algebra to study geometry in F^n just as you can for $\mathbf{R}^n$ (although it may not be possible to visualize what F^n is supposed to look like, and things like areas of triangles have values in F).

The abstract theory of finite dimensional vector spaces over F is the same, too. Linear independence, span, basis, dimension are all the same. Rank and nullity of a matrix, the same. Change of basis is the same.

Eigenvalues, eigenvectors, and eigenspaces may have problems over some fields. In fact, when you studied transformations $\mathbf{R}^n \to \mathbf{R}^n$, sometimes you had complex eigenvalues and their only eigenvectors were in $\mathbf{C}^n$. Likewise when looking at transformations $F^n \to F^n$ and the eigenvalues aren't in F, you'll may have to go to some field extension F' of F to find them and to F'^n to find the eigenvectors.

Likewise, canonical forms for matrices will depend on F.

2.2 Cyclic rings and finite fields

In this section we'll look at fields that are finite, and we'll discover that $\mathbf{Q}$ actually isn't the smallest field. Although they're smaller fields—they're finite—they won't be subfields of $\mathbf{Q}$.

First we'll look a bit at the concept of congruence modulo n, where n is a positive integer. Then we'll look at the ring of integers modulo n, denoted $\mathbf{Z}/n\mathbf{Z}$ or more simply $\mathbf{Z}_n$. We'll see why they're called cyclic rings. Finally, we'll look at the case where n is prime, and we'll denote it p then, where $\mathbf{Z}_p$ turns out to be a field, and we'll examine some of the cyclic fields.

Definition 2.4. Fix n, a positive integer. We say that two integers x and y are *congruent modulo* n if n evenly divides the difference $x-y$. We'll use the standard notation from number theory

$$x \equiv y \pmod{n}$$

to indicate that x is congruent to y modulo n, and the notation $n|m$ to indicate that the integer n divides the integer m (with no remainder). Then

$$x \equiv y \pmod{n} \quad \text{iff} \quad n\big|(x-y).$$

When n doesn't divide the difference $x - y$, we say a is not congruent to b, denoted $x \not\equiv y \pmod{n}$.

You're familiar with congruence modulo 12; it's what 12-hour clocks use.

We'll look at the general theory of equivalence relations before coming back to $\mathbf{Z}_n$.

2.2.1 Equivalence relations

Our primary example right now of an equivalence relation is congruence modulo n, $x \equiv y \pmod{n}$, but we'll have several other equivalence relations later.

There are various symbols used for equivalence relations, such as $\cong$, $\equiv$, $\approx$, $\asymp$, $\simeq$, $\sim$, and so forth. We'll stick with $\equiv$ for now.

Definition 2.5. An *equivalence relation* $\equiv$ on a set S is a relation that is reflexive, symmetric, and transitive.

A relation on a set S may be identified with a subset of the product $S \times S$. For an equivalence relation $\equiv$, this means $x \equiv y$ corresponds to the statement that the ordered pair (x, y) is an element of that subset.

Reflexive: $\forall x$, $x \equiv x$.

Symmetric: $\forall x, \forall y$, $x \equiv y$ implies $y \equiv x$.

Transitive: $\forall x, \forall y, \forall z$, $x \equiv y$ and $y \equiv z$ implies $x \equiv z$.

Theorem 2.6. Congruence modulo n is an equivalence relation.

Proof. For reflexivity, $x \equiv x \pmod{n}$ holds since $n\big|(x-x)$.

For symmetry, we need to show that $x \equiv y \pmod{n}$ implies $y \equiv x \pmod{n}$. But if $n\big|(x-y)$, then $n\big|(y-x)$.

For transitivity, suppose that $x \equiv y \pmod{n}$ and $y \equiv z \pmod{n}$. Then $n\big|(x-y)$ and $n\big|(y-z)$, so there exist k and m such that $nk = x - y$ and $nm = y - z$. Therefore $n(k+m) = x - z$, showing that $n\big|(x-z)$. Hence $x \equiv z \pmod{n}$. Q.E.D.

Equivalence classes and partitions of sets. An equivalence relation on a set determines a partition on that set, and conversely, as we'll see presently.

Definition 2.7. Given an equivalence relation on a set, an *equivalence class* of an element x, denoted $[x]$, is the set of all elements equivalent to x,

$$[x] = \{y \,|\, y \equiv x\}.$$

You can easily show the several properties of equivalence classes.

Theorem 2.8. If $\equiv$ is an equivalence class on a set S, then the following four statements are equivalent

1. $x \equiv y$.
2. $[x] = [y]$.
3. $x \in [y]$.
4. $[x] \cap [y] \neq \emptyset$.

Furthermore, for each $x \in S$, there is exactly one equivalence class containing x, namely, $[x]$.

Definition 2.9. A *partition* of a set S is a collection of nonempty subsets, called *parts*, of S which are pairwise disjoint and whose union is all of S. Thus, each element of S belongs to exactly one of the parts.

The above theorem shows that the equivalence classes form a partition. The converse is also true as you can easily show.

Theorem 2.10. For each equivalence class on a set, the equivalence classes partition the set. Conversely, a partition of a set determines an equivalence relation where two elements are equivalent if they're in the same part.

The set of equivalence classes is sometimes denoted $S/_\equiv$, and it's sometimes called a quotient set. Using equivalence classes to construct new sets of things is a common practice in mathematics and especially in algebra.

Keep in mind that you can always name an element of $S/_\equiv$ by naming an element of S, but two elements x and y of S will name the same element of $S/_\equiv$, that is, $[x] = [y]$, if $x \equiv y$.

The function $\gamma : S \to S/_\equiv$ defined by $\gamma(x) = [x]$ is called a *projection*, or the *canonical function*, from the set to its quotient set.

2.2.2 The cyclic ring $\mathbf{Z}_n$

Definition 2.11. The integers modulo n, $\mathbf{Z}_n$ is the set of equivalence classes of integers.

We'll denote these equivalence classes with square brackets subscripted by n. Thus, for instance, the element 0 in $\mathbf{Z}_6$ is really $[0]_6$, which we'll denote $[0]$ when modulo 6 is understood. This equivalence class is the set of all x such that $x \equiv 0 \pmod 6$. This $[0]_6 = \{\ldots, -18, -12, -6, 0, 6, 12, 18, \ldots\}$. Likewise the element 1 in $\mathbf{Z}_6$ is really the equivalence class of 1, which is the set

$$[1]_6 = \{x \in \mathbf{Z} \,|\, x \equiv 1 \pmod 6\} = \{\ldots, -17, -11, -5, 1, 7, 13, 19, \ldots\}.$$

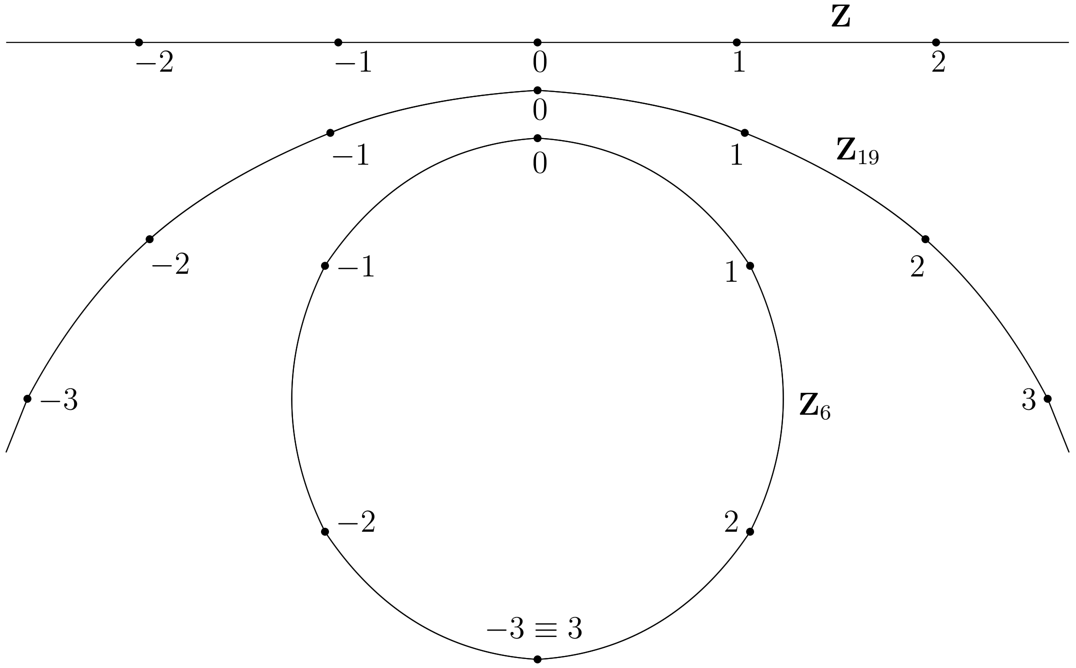
Note that $[1]_6 = [7]_6 = [13]_6$ all name the same equivalence class.

An equation in equivalence classes, such as $[x]_6 + [3]_6 = [5]_6$, is the same thing as an congruence, $x + 3 \equiv 5 \pmod 6$. The congruence notation is usually more convenient.

There are two ways you can think about integers modulo n. One is to think of them as regular integers from 0 through $n - 1$, do the arithmetic modulo n, and adjust your answer so it's in the same range. For example, we can take $\mathbf{Z}_6 = \{0,1,2,3,4,5\}$. Then, to do some computation, say $5(2-4) \pmod 6$, first compute $5(2-4)$ as integers to get -10, and then, since $-10 \equiv 2 \pmod 6$, say the answer is 2. That works very well for computation, but it's pretty messy when you're trying to do anything with variables or trying to prove anything in general.

A better way is to say that an element of $\mathbf{Z}_n$ is named by an integer, but two integers name the same element of $\mathbf{Z}_n$ if they're congruent modulo n. Thus, x and y name the same element of $\mathbf{Z}_n$ if $x \equiv y \pmod n$. This will work because congruence modulo n is an equivalence relation as we saw earlier.

In any case, it helps conceptually to think of the elements of $\mathbf{Z}_n$ as being arranged on a circle like we imagine the elements of $\mathbf{Z}$ being arranged on a line.



The operations on $\mathbf{Z}_n$. Our equivalence relation is congruence modulo n, so our equivalence classes are also called congruence classes.

Congruence modulo n is more than just an equivalence relation; it works well with addition, subtraction, and multiplication, as you can easily show.

Theorem 2.12. If $x \equiv y \pmod n$, and $u \equiv v \pmod n$, then $x+u \equiv y+v \pmod n$, $x-u \equiv y-v \pmod n$, and $xu \equiv yv \pmod n$.

These properties will allow us to define a ring structure on $\mathbf{Z}_n$, as done below.

But congruence modulo n doesn't work so well with division. Although $6 \equiv 0 \pmod 6$, it is not the case that $6/2 \equiv 0/2 \pmod 6$. Thus, we can't expect that $\mathbf{Z}_n$ will be a field, at least when $n = 6$.

Our job is to define addition, subtraction, and multiplication on $\mathbf{Z}_n$. Whenever a set is defined as a quotient set, as $\mathbf{Z}_n$ is, an extra step is required when defining an operation on it, as we'll see.

We would like to define addition on $\mathbf{Z}_n$ by saying $[x] + [u] = [x + u]$, that is, the sum of the equivalence class of x and the equivalence class of u should be the equivalence class of $x + u$. But what if we named the equivalence class x by some other integer, say y, and the equivalence of of u by some other integer v? How do we know we that $[y + v]$ is the same equivalence class as $[x + u]$? We can state this question in a couple of other ways. How do we know

$$[x] = [y] \text{ and } [u] = [v] \text{ implies } [x + u] = [y + v]?$$

That asks the question: how do we know

$$x \equiv y \pmod{n} \text{ and } u \equiv v \pmod{n} \text{ implies } x + u \equiv y + v \pmod{n}?$$

That's one of the properties of congruence mentioned above. That property says addition on $\mathbf{Z}_n$ is "well-defined".

Likewise, since multiplication works well with congruence,

$$x \equiv y \pmod{n} \text{ and } u \equiv v \pmod{n} \text{ imply } xu \equiv yv \pmod{n},$$

we can define multiplication on $\mathbf{Z}_n$ by $[x] \cdot [u] = [xu]$.

Furthermore, all the ring axioms will be satisfied in $\mathbf{Z}_n$ since they're satisfied in $\mathbf{Z}$. Thus, $\mathbf{Z}_n$ is a ring, and it's called a *cyclic ring*.

The projection $\gamma : \mathbf{Z} \to \mathbf{Z}_n$. The function $\gamma : \mathbf{Z} \to \mathbf{Z}_n$ defined by $\gamma(k) = [k]$ maps an integer to its equivalence class modulo n. We defined addition and multiplication in $\mathbf{Z}_n$

$$[x + u] = [x] + [u] \qquad \text{and} \qquad [xu] = [x]\,[u]$$

so γ preserves addition and multiplication. Furthermore, since $\gamma(1) = [1]$, it preserves 1. Therefore γ is a ring homomorphism. It is, of course, onto, so it is a ring epimorphism. It's called a *projection* or a *canonical homomorphism* to the quotient ring.

Later on, we'll generalize this construction to rings besides $\mathbf{Z}$ and their quotients, and we'll have projections for the generalizations, too.

The characteristic of a ring. What's weird about these cyclic rings is that if you start with 1 and add 1 over and over, you'll reach zero. For instance, in $\mathbf{Z}_5$, we have $1+1+1+1+1 = 5 \equiv 0 \pmod{5}$. This corresponds to the geometric interpretation of these cyclic rings being shaped like rings.

Definition 2.13. If some multiple of 1 equals 0 in a ring, then the *characteristic* of the ring is the smallest such multiple. If no multiple of 1 equals 0, then the characteristic is said to be 0.

We're primarily interested in characteristics when we're talking about fields, and we'll see soon that the characteristic of a field is either 0 or a prime number.

Example 2.14. The characteristic of $\mathbf{Z}_5$ is 5, and, in general, the characteristic of a finite cyclic ring $\mathbf{Z}_n$ is n.

2.2.3 The cyclic prime fields $\mathbf{Z}_p$

Since division doesn't work well with congruence, we can't expect $\mathbf{Z}_n$ to always have reciprocals, so we don't expect it to be a field. Let's first see when an element in $\mathbf{Z}_n$ is a unit, that is, it does have a reciprocal.

Theorem 2.15. An element k in $\mathbf{Z}_n$ is a unit if and only if k is relatively prime to n.

Proof. First, suppose that k is a unit in $\mathbf{Z}_n$. That means there exists l such that $kl \equiv 1 \pmod{n}$. Then $n|(kl-1)$, and hence n is relatively prime to k.

Second, suppose that k is relatively prime to n. Then, by the extended Euclidean algorithm, their greatest common divisor, 1, is a linear combination of k and n. Thus, there are integers x and y so that $1 = xk + yn$. Then $1 \equiv xk \pmod{n}$, and k does have a reciprocal, namely x, in $\mathbf{Z}_n$. Thus k is a unit in $\mathbf{Z}_n$. Q.E.D.

Theorem 2.16. The cyclic ring $\mathbf{Z}_n$ is a field if and only if n is prime.

Proof. Part of this theorem is a direct corollary of the previous one. Suppose n is prime. Then every nonzero element of $\mathbf{Z}_n$ is relatively prime to n. Therefore, $\mathbf{Z}_n$ is a field.

Next we'll show that if n is composite, the ring is not a field. Let n be the product of two integers m and k, both greater than 1. Then neither m nor k can have a reciprocal in $\mathbf{Z}_n$. Why not? Suppose that m^{-1} did exist in $\mathbf{Z}_n$. Then

$$\begin{aligned}(m^{-1}m)k &\equiv 1k \equiv k \pmod{n}\\ m^{-1}(mk) &\equiv m^{-1}n \equiv 0 \pmod{n}\end{aligned}$$

But $k \not\equiv 0 \pmod{n}$, a contradiction. So m^{-1} doesn't exist. Therefore, $\mathbf{Z}_n$ is not a field. Q.E.D.

Example 2.17. $\mathbf{Z}_2$. Note that there is only one nonzero element, namely 1, and it is its own inverse. The addition and multiplication tables for $\mathbf{Z}_2$ are particularly simple.

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

Note that subtraction is the same as addition in $\mathbf{Z}_2$ since $x - y \equiv x + y \pmod{2}$.

Example 2.18. $\mathbf{Z}_3$. Here, there are two nonzero elements, namely 1 and 2, but, for symmetry's sake, we'll call the two nonzero elements 1 and -1. Note that each of these two are their own inverses. The addition and multiplication tables are still pretty simple.

$+$	-1	0	1
-1	1	-1	0
0	-1	0	1
1	0	1	-1

$\cdot$	-1	0	1
-1	1	0	-1
0	0	0	0
1	-1	0	1

Example 2.19. $\mathbf{Z}_{13}$. What are the reciprocals of the 12 nonzero elements? We can name the nonzero elements as $\pm 1, \pm 2, \ldots, \pm 6$. You can verify that this table gives their inverses.

x	± 1	± 2	± 3	± 4	± 5	± 6
x^{-1}	± 1	∓ 6	∓ 4	∓ 3	∓ 5	∓ 2

For instance, the reciprocal of 2 is -6 since $2(-6) \equiv -12 \equiv 1 \pmod{13}$.

These fields, $\mathbf{Z}_p$ where p is prime, are the finite prime fields. But they're not all the finite fields.

Example 2.20. A field of order 9. We'll make an extension of $\mathbf{Z}_3$ to get a field of order 9. Note that -1 is not a square modulo 3. We can append $\sqrt{-1}$ to $\mathbf{Z}_3$ to get a field algebraic over it in exactly the same way we got $\mathbf{C}$ from $\mathbf{R}$. Let's use i as an abbreviation for $\sqrt{-1}$, as usual. Then

$$\mathbf{Z}_3(i) = \{x + yi \,|\, x, y \in \mathbf{Z}_3\}$$

Addition, subtraction, and multiplication give us no problems. We just have to check that nonzero elements have inverses. That's exactly as before.

$$\frac{1}{x+yi} = \frac{x-yi}{(x+yi)(x-yi)} = \frac{x-yi}{x^2+y^2} = \frac{x}{x^2+y^2} + \frac{-y}{x^2+y^2}i$$

Thus, if $x + yi$ is not 0 in $\mathbf{Z}_3(i)$, that is, not both of x and y are are congruent to 0 modulo 3, then $x^2 + y^2 \not\equiv 0 \pmod 3$, and the expression on the right gives $(x + yi)^{-1}$. Note that the characteristic of this field is 3 since $1 + 1 + 1$ is 0 in this field.

In fact, there are finite fields of order p^n for each power of a prime p. We'll prove that at some later time. These are called the Galois fields $GF(p^n)$. Note that cyclic prime field are the simplest Galois fields; $\mathbf{Z}_p$ is $GF(p)$.

2.2.4 Characteristics of fields, and prime fields

The characteristic of a ring was defined above, so we already have the definition for the characteristic of a field. Later when we look at the rest of the Galois fields, we'll see that the characteristic of the Galois field $GF(p^n)$ is p.

Those fields that have characteristic 0 all have $\mathbf{Q}$ as a subfield. The flawed proof we saw earlier included the mistaken assumption that all the elements $0, 1, 2, \ldots$ were distinct, which, as we've seen with these finite fields, isn't always the case. But we can correct the flawed proof to validate the following theorem. First, a definition.

Definition 2.21. A *prime field* is a field that contains no proper subfield. Equivalently, every element in it is a multiple of 1.

Theorem 2.22. *Each field F has exactly one of the prime fields as a subfield. It will have $\mathbf{Z}_p$ when it has characteristic p, but it will have $\mathbf{Q}$ if it has characteristic 0.*

The Frobenius endomorphism. Exponentiation to the power p has an interesting property when a ring R has prime characteristic p:

$$(x+y)^p = x^p + y^p$$

There are various ways to prove this. For instance, you can show that the binomial coefficient $\binom{p}{k}$ is divisible by p when $1 < k < p$.

This function $\varphi : R \to R$ defined by $\varphi(x) = x^p$ also preserves 1 and multiplication: $1^p = 1$ and $(xy)^p = x^p y^p$. Therefore, it is a ring endomorphism, called the *Frobenius* endomorphism.

We're most interested in the endomorphism when the ring is a field F of characteristic p. It's not particularly interesting when F is the prime field $\mathbf{Z}_p$ because it's just the identity function then. For other finite fields of characteristic p it will be an automorphism—it's a bijection since it's an injection on a finite set—and it's not the identity function for those fields.

Example 2.23. In the example above of the Galois field $GF(3^2) = \mathbf{Z}_3(i)$, the characteristic of the field is 3, so $\varphi(x+yi) = (x+yi)^3 = x^3 + (yi)^3 = x^3 - y^3 i = x - yi$. On the subfield $\mathbf{Z}_3$, φ is the identity, but not on all of $GF(3^2) = \mathbf{Z}_3(i)$, since $\varphi(i) = -i$.

2.3 Field Extensions, algebraic fields, the complex numbers

A lot of fields are found by extending known fields. For instance, the field of complex numbers $\mathbf{C}$ is extended from the field of real numbers $\mathbf{R}$. We'll look today at extending fields by adding square roots to known fields, the smallest kind of extension, called a *quadratic* extension. For example, $\mathbf{C}$ is a quadratic extension of $\mathbf{R}$.

2.3.1 An algebraic field

An *algebraic* number is a number x that is a root of a polynomial with rational coefficients. For instance, $x = \sqrt{2}$ is a root of the polynomial $x^2 - 2$. Algebraic fields are those that only contain algebraic numbers.

More generally, if x satisfies a polynomial equation $f(x) = 0$ where the polynomial f has coefficients in a field F, then we say x is *algebraic* over F. A field extension F' of F, all of whose elements are algebraic over F is said to be an *algebraic* extension of F.

Example 2.24. The field $\mathbf{Q}(\sqrt{2})$ is the smallest field that contains $\sqrt{2}$. In fact, its elements are all of the form

$$x + y\sqrt{2} \quad \text{where } x \in \mathbf{Q} \text{ and } y \in \mathbf{Q}.$$

It's pretty obvious that most of the field axioms hold. The only one that's not obvious is the existence of reciprocals, that is to say, the statement "$(x + y\sqrt{2})^{-1}$ is of the form $x' + y'\sqrt{2}$ where x' and y' are rational" is not so obvious. But the trick of "rationalizing the denominator" shows us how.

$$\frac{1}{x+y\sqrt{2}} = \frac{x-y\sqrt{2}}{(x+y\sqrt{2})(x-y\sqrt{2})} = \frac{x-y\sqrt{2}}{x^2-2y^2} = \frac{x}{x^2-2y^2} + \frac{-2y}{x^2-2y^2}\sqrt{2}$$

Thus, $\mathbf{Q}(\sqrt{2})$ is a field.

The trick was to multiply and divide by the *conjugate.* Let's give a notation to this conjugate: $\overline{x + y\sqrt{2}} = x - y\sqrt{2}$. Conjugation has some nice properties. It preserves all the elements of the base field $\mathbf{Q}$, that is, if $x \in \mathbf{Q}$, then $\overline{x} = x$. It preserves addition and multiplication, that is, if α and β are elements of $\mathbf{Q}(\sqrt{2})$, then $\overline{\alpha + \beta} = \overline{\alpha} + \overline{\beta}$ and $\overline{\alpha\beta} = \overline{\alpha}\,\overline{\beta}$. Finally, the operation of conjugation, $\overline{} : \mathbf{Q}(\sqrt{2}) \to \mathbf{Q}(\sqrt{2})$, is its own inverse, $\overline{\overline{\alpha}} = \alpha$. Thus, conjugation is a field automorphism. Furthermore, the elements α it fixes, $\overline{\alpha} = \alpha$, are just the elements of the base field $\mathbf{Q}$.

2.3.2 The field of complex numbers C

In the same way we just adjoined $\sqrt{2}$ to $\mathbf{Q}$ to get $\mathbf{Q}(\sqrt{2})$, we can adjoin $\sqrt{-1}$ to $\mathbf{R}$ to get $\mathbf{R}(\sqrt{-1})$, which is $\mathbf{C}$. Algebraically, the process is identical, but conceptually it's a little different because we thought that $\sqrt{2}$, being a real number, existed before we appended it to $\mathbf{Q}$, while it may not be so clear that $\sqrt{-1}$ exists before we append it to R. But $\sqrt{-1}$, usually denoted i, has the property $i^2 = -1$, so it is an algebraic number since it's the root of the polynomial $x^2 + 1$. In fact, $\mathbf{R}(i)$ consists of elements of the form

$$x + yi \quad \text{with} \quad x, y \in \mathbf{R}$$

as described by Euler. Addition and subtraction are "coordinatewise"

$$(x_1 + y_1 i) \pm (x_2 + y_2 i) = (x_1 + x_2) + (y_1 + y_2)i$$

while multiplication is only slightly more complicated

$$\begin{aligned}(x_1 + y_1 i)\,(x_2 + y_2 i) &= x_1x_2 + x_1y_2 i + x_2y_1 i + y_1y_2 i^2 \\ &= (x_1x_2 - y_1y_2) + (x_1y_2 + x_2y_1)i\end{aligned}$$

We can find reciprocals by rationalizing the denominator as we did above.

$$\frac{1}{x + yi} = \frac{x - yi}{(x + yi)(x - yi)} = \frac{x - yi}{x^2 + y^2} = \frac{x}{x^2 + y^2} + \frac{-y}{x^2 + y^2}i$$

We can define complex *conjugation* by $\overline{x + yi} = x - yi$. It's a field automorphism of $\mathbf{C}$, and its fixed subfield is $\mathbf{R}$.

We can also define a *norm* on $\mathbf{C}$ once we have conjugation. For $z = x + yi \in \mathbf{Q}$, let

$$|z|^2 = z\overline{z} = (x + yi)(x - yi) = x^2 + y^2.$$

Since $|z|^2$ is a nonnegative real number, it has a square root $|z|$.

A matrix representation of C. Consider the subset C of the matrix ring $M_2(\mathbf{R})$ consisting of matrices of the form

$$\begin{bmatrix} x & y \\ -y & x \end{bmatrix} \quad \text{where} \quad x, y \in \mathbf{R}.$$

You can easily show that this is a subring of $M_2(\mathbf{R})$ since the 0 matrix and the identity matrix are of this form, the sum and difference of matrices of this form are of this form, and so is the product as you can see here

$$\begin{bmatrix} x & y \\ -y & x \end{bmatrix} \begin{bmatrix} u & v \\ -v & u \end{bmatrix} = \begin{bmatrix} xu - yv & xv + yu \\ -yu - vx & -yv + xu \end{bmatrix}.$$

Thus, C is a subring of $M_2(\mathbf{R})$. Furthermore, it's a commutative subring even though $M_2(\mathbf{R})$ is not a commutative ring since the same product results when the two factors are interchanged:

$$\begin{bmatrix} u & v \\ -v & u \end{bmatrix} \begin{bmatrix} x & y \\ -y & x \end{bmatrix} = \begin{bmatrix} ux - vy & uy + vx \\ -vx - uy & -vy + ux \end{bmatrix}.$$

Furthermore C is a field because nonzero matrices in it have inverses. For suppose not both x and y are 0. Then

$$\begin{bmatrix} x & y \\ -y & x \end{bmatrix} \begin{bmatrix} \frac{x}{x^2+y^2} & \frac{-y}{x^2+y^2} \\ \frac{y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In fact, C is isomorphic to the complex field $\mathbf{C}$ as described above. The isomorphism is described by the one-to-one correspondence

$$\begin{bmatrix} x & y \\ -y & x \end{bmatrix} \leftrightarrow x + yi.$$

Note that a real number x corresponds to the matrix $\begin{bmatrix} x & 0 \\ 0 & x \end{bmatrix}$ while a purely imaginary number yi corresponds to the matrix $\begin{bmatrix} 0 & y \\ -y & 0 \end{bmatrix}$.

Note that complex conjugation in this representation is just matrix transposition.

This alternate representation of the complex numbers as matrices directly explains how a complex number acts as a linear transformation on the real plane $\mathbf{R}^2$. The complex number $x + yi$ maps a point (a, b) of $\mathbf{R}^2$ to the point $(ax + by, -ay + bx)$ since

$$\begin{bmatrix} x & y \\ -y & x \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} ax + by \\ -ay + bx \end{bmatrix}.$$

Matrix representations of various fields, rings, and groups are useful for two reasons. One is that they give us geometric interpretations for the elements as illustrated above. The other is that all the tools of linear algebra are available to us once we have the matrix representation.

2.3.3 General quadratic extensions

Now that we've seen a couple of quadratic extensions, let's see how it works in general.

Let F be a field and e an element of F that is not a square. In other words, the polynomial $x^2 - e$ has no roots in F. We'll consider ordered pairs $(a_1, a_2) \in F \times F$, but we'll write them as $a_1 + a_2\sqrt{e}$. We'll define addition coordinatewise

$$(a_1 + a_2\sqrt{e}) + (b_1 + b_2\sqrt{e}) = (a_1 + b_1) + (a_2 + b_2)\sqrt{e}$$

and define multiplication by

$$(a_1 + a_2\sqrt{e})\,(b_1 + b_2\sqrt{e}) = (a_1b_1 + ea_2b_2) + (a_1b_2 + a_2b_2)\sqrt{e}.$$

You can check that these definitions give us a ring. But, does it give us a field? As we did before, we'll find a reciprocal of a nonzero element $a_1 + a_2\sqrt{e}$

$$\frac{1}{a_1 + a_2\sqrt{e}} = \frac{a_1 - a_2\sqrt{e}}{(a_1 + a_2\sqrt{e})(a_1 - a_2\sqrt{e})} = \frac{a_1 - a_2\sqrt{e}}{a_1^2 - ea_2^2}$$

In order for this to be the reciprocal, all we have to do is show the denominator $a_1^2 - ea_2^2$ is not 0. In the case that $a_2 = 0$ we know $a_1 \neq 0$ since not both are 0, so in that case $a_1^2 - ea_2^2$ is not 0. That leaves us the case that $a_2 \neq 0$. Suppose that $a_1^2 - ea_2^2 = 0$. Then $ea_2^2 = a_1^2$, and dividing by a_2^2, we conclude $e = (a_1/a_2)^2$. But e is not a square in F. Thus $a_1^2 - ea_2^2$ is not 0 in this case, too. Therefore, we've found the reciprocal.

Thus, we have a field, $F(\sqrt{e})$.

When we look at more general field extensions, we'll have a lot more theory, and we won't have details to check as we did here. That theory will involve the concept of "ideals" in a ring.

2.4 Real numbers and ordered fields

We'll look now at **R**, the field of real numbers. What's so special about the real number field? For one thing, it's got an order on it; we can compare two real numbers x and y and say which is smaller or if they're equal. That's an extra structure on a field. We'll start by looking at this concept of ordered field.

That isn't enough to distinguish **R** from other fields. There are plenty of other ordered fields, such as **Q** and all the fields between **Q** and **R**.

2.4.1 Ordered fields

The easiest way to define an ordered field is by saying it's partitioned into positive elements, negative elements, and 0, and requiring a couple properties on these parts.

Definition 2.25 (Ordered field)**.** An *ordered field* consists of a field F along with a subset P whose elements are called *positive* such that

1. F is partitioned into three parts: P, $\{0\}$, and N where

   $$N = \{x \in F \mid -x \in P\}$$

   the elements of N are called *negative*;

2. the sum of two positive elements is positive; and

3. the product of two positive elements is positive.

Properties of ordered fields. You can show from this definition that

1. the sum of negative elements is negative
2. the product of a negative element and a positive element is negative
3. the product of two negative elements is positive
4. 1 is positive, and -1 is negative

Examples. **R**, **Q**, and all fields between them are ordered fields where the usual positive numbers in the field form P.

The binary order relations. From P we can define the binary order relations $<$, $\leq$, $>$, and $\geq$. For instance

$$x < y \quad \text{iff} \quad y - x \in P.$$

All the expected properties of these order relations follow. Here are a few.

1. Trichotomy: For each pair x, y, exactly one of the three relations $x < y$, $x = y$, or $x > y$ holds.
2. Transitivity: $x < y$ and $y < z$ imply $x < z$.
3. If x is positive and $y < z$, then $xy < xz$.
4. If x is negative and $y < z$, then $xy > xz$.
5. If $0 < x < y$, then $0 < 1/y < 1/x$.

Although **Q** and **R** are ordered fields, finite fields and **C** have no ordering.

Theorem 2.26. The characteristic of an ordered field is 0.

Proof. Suppose F is an ordered field of characteristic $p \neq 0$. Since 1 is positive, then any sum of 1s will be positive. Then p is positive. But p equals 0 which is not positive. A contradiction. Therefore an ordered field cannot have nonzero characteristic. Q.E.D.

Theorem 2.27. The complex field **C** has no order.

Proof. Suppose it did. Then either i or $-i$ would be positive. Since the square of a positive number is positive, therefore -1 is positive. But -1 is negative, a contradiction. Thus, **C** has no order. Q.E.D.

Example 2.28. An ordered extension of the real numbers with infinite elements and infinitesimal elements. We can give the field of rational functions $\mathbf{R}(x)$ an order as follows. First, we'll define when a polynomial $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ is positive, and that will be when its leading coefficient a_n is a positive real number. Next, we'll define when a rational function $f(x)/g(x)$ is positive, and that will be when f and g are both positive polynomials or both negative polynomials. It follows that $f(x)/g(x)$ is negative one of f

and g is positive and the other is negative. Only $0/g(x)$, which equals 0, won't be positive or negative. You can easily show that the sum and product of positive rational functions is positive.

The real numbers $\mathbf{R}$ is an ordered subfield of $\mathbf{R}(x)$, meaning that it's a subfield and its elements have the same order whether the order on $\mathbf{R}$ is used or the order on $\mathbf{R}(x)$ is used.

With this order, there are elements that are larger than any real number a, for example, $x > a$ since $x - a$ is positive. Likewise, there are positive elements that are smaller than any positive real number, $1/x$, for example.

2.4.2 Archimedean orders

The last example is an example of an ordered field with infinite elements and infinitesimals. Every ordered field F is an extension of $\mathbf{Q}$, so we can define an infinite element of F to be an element $x \in F$ greater than every rational number, and we can define a positive infinitesimal element as a positive $x \in F$ smaller than every positive rational number. Note that the reciprocal of an infinite element is an infinitesimal, and vice versa.

Definition 2.29. An *Archimedean ordered field* or, more simply, *Archimedean field* is simply an ordered field F without infinite elements or infinitesimals.

There are equivalent characteristics that could be used for the definition. Here are two. Each element of F is less than some integer. Each positive element of F is greater than the reciprocal of some positive integer.

Of course, the preceding example is a nonarchimedean field. Still, there are loads of Archimedean fields, namely $\mathbf{Q}$, $\mathbf{R}$, and all the intermediate fields. We still haven't answered the question about what makes $\mathbf{R}$ special. Before we go on, however, let's see how elements in an Archimedean field are determined by how they compare to rational numbers.

For an Archimedean field F, since F is ordered, it has characteristic 0, so it has as a subfield, indeed, an ordered subfield, the field of rational numbers $\mathbf{Q}$.

Theorem 2.30 (Density)**.** Between any two distinct elements of an Archimedean field, there lies a rational number.

Proof. Let $x < y$ in an Archimedean field. We're looking for a rational number $\frac{m}{n}$ between x and y. If x is negative while y is positive, then the rational number 0 lies between them. We can reduce the case where they're both negative to the case where they're both positive by noting that if $\frac{m}{n}$ lies between $-x$ and $-y$, then $-\frac{m}{n}$ lies between x and y.

So we may assume that both x and y are positive. If we can find some multiple n of them so that $ny - nx > 1$, then some integer m lies between ny and nx, but $nx < m < ny$ gives $x < \frac{m}{n} < y$. And we can find such a multiple since $y - x$ is greater than the reciprocal $\frac{1}{n}$ of some positive integer since the field is Archimedean. Q.E.D.

An element a of F partitions $\mathbf{Q}$ into two parts (L_a, R_a)

$$L_a = \{x \in \mathbf{Q} \,|\, x < a\} \text{ and } R_a = \{x \in \mathbf{Q} \,|\, x \geq a\}.$$

These two parts have a special property.

Definition 2.31. A *Dedekind cut* of the rational numbers is a partition of **Q** into two nonempty parts (L, R)—a left part L and a right part R—such that every element of L is less than every element of R. Furthermore, the left part does not have a greatest element.

Theorem 2.32. *An element a of an Archimedean field F is determined by its Dedekind cut (L_a, R_a). That is, if $(L_a, R_a) = (L_b, R_b)$, then $a = b$.*

Proof. If $a \neq b$, then there is a rational number between them, so that rational number will be in one left part but the other right part. Q.E.D.

In an Archimedean field F not every Dedekind cut has to determine an element. For example, in **Q**, the cut (L, R) where $L = \{x \,|\, x < 0 \text{ or } x^2 \leq 2\}$ and $R = \{x \,|\, x > 0 \text{ and } x^2 > 2\}$ is not the cut of any rational number. But that same cut for **R** is the cut of $\sqrt{2}$. The real numbers are special in that every cut is the cut of some real number.

Although there might not be a element of F for every cut, the cuts are enough to determine, along with the order on F and the field structure of **Q**, the field structure of F.

It helps in proofs to cut in half the information of a Dedekind cut from (L, R) to just L. It is sufficient to define a Dedekind cut just in terms of of the left part. You can prove the following lemma to simplify the statement and the proof of the following theorem.

Lemma 2.33. *If (L, R) is a Dedekind cut, then L has the following three properties*

i. *L is a nonempty, proper subset of **Q**;*

ii. *if $y \in L$ and $x \in \mathbf{Q}$ such that $x < y$, then $x \in L$; and*

iii. *for each $x \in C$, there exists $y \in C$ such that $x < y$*

Conversely, if L has these three properties, then (L, R) is a cut where R is the complement of L.

Theorem 2.34. *In an Archimedean field F, addition and multiplication are determined by Dedekind cuts in the sense that If a and b are two elements of F, then the left part of their sum $a + b$ is determined by their left parts*

$$L_{a+b} = \{x + y \,|\, x \in L_a \text{ and } y \in L_b\}.$$

If a and b are two positive elements of F, then the left part of their product is determined by their left parts

$$L_{ab} = \{xy \,|\, x \in L_a, x > 0, y \in L_b \text{ and } y > 0\} \cup \{x \,|\, x \leq 0\}.$$

2.4.3 Complete ordered fields

There are various definitions given for complete ordered fields, all logically equivalent. Here's one.

Definition 2.35. A *complete ordered field* is an Archimedean field that cannot be extended to a larger Archimedean field. Equivalently, every Dedekind cut determines an element of the field.

Completeness is the final property that characterizes **R**. Actually, right now we haven't proved that there is *at least* one complete ordered field, and we haven't proved that there is *only* one complete ordered field. Once we do, we can finally properly define **R**.

Existence of a complete ordered field We'll start by stating the theorem which gives the components for one way of constructing a complete ordered field F. To make it complete, we just have to make sure that every Dedekind cut determines an element of the field. The way to do that, of course, to define the field to be the cuts, and the definition of the operations of addition and multiplication are determined by the cuts as seen in the last theorem.

Theorem 2.36. There is a complete ordered field F. It's elements are Dedekind cuts of $\mathbf{Q}$. If L_1 and L_2 are left parts of two cuts, then the left part of the sum is determined by the left part

$$L_+ = \{x + y \,|\, x \in L_1 \text{ and } y \in L_2\}.$$

If L is the left part a positive cut (one that contains at least one positive rational number), then its negation is determined by the left part

$$L_- = \{-x \,|\, x \notin L\}$$

except, if this L_- has a largest element, that largest element is removed. If L_1 and L_2 are left parts of two positive cuts, then the left part of the product is determined by the left part

$$L_\times = \{xy \,|\, x \in L_1, x > 0, y \in L_2 \text{ and } y > 0\} \cup \{x \,|\, x \le 0\}.$$

There are many details to show to verify that R is a complete ordered field. First, that the sets L_+, L_-, and $L_\times$ are left parts. then the field axioms need to be verified, then the order axioms, then that's it's an Archimedean field. The last step, that it's complete is almost obvious from the construction. No one of these steps is difficult, but there are many details to check.

There are alternate ways to construct complete ordered fields. One is by means of Cauchy sequences. The spirit is different, but the result is the same, since, as we're about to see, there is only one complete ordered field.

Uniqueness of the complete ordered field We have to somehow exclude the possibility that there are two different Archimedean fields that can't be extended to larger Archimedean fields.

We don't want to count two isomorphic fields as being different, since, in essence, they're the same field but the names of the elements are just different. So, what we want is the following theorem.

Theorem 2.37. Any two complete ordered fields are isomorphic as ordered fields. Furthermore, there is only one isomorphism between them.

Proof. We may treat the field $\mathbf{Q}$ as a subfield of the two complete ordered fields F_1 and F_2. Then as a Dedekind cut determines an element $a_1 \in F_1$ and an element a_2 in F_2, we have a bijection $F_1 \to F_2$. You only need to verify that preserves addition and multiplication, which it does, since in an Archimedean ring, addition and multiplication are determined by Dedekind cuts. Q.E.D.

R is the complete ordered field We now know that there is only one complete ordered field up to isomorphism. Any such complete ordered field may be taken as the real numbers.

2.5 Skew fields (division rings) and the quaternions

> Sir William Rowan Hamilton, who early found that his road [to success with vectors] was obstructed—he knew not by what obstacle—so that many points which seemed within his reach were really inaccessible. He had done a considerable amount of good work, obstructed as he was, when, about the year 1843, he perceived clearly the obstruction to his progress in the shape of an old law which, prior to that time, had appeared like a law of common sense. The law in question is known as the *commutative* law of multiplication.
>
> Kelland and Tait, 1873

2.5.1 Skew fields (division rings)

Skew fields, also called division rings, have all the properties of fields except that multiplication need not be commutative. When multiplication is not assumed to be commutative, a couple of the field axioms have have to be stated in two forms, a left form and a right form. In particular, we require

1. there is a multiplicative identity, an element of F denoted 1, such that $\forall x, 1x = x = x1$;
2. there are multiplicative inverses of nonzero elements, that is, $\forall x \neq 0, \exists y, xy = 1 = yx$; and
3. multiplication distributes over addition, that is, $\forall x, \forall y, \forall z, x(y+z) = xy + xz$ and $\forall x, \forall y, \forall z, (y+z)x = yx + zx$.

All the other axioms remain the same, except we no longer require commutative multiplication.

The most important skew field is the quaternions, mentioned next. Waring showed that there were no finite skew fields that weren't fields (a difficult proof).

2.5.2 The quaternions H

We're not going to study skew fields, but one is of particular importance, the quaternions, denoted **H**. The letter **H** is in honor of Hamilton, their inventor.

We can define a quaternion a as an expression

$$a = a_0 + a_1 i + a_2 j + a_3 k$$

where a_0, a_1, a_2, and a_3 are real numbers and i, j, and k are formal symbols satisfying the properties

$$i^2 = j^2 = k^2 = -1$$

and

$$ij = k, jk = i, ki = j.$$

The i, j, and k are all square roots of -1, but they don't commute as you can show from the definition that

$$ji = -k, kj = -i, ik = -j.$$

This doesn't lead to a commutative multiplication, but note that if a is real (i.e., its pure quaternion parts a_1, a_2, and a_3 are all 0), then a will commute with any quaternion b.

Addition and subtraction are coordinatewise just like in $\mathbf{C}$. Here's multiplication.

$$\begin{aligned} & (a_0 + a_1 i + a_2 j + a_3 k)\,(b_0 + b_1 i + b_2 j + b_3 k) \\ = \; & (a_0 b_0 - a_1 b_1 - a_2 b_2 - a_3 b_3) \\ + \; & (a_0 b_1 + a_1 b_0 + a_2 b_3 - a_3 b_2) i \\ + \; & (a_0 b_2 - a_1 b_3 + a_2 b_0 + a_3 b_1) j \\ + \; & (a_0 b_3 + a_1 b_2 - a_2 b_1 - a_3 b_0) k \end{aligned}$$

It's easy to check that all the axioms for a noncommutative ring are satisfied. The only thing left to in order to show that $\mathbf{H}$ is a skew field is that reciprocals exist. We can use a variant of rationalizing the denominator to find the reciprocal of a quaternion.

$$\begin{aligned} \frac{1}{a_0 + a_1 i + a_2 j + a_3 k} &= \frac{a_0 - a_1 i - a_2 j - a_3 k}{(a_0 - a_1 i - a_2 j - a_3 k)(a_0 + a_1 i + a_2 j + a_3 k)} \\ &= \frac{a_0 - a_1 i - a_2 j - a_3 k}{a_0^2 + a_1^2 + a_2^2 + a_3^2} \end{aligned}$$

Thus, a nonzero quaternion $a_0 + a_1 i + a_2 j + a_3 k$, that is, one where not all of the real numbers a_0, a_1, a_2, and a_3 are 0, has an inverse, since the denominator $a_0^2 + a_1^2 + a_2^2 + a_3^2$ is a nonzero real number.

The expression $a_0 - a_1 i - a_2 j - a_3 k$ used to rationalize the denominator is the *conjugate* of the original quaternion $a_0 + a_1 i + a_2 j + a_3 k$. It's worthwhile to have a notation for it.

$$\overline{a_0 + a_1 i + a_2 j + a_3 k} = a_0 - a_1 i - a_2 j - a_3 k,$$

as we do for $\mathbf{C}$. We'll also define the *norm* of a quaternion a by $|a|^2 = a\overline{a}$. It's a nonnegative real number, so it has a square root $|a|$.

Thus, if a is a nonzero quaternion, then its inverse is $1/a = \overline{a}/|a|^2$.

For $\mathbf{C}$, the field of complex numbers, conjugation was a field automorphism, but for $\mathbf{H}$, it's not quite an automorphism. It has all of the properties of an automorphism except one. It preserves 0, 1, addition and subtraction $\overline{a \pm b} = \overline{a} \pm \overline{b}$, and reciprocation $\overline{1/a} = 1/\overline{a}$, but it reverses the order of multiplication $\overline{ab} = \overline{b}\,\overline{a}$. We'll call such a thing an *antiautomorphism*.

Theorem 2.38. The norm of a product is the product of the norms.

Proof. $|ab|^2 = ab\overline{ab} = ab\overline{b}\overline{a} = a|b|^2\overline{a} = a\overline{a}|b|^2 = |a|^2\,|b|^2$. Q.E.D.

If we unpack the equation $|a|^2\,|b|^2 = |ab|^2$, we'll get as a corollary Lagrange's identity on real numbers which shows how to express the product of two sums of four squares as the sum of four squares.

Corollary 2.39 (Lagrange)**.** The product of the sum of four squares of integers is a sum of four squares of integers

$$\begin{array}{rl} & (a_0^2 + a_1^2 + a_2^2 + a_3^2)\,(b_0^2 + b_1^2 + b_2^2 + b_3^2) \\ = & (a_0b_0 - a_1b_1 - a_2b_2 - a_3b_3)^2 \\ + & (a_0b_1 + a_1b_0 + a_2b_3 - a_3b_2)^2 \\ + & (a_1b_2 + a_2b_1 + a_3b_1 - a_1b_3)^2 \\ + & (a_2b_3 + a_3b_2 + a_1b_2 - a_2b_1)^2 \end{array}$$

Note that this equation not only works for real numbers, but also for integers, indeed when the coefficients lie in any commutative ring. Lagrange used this identity to show that every nonnegative integer n is the sum of four squares. The identity above is used to reduce the general case to the case when n is prime. Lagrange still had work to do to take care of the prime case.

A matrix representation for H. There are various matrix representations for **H**. This one will make **H** a subring of the real matrix ring $M_4(\mathbf{R})$. We'll represent 1 by the identity matrix, and i, j, and k by three other matrices which, you can verify, satisfy $i^2 = j^2 = k^2 = -1$ and $ij = k, jk = i, ki = j$.

$$1 \leftrightarrow \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad i \leftrightarrow \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$j \leftrightarrow \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \qquad k \leftrightarrow \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Then a generic quaternion $a + bi + cj + dk$ corresponds to the matrix

$$\begin{bmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{bmatrix}$$

Quaternions and geometry. Each quaternion a is the sum of a real part a_0 and a pure quaternion part $a_1i + a_2j + a_3k$. Hamilton called the real part a *scalar* and pure quaternion part a *vector*. We can interpret $a_1i + a_2j + a_3k$ as a vector $\mathbf{a} = (a_1, a_2, a_3)$ in $\mathbf{R}^3$. Addition and subtraction of pure quaternions then are just ordinary vector addition and subtraction.

Hamilton recognized that the product of two vectors (pure quaternions) had both a vector component and a scalar component (the real part). The vector component of the product $\mathbf{ab}$ of two pure quaternions Hamilton called the *vector product*, now often denoted $\mathbf{a} \times \mathbf{b}$ or $\mathbf{a} \wedge \mathbf{b}$, and called the *cross product* or the *outer product.* The negation of the scalar component Hamilton called the *scalar product*, now often denoted $\mathbf{a} \cdot \mathbf{b}$, $(\mathbf{a}, \mathbf{b})$, $\langle \mathbf{a}, \mathbf{b} \rangle$, or $\langle \mathbf{a} | \mathbf{b} \rangle$ and called the *dot product* or the *inner product.* Thus

$$\mathbf{ab} = \mathbf{a} \times \mathbf{b} - \mathbf{a} \cdot \mathbf{b}.$$

Hamilton's quaternions were very successful in the 19th century in the study of three-dimensional geometry.

Here's a typical problem from Kelland and Tait's 1873 *Introduction to Quaternions.* If three mutually perpendicular vectors be drawn from a point to a plane, the sum of the reciprocals of the squares of their lengths is independent of their directions.

Matrices were invented later in the 19th century. (But determinants were invented earlier!) Matrix algebra supplanted quaternion algebra in the early 20th century because (1) they described linear transformations, and (2) they weren't restricted to three dimensions.

Exercise 2.10. Show that $\mathbf{H}$ can be represented as a subring of the complex matrix ring $M_2(\mathbf{C})$ where

$$1 \leftrightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad i \leftrightarrow \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}$$
$$j \leftrightarrow \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \qquad k \leftrightarrow \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}$$

so that a generic quaternion $a + bi + cj + dk$ corresponds to the matrix

$$\begin{bmatrix} a+bi & c+di \\ -c+di & a-bi \end{bmatrix}$$

Chapter 3

Rings

Rings have the three operations of addition, subtraction, and multiplication, but don't need division. Most of our rings will have commutative multiplication, but some won't, so we won't require that multiplication be commutative in our definition. We will require that every ring have 1. The formal definition for rings is very similar to that for fields, but we leave out a couple of the requirements.

3.1 Introduction to rings

A *ring* is a set equipped with two binary operations, one called *addition* and the other called *multiplication*, denoted in the usual manner, which are both associative, addition is commutative, both have identity elements (the additive identity denoted 0 and the multiplicative identity denoted 1), addition has inverse elements (the inverse of x denoted $-x$), and multiplication distributes over addition. If, furthermore, multiplication is commutative, then the ring is called a *commutative ring.*

3.1.1 Definition and properties of rings

Here's a more complete definition.

Definition 3.1. A *ring R* consists of

1. a set, also denoted R and called the *underlying set* of the ring;

2. a binary operation $+ : R \times R \to R$ called *addition*, which maps an ordered pair $(x, y) \in R \times R$ to its *sum* denoted $x + y$;

3. another binary operation $\cdot : R \times R \to R$ called *multiplication*, which maps an ordered pair $(x, y) \in R \times R$ to its *product* denoted $x \cdot y$, or more simply just xy;

   such that

4. addition is commutative, that is, $\forall x, \forall y, x + y = y + x$;

5. addition is associative, that is, $\forall x, \forall y, (x + y) + z = x + (y + z)$;

6. multiplication is associative, that is, $\forall x, \forall y, (xy)z = x(yz)$;

7. there is an additive identity, an element of F denoted 0, such that $\forall x, 0 + x = x$;

8. there is a multiplicative identity, an element of F denoted 1, such that $\forall x, 1x = x$;

9. there are additive inverses, that is, $\forall x, \exists y, x + y = 0$; and

10. multiplication distributes over addition, that is, $\forall x, \forall y, \forall z, x(y + z) = xy + xz$.

When multiplication is also commutative, that is, $\forall x, \forall y, xy = yx$, the ring is called a *commutative ring*. The conditions for a ring are often call the *ring axioms.*

Subtraction, multiples, and powers. As we did with fields, we can define subtraction, integral multiples, and nonnegative integral powers. We won't have division or negative integral powers since we don't have reciprocals.

As before, we define subtraction in terms of negation. The *difference* of two elements x and y is $x - y = x + (-y)$. The expected properties of subtraction all follow from the ring axioms. For instance, multiplication distributes over subtraction.

Likewise, we can define integral multiples of elements in a ring. Define $0x$ as 0, then inductively define $(n + 1)x = x + nx$ when $n \geq 0$. Then if $-n$ is a negative integer, define $-nx$ as $-(nx)$. The usual properties of multiples, like $(m + n)x = mx + nx$ still hold.

Furthermore, we can define positive integral powers of x. Define x^1 as x for a base case, and inductively, $x^{n+1} = xx^n$. Thus nx is the product of n x's. For instance, $x^3 = xxx$.

Examples 3.2 (rings)**.** Of course, all fields are automatically rings, but what are some other rings? We've talked about some others already, including

1. the ring of integers $\mathbf{Z}$ which includes all integers (whole numbers)—positive, negative, or 0.

2. the ring of polynomials $R[x]$ with coefficients in a commutative ring R.

3. the matrix ring $M_n(R)$ of $n \times n$ matrices with entries in a commutative ring R. This example is a noncommutative ring when $n \geq 2$.

4. the ring of upper triangular matrices is a subring of $M_n(R)$.

5. the cyclic ring $\mathbf{Z}_n$, the ring of integers modulo n, where n is a particular integer.

6. the ring $\mathcal{P}(S)$ of subsets of a set S where $A + B$ is the symmetric difference and AB is the intersection of two subsets A and B.

Properties that follow from the ring axioms. There are numerous useful properties that from the axioms, but not so many as follow from the field axioms. Here's a list of several of them.

1. 0 is unique. That is, there is only one element x of a ring that has the property that $\forall y, x + y = y$. Likewise, 1 is unique.

2. Multiplication distributes over subtraction. $x(y-z)=xy-xz$ and $(y-z)x=yx-zx$.

3. $-0=0$.

4. $0x=0$.

5. $(-1)x=-x$, $(-x)y=-(xy)=x(-y)$, and $(-x)(-y)=xy$.

There are some expected properties that are not included here. I'll show why not using examples from $\mathbf{Z}_6$.

1. If the product of two elements is 0, $xy=0$, it does not follow that either $x=0$ or $y=0$. For example, in $\mathbf{Z}_6$ the product of 2 and 3 is 0.

2. Cancellation does not always work. That is, if $xy=xz$ and $x\neq 0$, it doesn't follow that $y=z$. For example, in $\mathbf{Z}_6$, $3\cdot 2=3\cdot 4$, but $2\neq 4$.

3.1.2 Products of rings

If R_1 and R_2 are two rings, you can construct their product ring R. The underlying set of R is the product $R_1\times R_2$ of the underlying sets of the two rings, and addition, subtraction, and multiplication are coordinatewise. Thus,

$$(x_1,x_2)\pm(y_1,y_2)=(x_1\pm y_1,x_2\pm y_2)\qquad\text{and}\qquad (x_1,x_2)\,(y_1,y_2)=(x_1y_1,x_2y_2).$$

The additive identity in $R_1\times R_2$ is $0=(0,0)$, and the multiplicative identity is $1=(1,1)$. Since all the operations are performed coordinatewise, the ring axioms are satisfied in $R_1\times R_2$, so it's a ring.

The projection functions $\pi_1 : R_1\times R_2\to R_1$ and $\pi_2 : R_1\times R_2\to R_2$ defined by $\pi_1(x_1,x_2)=x_1$ and $\pi_2(x_1,x_2)=x_2$ are both ring homomorphisms. They preserve addition, multiplication, and 1.

Products of more than 2 rings can be defined analogously, even products of infinitely many rings.

Although the products of rings are rings, the products of fields aren't fields, but just rings.

3.1.3 Integral domains

Much of the time we will want the cancellation property that was mentioned above to hold, so we'll give a special name to commutative rings that satisfy them. It will help if we make a couple of definitions.

Definition 3.3. A nonzero element x in a commutative ring is a *zero-divisor* if there exists a nonzero y such that $xy=0$. We'll say a commutative ring satisfies the *cancellation law* if

$$\forall x\neq 0, \forall y, \forall z,\ xy=xz \text{ implies } y=z.$$

We found in the example above that 2 and 3 are zero-divisors in $\mathbf{Z}_6$, and that $\mathbf{Z}_6$ did not satisfy the cancellation law. You can examine $\mathbf{Z}_n$ to determine which nonzero elements are zero-divisors and which have reciprocals.

There's a connection between zero-divisors and the cancellation law.

Theorem 3.4. A commutative ring satisfies the cancellation law if and only if it has no zero-divisors.

Proof. Suppose the ring satisfies the cancellation law. Let x be a nonzero element in the ring. If $xy = 0$, then $xy = x0$, so by that cancellation law, $y = 0$. Then x can't be a zero-divisor. Thus the ring has no zero-divisors.

Next suppose that the ring has no zero-divisors. We'll show it satisfies the cancellation law. If $x \neq 0$ and $xy = xz$, then $x(y - z) = 0$, and since x is not a zero divisor, therefore $y - z = 0$, so $y = z$. Thus the ring satisfies the cancellation law. Q.E.D.

Group rings You can form a ring $\mathbf{Z}G$ out of a group G as follows. Assume that G is written multiplicatively. The finite formal sums of elements of G are the elements of $\mathbf{Z}G$. Thus, if n is a nonnegative integer and $a_1, \ldots, a_n \in G$, then the formal sum $x_1a_1 + \cdots + x_na_n$ names an element of the group ring $\mathbf{Z}G$. Addition is coordinatewise. Multiplication uses the group operation.

This definition can be generalizes so that group rings have their coordinates in any commutative ring R, not just $\mathbf{Z}$. This results in a group ring RG.

Exercise 3.1. Let G be the two element cyclic group $G = \{1, a\}$ where $a^2 = 1$. A typical element of $\mathbf{Z}G$ is $x + ya$ where $x, y \in \mathbf{Z}$. Multiplication is defined by $(x_1 + y_1a)(x_2 + y_2a) = (x_1y_1 + x^2y_2) + (x_1y_2 + x_2y_1)a$. Show that the square of any nonzero element in $\mathbf{Z}G$ is not zero, but show that $\mathbf{Z}G$ does have zero-divisors by finding a pair.

Definition 3.5 (integral domain)**.** An *integral domain* is a commutative ring D in which $0 \neq 1$ that satisfies one of the two equivalent conditions: it has no zero-divisors, or it satisfies the cancellation law.

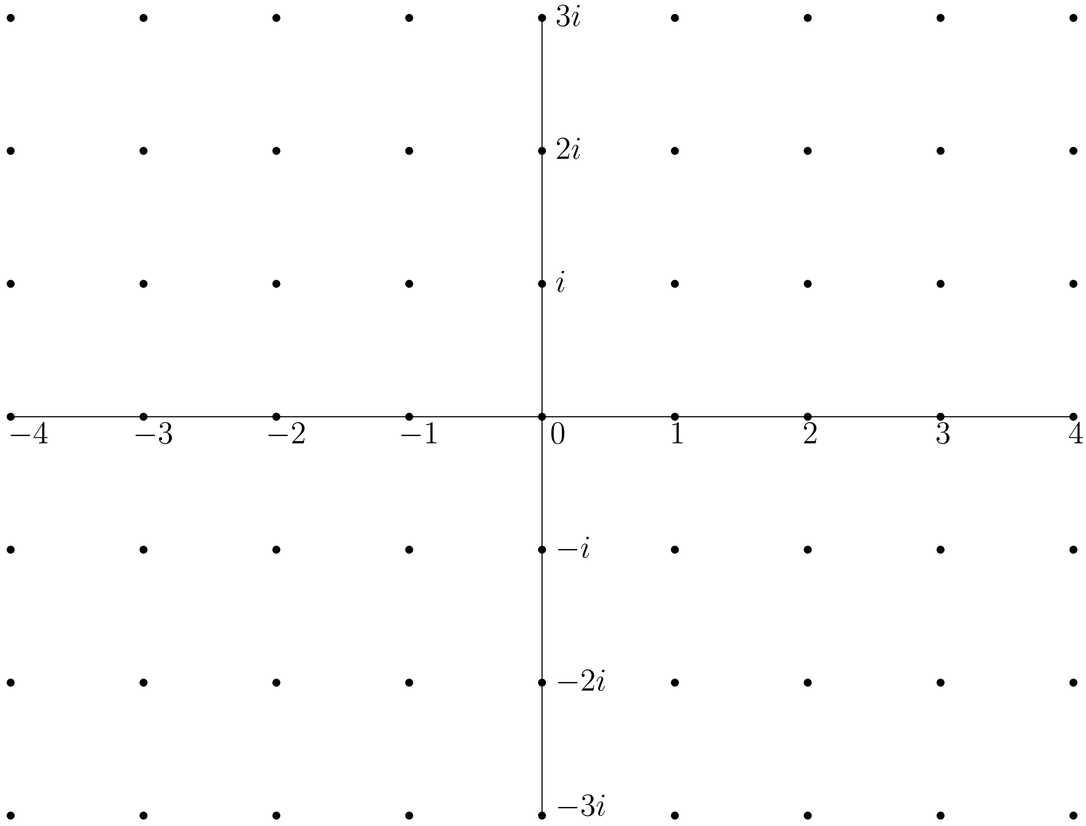
All the fields and most of the examples of commutative rings we've looked at are integral domains, but $\mathbf{Z}_n$ is not an integral domain if n is not a prime number.

Note that any subring of a field or an integral domain will an integral domain since the subring still won't have any zero-divisors.

Products of rings. Products of (nontrivial) rings are never integral domains since they always have the zero divisors $(1, 0)$ and $(0, 1)$ whose product is 0.

3.1.4 The Gaussian integers, $\mathbf{Z}[i]$

One important example of an integral domain is that of the Gaussian integers $\mathbf{Z}[i]$. Its elements are of the form $x + yi$ where $x, y \in \mathbf{Z}$, so they can be viewed as a lattice of points in the complex plane. You can check that $\mathbf{Z}[i]$ is closed under addition, subtraction, multiplication, and includes 1, so it is a subring of the field $\mathbf{C}$. Therefore, it's an integral domain.



There are four units (elements having reciprocals) in the Gaussian integers. Besides 1 and -1, i and $-i$ are also units. Note that $(1+i)(1-i)=2$, so 2 is not prime in $\mathbf{Z}[i]$ even though it is prime in $\mathbf{Z}$.

We'll come back to $\mathbf{Z}[i]$ when we study Euclidean domains.

3.1.5 Finite fields again

We won't find any examples of finite integral domains that aren't fields because there aren't any.

Theorem 3.6. If R is a finite integral domain, then R is a field.

Proof. Let x be a nonzero element of R. Consider the positive powers of x:

$$x, x^2, x^3, ..., x^n \ldots.$$

Since there are infinitely many powers, but only finitely many elements in R, therefore at least two distinct powers are equal. Let, then, $x^m = x^n$ with $m < n$. Cancel x^m from each side of the equation to conclude $x^{n-m} = 1$. Therefore, the reciprocal of x is x^{n-m-1}. Therefore, every nonzero element has an inverse. Q.E.D.

This theorem can be used to give a short proof that $\mathbf{Z}_p$ is a field when p is a prime, since it's easy to show that $\mathbf{Z}_p$ is an integral domain. We'll show it has no zero-divisors. Suppose that $xy \equiv 0 \pmod{p}$. Then $p|xy$. But if a prime divides a product, it divides one of the factors, so either $p\big|x$ or $p\big|y$, in other words, either $x \equiv 0 \pmod p$ or $y \equiv 0 \pmod p$. Thus, $\mathbf{Z}_p$ is an integral domain, and hence, by the above theorem, it's a field.

Our earlier, more complicated proof used the extended Euclidean algorithm to find an inverse for x. That's actually a much more efficient way to find the inverse than to look through the powers of x.

3.2 Factoring $\mathbf{Z}_n$ by the Chinese remainder theorem

We'll look at the structure of the cyclic ring $\mathbf{Z}_n$ when n is composite in more detail. In particular, when n is not a power of a prime number, then $\mathbf{Z}_n$ is a product of smaller cyclic rings.

3.2.1 The Chinese remainder theorem

Theorem 3.7. Suppose that $n = km$ where k and m are relatively prime. Then

$$\mathbf{Z}_n \cong \mathbf{Z}_k \times Z_m.$$

More generally, if n is the product $k_1 \cdots k_r$ where the factors are pairwise relatively prime, then

$$\mathbf{Z}_n \cong \mathbf{Z}_{k_1} \times \cdots \times \mathbf{Z}_{k_r} = \prod_{i=1}^{r} \mathbf{Z}_{k_i}.$$

In particular, if the prime factorization of n is $n = p_1^{e_1} \cdots p_r^{e_r}$. Then the cyclic ring $\mathbf{Z}_n$ factors as the product of the cyclic rings $\mathbf{Z}_{p_i^{e_i}}$, that is,

$$\mathbf{Z}_n \cong \prod_{i=1}^{r} \mathbf{Z}_{p_i^{e_i}}.$$

Proof. The third statement is a special case of the second.

The second follows from the first by induction.

In one direction, $\mathbf{Z}_n \to \mathbf{Z}_k \times \mathbf{Z}_m$, the function giving the isomorphism is fairly obvious; it's built of the two functions $\mathbf{Z}_n \to \mathbf{Z}_k$ and $\mathbf{Z}_n \to \mathbf{Z}_m$ that are easy to describe.

There is an obvious candidate for a ring function $\mathbf{Z}_n \to \mathbf{Z}_k$, namely $[x]_n \mapsto [x]_k$ by which is meant the equivalence class of x modulo n is sent to the equivalence class of x modulo k.

First, we have to check that this function is well defined. Suppose $[x]_n = [y]_n$. Then $x \equiv y \pmod n$, so $n\big|(x-y)$. But $k\big|n$, therefore $k\big|(x-y)$. Hence, $x \equiv y \pmod k$, and $[x]_k = [y]_k$. So the function is well-defined.

You can check the rest, that this function preserves the ring operation so that it's a ring homomorphism.

Putting together the two ring homomorphisms $\mathbf{Z}_n \to \mathbf{Z}_k$ and $\mathbf{Z}_n \to \mathbf{Z}_m$ we have a ring homomorphism

$$\begin{array}{rcl} \mathbf{Z}_n & \to & \mathbf{Z}_k \times \mathbf{Z}_m \\ [x]_n & \mapsto & ([x]_k, [x]_m) \end{array}$$

In order to show that this is an isomorphism, all we need to do is to show that it's a bijection, and for that, all we need to do is to show that it's an injection since the sets Z_n and $Z_k \times Z_n$ have the same cardinality.

Suppose that $[x]_n$ and $[y]_n$ are sent to the same element in $\mathbf{Z}_k \times \mathbf{Z}_m$. Then $[x]_k = [y]_k$ and $[x]_m = [y]_m$, that is, $k\big|(x-y)$ and $m\big|(x-y)$. Since they both divide $x-y$, so does their least common multiple. But they're relatively prime, so their LCM is their product, n. Thus $n\big|(x-y)$, so $[x]_n = [y]_n$. Therefore, this is a one-to-one function, hence a one-to-one correspondence. Thus, the ring homomorphism is an isomorphism. Q.E.D.

The inverse. Well, since it's a bijection, it shouldn't be too hard to find its inverse $\mathbf{Z}_k \times \mathbf{Z}_m \to \mathbf{Z}_n$. In other words, solve for x (mod n) the pair of simultaneous congruences

$$x \equiv a \pmod{k}$$
$$x \equiv b \pmod{m}$$

We can find a solution with the extended Euclidean algorithm. Since $\text{GCD}(m,k) = 1$, therefore 1 is a linear combination of m and k, that is, there are integers s and t so that $sm + tk = 1$. Multiply by $b-a$ to conclude $s(b-a)m + t(b-a)k = b-a$. Therefore, $t(b-a)k + a = b - s(b-a)m$. Let that be x. Then $x \equiv a \pmod{k}$ and $x \equiv b \pmod{m}$ as required.

Problems like this in indeterminate analysis were solved in ancient China and in ancient India. The earliest appeared in *Sunzi suanjing (Master Sun's Mathematical Manual)* in the about the fourth century C.E. in China. In 1247 Qin Jiushao gave a general method for solving linear congruences in his *Shushu jiuzhang (Mathematical Treatise in Nine Sections).*

3.2.2 Brahmagupta's solution

In India in the seventh century C.E., Brahmagupta also gave a general algorithm for solving these linear congruences in his *Brāhmasphuṭasiddhānta (Correct Astronomical System of Brahma).* If more than two congruences were given, he first reduced the problem to solving pairs of congruences as we did above. His solution is the one described above.

As an example, find x (mod 210) if

$$x \equiv 11 \pmod{45}$$
$$x \equiv 4 \pmod{56}$$

Here's how he did it in modern notation, explained with the numerical example above.

We're looking for a value of x so that $x = 45s + 11 = 56t + 4$ for some integers s and t. So we need s and t so that $45s + 7 = 56t$. That reduces to $45(s-t) + 7 = 11t$. Let $s' = s - t$. To solve $45s' + 7 = 11t$, since $45 = 4 \cdot 11 + 1$, reduce it to $s' + 7 = 11(t - 4s')$. Let $t' = t - 4s'$. We can solve $s' + 7 = 11t'$ by setting $s' = 4$ and $t' = 1$. Substituting these in the defining equations, we find $t = t' + 4s' = 17$, and $s = s' + t = 21$. Therefore, $x = 45s + 11 = 956$, the answer.

Of course, Brahmagupta did not use variables. His is solution was described as a fairly simple algorithm that just used the four arithmetic operations.

3.2.3 Qin Jiushao's solution

The algorithm that Qin Jiushao described was fairly different and applied directly to many linear congruences so long as the moduli were pairwise relatively prime. Let's illustrate it with the system of three congruences

$$\begin{aligned} x &\equiv 45 \pmod{121} \\ x &\equiv 31 \pmod{63} \\ x &\equiv 30 \pmod{100} \end{aligned}$$

Since the moduli are pairwise relatively prime, we can find a unique solution to this system modulo 762300, the product of the moduli.

Step 1. For each modulus, find a reciprocal of the product of the remaining moduli modulo the given modulus. For the first modulus, 121, that means we need the reciprocal of 6300 modulo 121, that is, we need to solve

$$6300y \equiv 1 \pmod{121}.$$

That's the same as $8y \equiv 1 \pmod{121}$. The extended Euclidean algorithm gives us $1 = (-15) \cdot 8 + 1 \cdot 121$, so $y = -15$ is a solution.

For the second modulus, 63, we need the reciprocal of 12100 modulo 63. That's the same as the reciprocal of 4 modulo 63, which is 16.

For the third modulus, 100, we need the reciprocal of 7623 modulo 100. That's the same as the reciprocal of 23 modulo 100. By the extended Euclidean algorithm, $(-13) \cdot 23 + 3 \cdot 8 = 1$, so -13 is the reciprocal of 23 modulo 100.

Step 2. To get x sum three products abc, one for each congruence, where a is the constant in the congruence, b is the product of the other moduli, and c is the reciprocal found in the previous step. That gives us

$$\begin{aligned} & 45 \cdot 6300 \cdot (-15) \\ + \; & 31 \cdot 12100 \cdot 16 \\ + \; & 30 \cdot 7623 \cdot (-13) \\ = \; & -283515 + 6001600 - 2972970 = 2745115 \end{aligned}$$

and then reduce this number modulo the product 762300 of all three moduli. That gives a final answer of $x \equiv 458215 \pmod{762300}$.

3.3 Boolean rings

> Representing by x the class "men," and by y "Asiatics," let z represent the adjective "white" to the collection of men expressed by the phrase "Men except Asiatics," is the same as to say "White men except white Asiatics." Hence we have
>
> $$z(x - y) + zx - zy.$$
>
> This is also in accordance with the laws of ordinary algebra.

> George Boole, 1854. *An Investigation of the Laws of Thought on which are founded the mathematical theories of logic and probabilities.*

George Boole (1815-1864). Boole wanted to bring logic into the realm of mathematics, which he did by algebrizing it.

We'll incorporate his investigations in our study of ring theory, but change his notation slightly. Boole did not allow a sum of two things unless they were disjoint, so $x + x$ had no meaning for him. We'll just take $+$ to be an exclusive or (symmetric difference), so $x+x$ will be 0 for us.

3.3.1 Introduction to Boolean rings

We saw before that powerset $\mathcal{P}(S)$ of a set S becomes a ring when we define $A+B$ to be the symmetric difference and AB to be the intersection of two subsets A and B. The 0 element of the ring is the emptyset $\emptyset$, while the 1 element is S. The complement of a subset A is $1 - A$ (which equals $1 + A$).

This ring has some unusual properties.

Definition 3.8. An element e of a ring is said to be *idempotent* when $e^2 = e$. If every element in a ring is idempotent, then the ring is called a *Boolean ring.*

The ring $\mathcal{P}(S)$ is evidently an example of a Boolean ring.

Notice that 0 and 1 are always idempotent in any ring. Also, even though a Boolean ring is not required to be commutative, you can fairly easily show it is commutative since every element is idempotent, so a requirement of commutativity could be included in the definition, but it's redundant. One more special property of a Boolean ring is that its characteristic is 2. So negation does nothing, $-x = x$, and subtraction is the same as addition in a Boolean ring, $x - y = x + y$.

The following table compares common notations in logic and set theory to the notation in a Boolean ring. Here, P and Q are propositions or predicates, A and B are subsets of a set Ω, and x and y are elements of a Boolean ring. These are just a few correspondences. You can add many more.

Logic	Set theory	Algebra
T (true)	Ω	1
F (false)	$\emptyset$	0
$P \wedge Q$ (and)	$A \cap B$	xy
$P \vee Q$ (inclusive or)	$A \cup B$	$x + y + xy$
$P \oplus Q$ (exclusive or)	$A \oplus B$	$x + y$
$\neg P$ (not)	$\overline{A}$	$1 + x$
$P \iff Q$	$A = B$	$x = y$
$P \implies Q$	$A \subseteq B$	$xy = x$
$T \vee Q \iff Q$	$\Omega \cap B = B$	$1y = y$
$F \wedge Q \iff Q$	$\emptyset \cup B = B$	$0 + y = y$
$F \vee Q \iff F$	$\emptyset \cap B = \emptyset$	$0y = 0$
$P \vee Q \iff Q \vee P$	$P \cap Q = Q \cap P$	$xy = yx$

3.3.2 Factoring Boolean rings

Suppose that a set S is partitioned into subsets $S_1, S_2, ..., S_n$. That means S is the union of all these subsets, and they are pairwise disjoint. Then the ring $\mathcal{P}S$ is isomorphic to a product of the rings $\mathcal{P}S_i$. The function

$$\begin{array}{rcl} \mathcal{P}(S) & \cong & \mathcal{P}(S_1) \times \mathcal{P}(S_2) \times \cdots \times \mathcal{P}S_n \\ A & \mapsto & (AS_1, AS_2, \ldots, AS_n) \end{array}$$

gives the ring homomorphism in one direction, and it's a bijection since A is the disjoint union of the terms on the right.

In fact, this works even when S is partitioned into arbitrarily many subsets. Since S is the disjoint union of its singletons $S = \cup_{x\in S}\{x\}$, therefore $\mathcal{P} = \prod_{x\in S}\mathcal{P}(\{x\})$. In other words, $\mathcal{P}$ is a power of the 2-element ring.

Factoring works in a general Boolean ring as well as those of the form $\mathcal{P}(S)$. Let R be a Boolean ring, and e any idempotent in it other than 0 or 1. Let $\overline{e} = 1 - e$, so that $1 = e + \overline{e}$ from which it follows that $x = xe + x\overline{e}$ for all $x \in R$. Let $R_e = \{xe \,|\, x \in R\}$, and let $R_{\overline{e}} = \{xe \,|\, x \in R\}$. You can check that both R_e and $R_{\overline{e}}$ are Boolean rings, where the multiplicative identities are e and $\overline{e}$, respectively. Furthermore,

$$\begin{array}{rcccc} R & \cong & R_e & \times & R_{\overline{e}} \\ x & \mapsto & (xe & , & x\overline{e}) \end{array}$$

3.3.3 A partial order for a Boolean ring

If we define $x \subseteq y$ to mean $xy = y$, then our Boolean ring will have a partial ordering. Recall that a partial ordering on a set is reflexive, antisymmetric, and transitive.

1. Reflexive: $x \subseteq x$, since $x^2 = x$.

2. Antisymmetric: $x \subseteq y$ and $y \subseteq x$ imply $x = y$, since $xy = x$ and $yx = y$ imply $x = y$.

3. Transitive: $x \subseteq y$ and $y \subseteq z$ imply $x \subseteq z$, since $xy = x$ and $yz = y$ imply $xz = x$. (*Proof:* $xz = (xy)z = x(yz) = xy = x$.)

In this partial order, the product xy is the *meet* $x \wedge y$ of x and y, that is, it's the largest element z such that $z \subseteq x$ and $z \subseteq y$. Likewise, $x + y + xy$ is the *join* $x \wedge y$ of x and y, that is, it's the smallest element z such that $x \subseteq z$ and $y \subseteq z$. A partial order that has meets and joins of pairs of elements is called a *lattice*. Not all lattices have the distributive properties

$$(x \vee y) \wedge z = (x \wedge z) \vee (y \wedge z) \text{ and } (x \wedge y) \vee z = (x \vee z) \wedge (y \vee z)$$

but Boolean rings do, so Boolean rings are examples of *distributive lattices*

An element x of a ring that is not zero and not a unit is *irreducible* if whenever $x = yz$, either y or z is a unit. In other words, it can't be factored.

The only unit in a Boolean ring is 1. In the Boolean ring $\mathcal{P}(S)$ the irreducible elements are the subsets of S that are complements of singletons $S - \{a\}$. This gives us a way to recover S from $\mathcal{P}(S)$. I'll let you prove the following theorem.

Theorem. If R is a finite Boolean ring, then $R \cong \mathcal{P}(S)$ where

$$S = \{x \in R \,|\, 1 - x \text{ is irreducible}\}.$$

3.4 The field of rational numbers and general fields of fractions

Suppose that we already have constructed the integral domain of integers $\mathbf{Z}$, but for some reason do not have the field of rational numbers $\mathbf{Q}$. Then we could construct $\mathbf{Q}$ from $\mathbf{Z}$ since each rational number can be named by a pair of integers. We'll do that. The steps we use only depend on $\mathbf{Z}$ being an integral domain. That means that the construction we use can also be used to create a *field of fractions* F from any integral domain R. In the following, think of R as $\mathbf{Z}$ and F as $\mathbf{Q}$.

An equivalence relation on pairs of integers. First of all, a rational number $\frac{m}{n}$ can be named by a pair of integers (m, n) where the second integer n does not equal 0. But different pairs (m, n) and (k, l) can name the same integer $\frac{m}{n} = \frac{k}{l}$ if $ml = nk$. That suggests if we want to create rational numbers from integers, we'll need an equivalence relation.

We'll start with the set $R \times R_{\neq 0}$ of ordered pairs (m, n) of elements of an integral domain R with $n \neq 0$. Define a relation $\equiv$ on this set by

$$(m, n) \equiv (k, l) \quad \text{iff} \quad ml = nk.$$

You can easily verify that this relation is an equivalence relation.

> Reflexivity: $(m, n) \equiv (m, n)$. That's valid since $mn = mn$.
>
> Symmetry: $(m, n) \equiv (k, l)$ implies $(k, l) \equiv (m, n)$. That's valid since $ml = nk$ implies $kn = lm$.
>
> Transitivity: $(m, n) \equiv (k, l)$ and $(k, l) \equiv (s, t)$ imply $(m, n) \equiv (s, t)$. We need to show that $ml = nk$ and $kt = ls$ imply $mt = ns$. Multiply the first equation by t and the second by n. Then $mlt = nkt$ and $nkt = nls$, so $mlt = nls$. But R is an integral domain, so cancellation is valid when $l \neq 0$, so $mt = ns$.

Thus, $\equiv$ is an equivalence relation on $R \times R_{\neq 0}$. Let F be the quotient set $F_{\equiv}$, and denote an element $[(m, n)]$ of F by $\frac{m}{n}$.

So far, we've got the underlying set for our proposed field F, but we don't have the operations for a field. Before we define them (and show they're well-defined), let's verify that the function $R \to R \times R_{\neq 0} \to F$ which sends an element m of R first to $(m, 1)$ then to $\frac{m}{1}$ is a one-to-one function. Suppose that $\frac{m}{1} = \frac{n}{1}$. That means $m1 = 1n$, so $m = n$. Thus we may interpret $R \to F$ as making R a subset of F by identifying m with $\frac{m}{1}$.

Addition on F. We'd like to define the sum

$$\frac{m}{n} + \frac{k}{l} \quad \text{as} \quad \frac{ml + nk}{nl},$$

but as our fractions are really equivalence classes, we need to show that's well defined. In detail, we need to show that

$$\frac{m}{n} = \frac{m'}{n'} \quad \text{and} \quad \frac{k}{l} = \frac{k'}{l'} \quad \text{imply} \quad \frac{ml + nk}{nl} = \frac{m'l' + n'k'}{n'l'}.$$

That reduces to showing that

$$mn' = nm' \qquad \text{and} \qquad kl' = lk' \qquad \text{imply} \qquad (ml + nk)n'l' = nl(m'l' + n'k').$$

But that can be shown by multiplying the first equation by ll', the second by nn' and adding the two resulting equations. Thus, this addition on F is well-defined.

Multiplication on F. We'd like to define the product

$$\frac{m}{n}\frac{k}{l} \qquad \text{as} \qquad \frac{mk}{nl},$$

We need to show that's well defined. You'll find that the proof is easier than the one above for addition.

Next, we need to verify that with these definitions F satisfies the field axioms.

Commutativity of addition. $\dfrac{m}{n} + \dfrac{k}{l} = \dfrac{k}{l} + \dfrac{m}{n}$. That's easily verified since $\dfrac{ml + nk}{nl} = \dfrac{kn + lm}{ln}$. (That depends on commutativity of addition and multiplication in R.)

Commutativity of multiplication. $\dfrac{m}{n}\dfrac{k}{l} = \dfrac{k}{l}\dfrac{m}{n}$. That's easily verified since $\dfrac{mk}{nl} = \dfrac{km}{ln}$.

Associativity of addition. You can easily show it, but it's a big mess.

Associativity of multiplication. Pretty easy.

Additive identity. $\dfrac{0}{1} + \dfrac{k}{l} = \dfrac{k}{l}$. Easy.

Multiplicative identity $\dfrac{1}{1}\dfrac{k}{l} = \dfrac{k}{l}$. Easy.

Negation. $\dfrac{m}{n} + \dfrac{-m}{n} = \dfrac{0}{1}$. Pretty easy.

Reciprocation. For $\dfrac{m}{n} \neq \dfrac{0}{1}$, $\dfrac{m}{n}\dfrac{n}{m} = \dfrac{1}{1}$. Pretty easy.

Multiplication distributes over addition. Easy but messy.

$0 \neq 1$. We need to show that $\dfrac{0}{1} \neq \dfrac{1}{1}$ in F. But that's the same as $0 \cdot 1 \neq 1 \cdot 1$ in the integral domain R, and part of the definition of integral domain requires $0 \neq 1$.

Thus, F is a field.

We'll summarize this result as a theorem.

Theorem 3.9. An integral domain R is a subring of a field F, called the *field of fractions*, where each element of F can be represented as $\frac{m}{n}$ where m and n are elements of R and $n \neq 0$.

Corollary 3.10. An integral domain either has characteristic 0 or prime characteristic p.

Proof. It has the same characteristic as its field of fractions which is either 0 or a prime number. Q.E.D.

Examples 3.11. The primary example of this is the construction of $\mathbf{Q}$ from $\mathbf{Z}$.

For another example, take the Gaussian integers $\mathbf{Z}[i]$ for the integral domain R. Then the field of fractions F is the field $\mathbf{Q}(i)$.

Yet for another example, take the polynomial ring $F[x]$ with coefficients in a field F. It's an integral domain, and its field of fractions is the rational function field $F(x)$ with coefficients in F.

Stopping short of inverting all elements. Sometimes you may want to create reciprocals for some elements of an integral domain, but not for all elements. This can be done by a minor modification of the above process. Suppose, for instance, that you want to extend $\mathbf{Z}$ to include the reciprocal of 2 but not of any other prime number. That would lead to the *domain of dyadic rationals* $\mathbf{Z}[\frac{1}{2}]$ where the denominators are powers of 2.

On the other hand, if you want to extend $\mathbf{Z}$ to include the reciprocals of all the primes except 2, just include odd denominators. This is called localizing $\mathbf{Z}$ at 2.

These other constructions are useful, but we won't use them ourselves.

3.5 Categories and the category of rings

Categories are higher order algebraic structures. We'll look at the category of rings in which the objects of the category are all the rings. The purpose of a category is to study the interrelations of its objects, and to do that the category includes morphisms between the objects. In the case of the category of rings, the morphisms are the ring homomorphisms.

We'll start with the formal definition of categories. We'll use the category of rings both to illustrate categorical concepts and to study rings. Category theory was developed by Eilenberg and Mac Lane in the 1940s.

3.5.1 The formal definition of categories

Unlike fields, rings, and groups, we won't require that categories build on sets. In a category the collection of all its objects won't be a set because the collection is larger than any set. That's not a problem since theories don't have to be built on set theory. Indeed, set theory itself is not built on set theory.

Definition 3.12. A *category* $\mathcal{C}$ consists of

1. *objects* often denoted with uppercase letters, and

2. *morphisms* (also called *maps* or *arrows*) often denoted with lowercase letters.

3. Each morphism f has a *domain* which is an object and a codomain which is also an object. If the domain of f is A and the codomain is B, then we write $f : A \to B$ or $A \xrightarrow{f} B$. The set of all morphisms from A to B (if it is a set) is denoted $\operatorname{Hom}(A, B)$.

4. For each object A there is a morphism $1_A : A \to A$ called the *identity morphism* on A.

5. Given two morphisms $A \xrightarrow{f} B$ and $B \xrightarrow{g} C$ where the codomain of one is the same as the domain of the other there is another morphism $A \xrightarrow{g\circ f} C$ called the *composition* of the two morphisms. This composition is illustrated by the commutative diagram

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[dr, "g \circ f"'] & B \arrow[d, "g"] \\
 & C
\end{tikzcd}$$

6. for all $A \xrightarrow{f} B$, $f \circ 1_A = f$ and $1_B \circ f = f$. These compositions are illustrated by the two commutative diagrams

$$\begin{tikzcd}
A \arrow[d, "1_A"'] \arrow[dr, "f \circ 1_A"] & \\
A \arrow[r, "f"'] & B
\end{tikzcd}
\qquad
\begin{tikzcd}
A \arrow[r, "f"] \arrow[dr, "1_B \circ f"'] & B \arrow[d, "1_B"] \\
 & B
\end{tikzcd}$$

7. for all $A \xrightarrow{f} B$, $B \xrightarrow{g} C$, and $C \xrightarrow{h} D$, $(h \circ g) \circ f = h \circ (g \circ f)$. In the diagram below, if the two triangles in the diagram each commute, then the parallelogram commutes.

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[dr, "g \circ f"'] & B \arrow[d, "g"] \arrow[dr, "h \circ g"] & \\
 & C \arrow[r, "h"'] & D
\end{tikzcd}$$

A diagram of objects and morphisms in a category is said to *commute*, or be a *commutative diagram* if any two paths of morphisms (in the direction of the arrows) between any two objects yield equal compositions.

Isomorphisms in a category $\mathcal{C}$. Although only morphisms are defined in a category, it's easy to determine which ones are isomorphisms. A morphism $f : A \to B$ is an *isomorphism* if there exists another morphism $g : B \to A$, called its *inverse*, such that $f \circ g = 1_A$ and $g \circ f = 1_B$.

Examples 3.13 (The categories of sets, groups, rings, and fields)**.** Although we're more interested in the category of rings right now, the category $\mathcal{S}$ of sets is also relevant. An object in $\mathcal{S}$ is a set, and a morphism in $\mathcal{S}$ is a function. The domain and codomain of a morphism are just the domain and codomain of the function, and composition is composition. Isomorphisms are bijections.

The objects of the category $\mathcal{G}$ of groups are groups, and the morphisms of $\mathcal{G}$ are group homomorphisms.

The objects of the category $\mathcal{R}$ of rings are rings, and the morphisms of $\mathcal{G}$ are ring homomorphisms.

The objects of the category of fields are fields, and its morphisms are field homomorphisms, which are just ring homomorphisms. The category of fields is a subcategory of the category of rings.

3.5.2 The category $\mathcal{R}$ of rings

Recall that a ring homomorphism $f : A \to B$ between rings is a function that preserves addition, multiplication, and 1. The category of rings has as its objects all rings and as its morphisms all ring homomorphisms. The identity morphism 1_A on a ring is the identity homomorphism, and composition is the usual composition of homomorphisms. Thus, we have a category $\mathcal{R}$ of rings.

If this were all there was to category theory, there wouldn't be much point to it. But by emphasizing the morphisms and deemphasizing elements in rings we can identify what's important about certain rings and certain ring constructions. We'll look at products of rings first to see what characterizes them. We'll also look at a couple of special rings, namely $\mathbf{Z}$ and $\mathbf{Z}[x]$, for characterizing properties of them. We'll also see how to characterize monomorphisms.

The universal property of products. Recall that the product $R_1 \times R_2$ of two rings is consists of ordered pairs (x_1, x_2) with $x_1 \in R_1$ and $x_2 \in R_2$, and the ring operations for $R_1 \times R_2$ are performed coordinatewise. Furthermore, we have the projection ring homomorphisms $R_1 \times R_2 \xrightarrow{\pi_1} R_1$ and $R_1 \times R_2 \xrightarrow{\pi_2} R_2$ which pick out the two coordinates.

This product has the universal property that for each ring S and ring homomorphisms $S \xrightarrow{f_1} R_1$ and $S \xrightarrow{f_2} R_1$, there exists a unique ring homomorphism $S \to R_1 \times R_2$, which we will denote (f_1, f_2) such that $f_1 = \pi_1 \circ (f_1, f_2)$ and $f_2 = \pi_2 \circ (f_1, f_2)$, as illustrated by the diagram below.

$$\begin{tikzcd}
 & & R_1 \\
S \arrow[urr, "f_1"] \arrow[r, "(f_1,f_2)"] \arrow[drr, "f_2"'] & R_1 \times R_2 \arrow[ur, "\pi_1"'] \arrow[dr, "\pi_2"] & \\
 & & R_2
\end{tikzcd}$$

In fact, the product is characterized by this universal property in the sense that if another ring R has this universal property, then there is a ring isomorphism $R \to R_1 \times R_2$. In more detail, if $R \xrightarrow{p_1} R_1$ and $R \xrightarrow{p_2} R_2$ have this product property (namely, that for each ring S and ring homomorphisms $S \xrightarrow{f_1} R_1$ and $S \xrightarrow{f_2} R_1$, there exists a unique ring homomorphism $S \xrightarrow{f} R$ such that $f_1 = p_1 \circ f$ and $f_2 = p_2 \circ f$), then there exists a unique ring isomorphism $R \xrightarrow{h} R_1 \times R_2$ such that $\pi_1 \circ h = p_1$ and $\pi_2 \circ h = p_2$.

Although this characterization of products was described for the category of rings, it is the definition for the product of two objects in any category. A product $R_1 \times R2$ is characterized by the property that a morphism to the product correspond to a pair of morphisms to the factors. The product of two sets in the category $\mathcal{S}$ of sets has this same universal property as does the product of two groups in the category $\mathcal{G}$ of groups. There are, however, no products in the category of fields.

Z is the initial object in the category of rings. We can also use category theory to pin down what's so special about the ring $\mathbf{Z}$. It has the property that given any ring R, there is a unique ring homomorphism $\mathbf{Z} \xrightarrow{f} R$, and it's defined by $f(n) = n$. An object in a category with that property is called the *initial object* in the category. Any two initial objects in a category are isomorphic.

The universal property of the polynomial ring $\mathbf{Z}[x]$. Given any ring R and any element $x \in R$, there is a unique ring homomorphism $\mathbf{Z}[x] \to R$ that maps x to a. This homomorphism is just evaluation at a, and a polynomial $f(x)$ is mapped to the element $f(a)$ in R.

3.5.3 Monomorphisms and epimorphisms in a category

Although we defined a monomorphism $f : A \to B$ as a one-to-one homomorphism, we can characterize monomorphisms entirely in terms of category theory.

Definition 3.14. A morphism $f : A \to B$ is *monic*, or a *monomorphism*, when if g and h are any two morphisms from any another object C to A such that $f \circ g = f \circ h$, then $g = h$.

$$\begin{tikzcd} C \arrow[r, shift left, "g"] \arrow[r, shift right, "h"'] & A \arrow[r, "f"] & B \end{tikzcd}$$

A monomorphism in the category S of sets is an injection.

This definition agrees with our previous definition for ring monomorphism in terms of elements, and one way to see the correspondence is to let C be $\mathbf{Z}[x]$.

Epimorphisms. The concept of epimorphism is dual to that of monomorphism. If we change the direction of all the arrows in the definition of monomorphism, we'll get the definition of epimorphism.

Definition 3.15. A morphism $f : A \to B$ is *epic*, or an *epimorphism*, when if g and h are any two morphisms from B to any another object C such that $g \circ f = h \circ f$, then $g = h$.

$$\begin{tikzcd} A \arrow[r, "f"] & B \arrow[r, shift left, "g"] \arrow[r, shift right, "h"'] & C \end{tikzcd}$$

In the category S of sets, an epimorphism is a surjection.

In the category R of rings, it's easy enough to show that if f is a surjective ring homomorphism, then f is an epimorphism, but there are more epimorphisms.

Example 3.16. Consider the inclusion function $\iota : \mathbf{Z} \to \mathbf{Q}$. We'll show that it's an epimorphism.

Let g and h be any two morphisms from $\mathbf{Q}$ to any another ring C such that $g \circ \iota = h \circ \iota$. Then $g(n) = h(n)$ for any integer n. Let $\frac{m}{n}$ be a rational number. Then $g(m) = h(m)$ and $g(n) = h(n)$. So,

$$g(\tfrac{m}{n})g(n) = g(\tfrac{m}{n}n) = g(m) = h(m) = h(\tfrac{m}{n}n) = h(\tfrac{m}{n})h(n) = h(\tfrac{m}{n})g(n).$$

Cancel the $g(n)$ at the ends of the continued equation to conclude $g(\frac{m}{n}) = h(\frac{m}{n})$. Thus, $g = h$.

Therefore, the ring homomorphism $\iota : \mathbf{Z} \to \mathbf{Q}$ is an epimorphism in $\mathcal{R}$, the category of rings. It is also a monomorphism. But it is not an isomorphism.

In many categories, if a morphism is both monic and epic, then it's also an isomorphism. That's true in $\mathcal{S}$ and in the category $\mathcal{G}$ of groups, but not for $\mathcal{R}$. This example shows that $\mathcal{R}$ is a somewhat unusual category.

3.6 Kernels, ideals, and quotient rings

These three concepts are closely related. For a ring homomorphism $f : R \to S$, the inverse image of 0 is a subset of R called the kernel of f and denoted $\operatorname{Ker} f$. It can't be just any subset, as we'll see, since it's closed under addition and multiplication by elements of R. A subset with those properties we'll call an ideal of R. Every ideal I of R is the kernel of some ring homomorphism $f : R \to S$. We'll use an ideal I of a ring R to define a quotient ring R/I and a projection $\gamma : R \to R/I$. These projections will be generalizations of the projections $\mathbf{Z} \to \mathbf{Z}_n$ that we studied earlier.

3.6.1 Kernels of ring homomorphisms

Definition 3.17. Let $f : R \to S$ be a ring homomorphism. Those elements of R that are sent to 0 in S form the *kernel* of f.

$$\operatorname{Ker} f = f^{-1}(0) = \{x \in R \,|\, f(x) = 0\}.$$

We'll look at properties of this kernel and what it tells us about the function f.

Example 3.18. It's a good idea to have in mind an example or two whenever a new concept is defined. The definition of the kernel of a ring homomorphism is given above, and a good example for it is the ring homomorphism $f : \mathbf{Z} \to \mathbf{Z}_n$ where n is a fixed integer. That's an especially good example we can use it throughout this discussion of rings, ideals, and quotient rings.

For that $f : \mathbf{Z} \to \mathbf{Z}_n$, an element $x \in \mathbf{Z}$ is in $\operatorname{Ker} f$ if it is sent to $[0]_n$, the 0 element in the ring $\mathbf{Z}_n$, that is, if $[x]_n = [0]_n$, or, more simply, if $n \big| x$. Therefore, the kernel of f consists of the multiples of n. A standard notation for the multiples of an integer n is $n\mathbf{Z}$. Thus, $\operatorname{Ker} f = n\mathbf{Z}$.

Kernels aren't just any subsets of R; they have some special properties. We have, of course, $0 \in \operatorname{Ker} f$, since $f(0) = 0$, Also, if x and y are both in $\operatorname{Ker} f$, then $f(x+y) = f(x) + f(y) = 0 + 0 = 0$, so their sum $x + y$ is also in $\operatorname{Ker} f$. Furthermore, if $x \in \operatorname{Ker} f$ and y is any element of R, then $f(xy) = f(x)f(y) = 0f(y) = 0$, so $xy \in \operatorname{Ker} f$, and likewise $yx \in \operatorname{Ker} f$.

Besides telling us what elements are sent to 0 by f, the kernel of f also tells us when two elements are sent to the same element. Since $f(x) = f(y)$ if and only if $f(x - y) = 0$, therefore, f will send x and y to the same element of S if and only if $x - y \in \operatorname{Ker} f$.

3.6.2 Ideals of a ring

The properties of kernels of homomorphisms that we just found we'll use to define ideals of rings. Historically, ideals had a different purpose, but we'll get to that purpose later. The word "ideal" is short for ideal number or ideal element.

Definition 3.19. An *ideal* I of a ring R is a subset that (1) includes 0, (2) is closed under addition, and (3) is closed under multiplication by elements of R. We can summarize these requirements symbolically by $0 \in I$, $I + I \subseteq I$, $RI \subseteq I$, and $IR \subseteq I$.

Both of the last two requirements, $RI \subseteq I$ and $IR \subseteq I$. are needed when R is a noncommutative ring. Most of the time we'll be dealing with commutative rings so one will do.

Note that $\{0\}$ is always an ideal in a ring R. We'll usually just denote it 0. Also, the entire ring R is an ideal, but not a proper ideal. (A *proper ideal* is any ideal $I \neq R$.)

Principal ideals and ideals generated by a set. The simplest examples of ideals are what are called principal ideals. Let a be an element of a commutative ring R. The set of all multiples of a,

$$(a) = \{xa \,|\, x \in R\},$$

is an ideal of R, as you can easily check. These ideals are called *principal ideals* because they are generated by one element. An alternate notation for the principal ideal generated by the element a is Ra or aR.

Note that (0), the ideal generated by 0, is just the 0 ideal, while (1), the ideal generated by 1, is all of R.

Sometimes it takes more than one element to generate an ideal. Let A be a subset of a commutative ring R. The smallest ideal that contains A is called the *ideal generated by A*. It must contain all linear combinations of elements of A since an ideal is closed under addition and closed under multiplication by elements of R, but that's enough. Usually, we're only interested in generating an ideal from a finite number of elements $A = \{a_1, a_2, \ldots, a_k\}$. Then the ideal generated by A is

$$(a_1, a_2, \ldots, a_k) = \{x_1 a_1 + \cdots + x_k a_k \mid \text{ each } x_i \in R\}.$$

An example of an ideal generated by two elements but not principal (not by one element) is $(5, x^2)$ in $\mathbf{Z}[k]$, the polynomial ring with integral coefficients.

3.6.3 Quotients rings, R/I

As mentioned above the kernel of a ring homomorphism f tells us when two elements are sent to the same element: $f(x) = f(y)$ if and only if $x - y \in \operatorname{Ker} f$. We can use $\operatorname{Ker} f$ to construct a "quotient ring" $R/\operatorname{Ker} f$ by identifying two elements x and y in R if their difference lies in $\operatorname{Ker} f$. In fact, we can do this not just for kernels of homomorphisms, but for any ideal I. That is, we can use an ideal I of R to determine when two elements x and y are to be identified, $x \equiv y$, and we'll end up with a ring R/I. The identification is called a congruence. This concept of congruence generalizes congruence modulo n on $\mathbf{Z}$.

Definition 3.20. A *congruence* $\equiv$ on a ring R is an equivalence relation such that for all $x, x', y, y' \in R$,

$$x \equiv x' \text{ and } y \equiv y' \text{ imply } x + y \equiv x' + y' \text{ and } xy \equiv x'y'.$$

The equivalence classes for a congruence are called *congruence classes*.

Theorem 3.21. If $\equiv$ is a congruence on a ring R, then the quotient set $R/_\equiv$, that is, the set of congruence classes, is a ring where addition is defined by $[x] + [y] = [x + y]$ and multiplication by $[x][y] = [xy]$.

Proof. First we need to show that the proposed definitions are actually well defined. That is, if a different representative x' is chosen from the congruence class $[x]$ and y' from $[y]$, then the same classes $[x' + y']$ and $[x'y']$ result. That is

$$[x] = [x'] \text{ and } [y] = [y'] \text{ imply } [x + y] = [x' + y'] \text{ and } [xy = xy'].$$

That's the same as the requirements met in the definition of congruence (which explains why they are in the definition).

Also, each of the axioms for a ring need to be verified, but they're all automatic. Here's commutativity of addition, for example.

$$[x] + [y] = [x + y] = [y + x] = [y] + [x].$$

We could say that the quotient ring inherits the properties from the ring. Q.E.D.

In the next theorem we'll see that an ideal I determines a congruence. We'll write the congruence $x \equiv y \pmod{I}$ rather than just $x \equiv y$ when we want to emphasize the role of I. The congruence classes may be written $[x]$ or $[x]_I$, or $x + I$. The last notation is a good one since $[x] = \{x + y \mid y \in I\}$.

Theorem 3.22 (Congruence modulo an ideal)**.** Let I be an ideal of a ring R. A congruence, called *congruence modulo I*, is defined by

$$x \equiv y \pmod{I} \text{ if and only if } x - y \in I.$$

The quotient ring, $R/_\equiv$, is denoted R/I.

Proof. First, we need to show that it's an equivalence relation.

Reflexivity. $x \equiv x \pmod{I}$. That's okay since $x - x = 0 \in I$.

Symmetry. $x \equiv y \pmod{I}$ implies $y \equiv x \pmod{I}$. That's okay because if $x - y \in I$, then $y - x = -(x - y) \in I$.

Transitivity. $x \equiv y \pmod{I}$ and $y \equiv z \pmod{I}$ imply $x \equiv z \pmod{I}$. That's okay, too. If $x - y \in I$ and $y - z \in I$, then so is their sum $x - z \in I$.

Thus, it's an equivalence relation. Next to show that

$$x \equiv x' \pmod{I} \text{ and } y \equiv y' \pmod{I} \text{ imply } x + y \equiv x' + y' \pmod{I} \text{ and } xy \equiv x'y' \pmod{I}.$$

That requirement reduces to the statement

$$x - x' \in I \text{ and } y - y' \in I \text{ imply } (x + y) - (x' + y') \in I \text{ and } (xy - x'y') \in I,$$

which, you can check, follow from the definition of ideal. Q.E.D.

Example 3.23 (Cyclic rings)**.** As we saw above, $I = n\mathbf{Z}$ is an ideal of $\mathbf{Z}$. The congruence defined here is the same one we had before. Thus, $x \equiv y \pmod{I}$ means $x \equiv y \pmod{n}$. The quotient ring is $\mathbf{Z}/n\mathbf{Z}$, which we have studied before and denoted $\mathbf{Z}_n$ for short.

Comment 3.24. The ring structure on the quotient R/I was defined from the ring structure on R, the projection $\gamma : R \to R/I$ is a ring homomorphism. This ring R/I is called a *quotient ring* of R. (It is also sometimes called a factor ring, but that term should be restricted to the case when R factors as a product of rings, one of which is R/I. An example of that is the Chinese remainder theorem.)

Examples 3.25 (Quadratic field extensions.)**.** We've looked at $\mathbf{Q}(\sqrt{2})$, $\mathbf{C} = \mathbf{R}(i)$, and other quadratic field extensions. We can interpret them as quotient rings.

Let's take $\mathbf{Z}(\sqrt{2})$ first. Consider the ring $R = \mathbf{Q}[x]$ of polynomials with rational coefficients. An ideal in R is the principal ideal $I = (x^2 - 2)$ generated by the polynomial $x^2 - 2$. In the quotient ring $R/I = \mathbf{Q}[x]/(x^2 - 2)$, we have $x^2 - 2 \equiv 0 \pmod{I}$, that is, $x^2 \equiv 2 \pmod{I}$, so in R/I, we find that 2 does have a square root, namely x. Since in R/I every polynomial $a_n x^n + \cdots + a_1 x + a_0$ is congruent to a polynomial of degree 1 (because $x^2 \equiv 2$), but no two linear polynomials are congruent mod I (because $a_1 x + a_0 \equiv b_1 x + b_0 \pmod{I}$ implies $(a_1 - b_1)x + (a_0 - b_0) \in I$ so $a_1 = b_1$ and $a_0 = b_0$), therefore every element in R/I is uniquely represented as a linear polynomial $a_1 x + a_0$. If we denote x by the symbol $\sqrt{2}$, then we find $\mathbf{Q}[x]/(x^2 - 2)$ is the same field as $\mathbf{Q}(\sqrt{2})$ that we described before.

Likewise, $\mathbf{R}[x]/(x^2 + 1)$ is $\mathbf{C}$.

We'll find this construction of new rings as quotient rings is very useful, especially when we take quotients rings of polynomial rings like we did here.

The image of a ring homomorphism is isomorphic to the ring modulo its kernel. Let $f : R \to S$ be a ring homomorphism. The image of f, denoted $f(R)$, is the set

$$f(R) = \{f(x) \in S \,|\, x \in R\}.$$

It is a subring of S, as you can easily verify. You can also show the following isomorphism theorem.

Theorem 3.26. If $f : R \to S$ is a ring homomorphism then the quotient ring $R/\operatorname{Ker} f$ is isomorphic to the image ring $f(R)$, the isomorphism being given by

$$\begin{array}{rcl} R/\operatorname{Ker} f & \to & f(R) \\ x + \operatorname{Ker} f & \mapsto & f(x) \end{array}$$

This gives us two ways to look at the image, either as a quotient ring of the domain R or as a subring of the codomain S.

Furthermore, we can now treat a ring homomorphism $f : R \to S$ as a composition of three ring homomorphisms. The first is the projection from R onto its quotient ring $R/\operatorname{Ker} f$, the second is the isomorphism $R/\operatorname{Ker} f \cong f(R)$, and the third is the inclusion of the image $f(R)$ as a subring of S.

3.6.4 Prime and maximal ideals

Sometimes it occurs that R/I is not just a ring, but either an integral domain or even a field. Those results occur when the ideal I is a prime ideal or a maximal ideal, respectively, as we'll define now.

Definition 3.27. An ideal I in a commutative ring R is said to be a *prime ideal* if R/I is an integral domain. Equivalently, I is a prime ideal if (1) $I \neq R$, and (2) $\forall x, y \in R$, if $xy \in I$, then either $x \in I$ or $y \in I$. An ideal I is said to be *maximal* it's a proper ideal, but it is not contained in any larger proper ideal.

Example 3.28. The ideals of $\mathbf{Z}$ that are prime are those of the form $p\mathbf{Z}$ where p is a prime number, and the 0 ideal. In fact, $p\mathbf{Z}$ are maximal ideals, but 0 is not maximal.

In a field F there is only one proper ideal, namely 0.

In an integral domain, the 0 ideal is a prime ideal, and conversely, if 0 is an ideal in a commutative ring, then the ring is an integral domain.

Theorem 3.29. Every maximal ideal is prime.

Proof. Let I be a maximal ideal of a commutative ring R, and let $xy \in I$. Suppose $x \notin I$. Then $xR + I = \{xu + v \,|\, u \in R, v \in I\}$ is an ideal containing I. Since I is an maximal ideal, therefore $xR + I$ is not a proper ideal but all of R. Therefore $1 = xu + v$ for some $u \in R$, $v \in I$. Hence $y = yxu + yv \in Iu + I = I$. Thus, I satisfies the conditions to be a prime ideal. Q.E.D.

We won't show it right now, but we'll prove later Krull's theorem which says that every ideal is containted in a maximal ideal. We'll need to discuss the axiom of choice and Zorn's lemma before we can prove it.

Theorem 3.30. Let I be an ideal of a commutative ring R. Then I is a maximal ideal if and only if R/I is a field.

Proof. We'll use the notation $[x]$ for xI to stress that we're thinking of it as an element of R/I.

Suppose that I is a maximal ideal, and let $[x]$ be any nonzero element of R/I, that is $x \notin I$. As in the last proof, $xR + I = R$. Therefore $1 = xu + v$ for some $u \in R$, $v \in I$. Then, in R/I we have $[1] = [x][u] + [v] = [x][u] + [0] = [x][u]$. Therefore $[x]$ has a reciprocal, and R/I is a field.

Now suppose that R/I is a field. Let $x \notin I$. We'll show that $xR + I = R$ which will show that I is a maximal ideal. In R/I, $[x] \neq [0]$, so $[x]$ has an inverse $[y]$, $[x][y] = [1]$, so $1 - xy \in I$, so $1 \in xR + I$, hence $R = xR + I$. Q.E.D.

3.7 Krull's theorem, Zorn's Lemma, and the Axiom of Choice

We'd like to prove Krull's theorem that every ideal in a commutative ring is contained in a maximal ideal, but in order to do that in general we'll need something called Zorn's lemma. It's a statement that's logically equivalent to the better known axiom of choice.

3.7.1 Axiom of choice

In some sense, any theorem that relies on the axiom of choice is flawed since the axiom of choice is not constructive. So, for instance, after proving an ideal is a subideal of a maximal ideal, we won't have any way to identify that maximal ideal.

The axiom of choice. This is an axiom of set theory. We haven't discussed set theory, and we really don't have to now, but there are many axioms of set theory, most of which are fairly obvious and uncontroversial.

The axiom says that given any set S, there exists a "choice function" γ which chooses from any nonempty set $T \subseteq S$ an element $\gamma(T) \in T$.

Here's a simple theorem that relies on the axiom of choice.

Theorem 3.31. Let $f : A \to B$ be a surjective function between sets A and B. Then there exists $g : B \to A$ such that $f \circ g$ is the identity function on B.

Proof. Let γ be a choice function for A. Then g is the function

$$g(y) = \gamma(f^{-1}(y)) = \gamma(\{x \mid f(x) = y\}).$$

Note that $f^{-1}(y)$ is not the empty set since f is surjective. Q.E.D.

3.7.2 Zorn's lemma

Although the axiom of choice is easy to state, it's not usually easy to use. Zorn's lemma, which is logically equivalent to the axiom of choice is hard to state, but easy to use. This lemma is applied to a nonempty collection $\mathcal{M}$ of subsets of a set S. A *chain* in $\mathcal{M}$ is a subset $\mathcal{C}$ of $\mathcal{M}$ such that $\forall A, B \in \mathcal{C}$, either $A \subseteq B$ or $B \subseteq A$. An *upper bound* of $\mathcal{C}$ is a subset B of S such that $\forall A \in \mathcal{C}, A \subseteq B$. A *maximal element* of $\mathcal{M}$ is a subset $B \in \mathcal{M}$ such that $\forall A \in M, B \not\subseteq A$.

Zorn's lemma. If every chain in $\mathcal{M}$ has an upper bound in $\mathcal{M}$, then $\mathcal{M}$ has a maximal element.

We won't prove that the Axiom of Choice is equivalent to Zorn's lemma because it would take too long, but let's see how we can use it.

Theorem 3.32 (Krull)**.** Let I be a proper ideal of a commutative ring R. Then there is a maximal ideal J such that $I \subseteq J$.

Proof. Consider the collection $\mathcal{M}$ of proper ideals of R that contain I. Note that $\mathcal{M}$ is nonempty since $I \in \mathcal{M}$. We'll show that every chain $\mathcal{C}$ in $\mathcal{M}$ has an upper bound in $\mathcal{M}$. Let $B = \bigcup_{A \in \mathcal{C}} A$. Certainly B is an upper bound for $\mathcal{C}$ since B is just the union of elements of $\mathcal{C}$.

We still have to show B is an ideal, which requires $RB \subseteq B$ and $B + B \subseteq B$ For the first, $RB = R \bigcup_{A \in \mathcal{C}} A = \bigcup_{A \in \mathcal{C}} RA = \bigcup_{A \in \mathcal{C}} A = B$. Now let $x, y \in B$. Then $x \in A_1$ for some $A_1 \in \mathcal{C}$ and $y \in A_2$ for some $A_2 \in \mathcal{C}$. But $\mathcal{C}$ is a chain, so either $A_1 \subseteq A_2$ or $A_2 \subseteq A_1$. In the first case, $x, y \in A_2$, so $x + y \in A_2 \subseteq B$, and in the second $x, y \in A_1$, so $x + y \in A_1 \subseteq B$. Thus, $B + B \subseteq B$.

Now we can apply Zorn's lemma. It implies $\mathcal{M}$ has a maximal element J. Clearly, $I \subseteq J$, and J is a proper ideal of R, but there are no larger proper ideals of R that contain J, so J is a maximal ideal. Q.E.D.

Note how we have not actually found J. There may be many different maximal ideals that contain I, and one was selected by a choice function, but we don't even know what the choice function is so we can't even determine J in principle.

There are many other applications of Zorn's lemma. For instance, you can prove that every vector space has a basis, even when the vector space is infinite dimensional.

3.8 Unique factorization domains, principal ideal domains, and Euclidean domains

Not every integral domain is as nice as the ring of integers. The ring of integers has three nice properties. One is unique factorization—every integer is uniquely a product of prime numbers. A second is that every ideal is a principal ideal. A third is that there is a division algorithm that is the basis of the Euclidean algorithm.

We'll look at these three properties and see how they are connected. There aren't many rings that have all these properties, and some rings have none of them. We'll investigate these properties and their interrelations.

3.8.1 Divisibility in an integral domain

We'll borrow the concepts of divisibility and greatest common divisor from **Z** and apply them to integral domains. We'll separate the concept of prime number in **Z** into two concepts since in some of the integral domains we'll look at they're actually different.

Definition 3.33. The following definitions apply to elements of an integral domain.

- Let a and b be nonzero elements. We'll say *a divides b*, written $a\big|b$ if there exists c such that $ac = b$.
- We'll say that d is *a greatest common divisor* of a and b, if d divides both a and b, and whenever another element e divides both a and b, then e divides d.
- An element x that is not zero and not a unit is *irreducible* if whenever $x = yz$, either y or z is a unit, otherwise it is *reducible*
- An element x is *prime* if whenever $x\big|yz$, then $x\big|y$ or $x\big|z$.

Note that we won't use the notation $d = \text{GCD}(a, b)$ when d is a greatest common divisor since there will be other greatest common divisors, that is, the greatest common divisor is only unique up to a unit. Later, when we look at principal ideal domains, we can use the notation $(c) = (a, b)$ for greatest common divisors which says the principal ideal (c) is the same as the ideal generated by a and b.

Several properties of divisibility follow directly from the definition just like they do with the integral domain is **Z**. (1). 1 divides every element. (2). Each element divides itself. (3). If $a\big|b$ then $a\big|bc$. (4). Divisibility is transitive. (5). If one element divides two other elements, then it divides both their sum and difference. (6). Cancellation: $a\big|b \iff ac\big|bc$.

3.8.2 Unique factorization domains

Unique factorization is a property that we expect. Given any element x in a ring D, we expect that we can factor it into 'atoms,' things that can't be cut further, and that there's only one way to do that. Of course, with our experience with the integers, we know that there's a bit of difficulty in stating the uniqueness part of the claim. For one thing, the order of the factors is variable, and, for another, there are units, like 1 and -1 that can be inserted to change the formal listing of the factors. Still, these are small things that we can deal with.

Definition 3.34. An integral domain is a *unique factorization domain* (UFD) if every element in it is a product of irreducible elements and it is a product of irreducible elements in only one way apart from the order of the product and factors of units.

The ring $\mathbf{Z}$ of integers is, of course, a unique factorization domain. An integer, such as 6 can be written in more than one way as a product of irreducible elements (primes, in the case of integers) $6 = 2 \cdot 3 = (-3) \cdot (-2)$, but the only difference is the order of the primes and the insertions of units in the factorization.

Recall that an ideal I in a commutative ring R is a prime ideal if R/I is an integral domain. Equivalently, I is a prime ideal if (1) $I \neq R$, and (2) $\forall x, y \in R$, if $xy \in I$, then either $x \in I$ or $y \in I$. You can easily prove the following theorem that relates prime elements to prime ideals.

Theorem 3.35. An element x is an integral domain D is prime if and only if the principal ideal (x) is a prime ideal.

Theorem 3.36. If an element in an integral domain is prime, then it irreducible.

Proof. Let x be prime. Suppose that $x = yz$. Then $x|yz$, so either $x|y$ or $x|z$. In the first case, $xw = y$ for some w. Therefore $xwz = yz = x$, and cancelling, $wz = 1$, so z is a unit. Likewise, in the second case y is a unit. Therefore x is irreducible. Q.E.D.

The converse of this theorem does not hold. That is, there are integral domains where not all irreducible elements are prime. We'll see that in this next example. But then a little later, we'll see that in principal ideal domains (about to be defined), irreducible elements are prime.

Example 3.37 (a nonUFD)**.** We'll find a number of other UFDs, but, it's important to know that not every integral domain has unique factorization. Consider the integral domain $R = \mathbf{Z}[\sqrt{10}]$. An element of it is of the form $x + y\sqrt{10}$ where x and y are integers. In this integral domain 9 can be factored in two ways.

$$9 = 3^2 = (\sqrt{10} + 1)(\sqrt{10} - 1),$$

but 3, $\sqrt{10} + 1$, and $\sqrt{10} - 1$ are all irreducible. This integral domain, and many others, are not UFDs. Although the three elements 3, $\sqrt{10}+1$, and $\sqrt{10}-1$ are irreducible, none divides any other, so none of them is prime, as you can see by the equation involving 9, above.

3.8.3 Principal ideal domains

A second nice property that the ring of integers has is that every ideal in $\mathbf{Z}$ is generated by a single element. If I is an ideal in $\mathbf{Z}$, then the GCD of all it's nonzero elements is an element of I and all other elements are multiples of this GCD. This will be our definition of a principal ideal domain (PID), and we'll show that every PID is a UFD. There are UFDs that aren't PIDs, for instance, $\mathbf{Z}[x]$, the ring of polynomials with integer coefficients is one; one nonprincipal ideal is generated by 2 and x.

Definition 3.38. An integral domain is a *principal ideal domain* (PID) if every ideal in the domain is principal, that is, generated by one element.

We'll show in a couple of steps that every PID is a UFD. The first one makes a connection between greatest common divisors and ideals.

Theorem 3.39. Let D be a principal ideal domain with nonzero elements a and b. The ideal $(a, b) = (c)$ for some element c since D is a PID. Then c is a greatest common divisor of a and b.

Proof. Since $a \in c$, therefore $c|a$. Likewise, $c|b$. We also know that $c \in (a, b)$, so $c = xa + yb$ for some elements x and y. To show that c is a greatest common divisor, suppose d is some other common divisor of a and b. Then $a = ud$ and $b = vd$ for some elements u and v. Now,

$$c = xa + yb = xud + yvd = (xu + yv)d.$$

Therefore, $d|c$. Thus c is a greatest common divisor of a and b. Q.E.D.

Theorem 3.40. In a principal ideal domain, irreducible elements are prime.

Proof. Suppose that p is irreducible and $p|ab$. We'll show either $p|a$ or $p|b$. Suppose $p \not| a$. Then the ideal (p, a) is (1) since p is irreducible. Since $1 \in (p, a)$, $1 = xp + ya$ for some elements x and y. Therefore, $b = bxp + aby$. But $p|ab$, so $p|b$. Q.E.D.

We'll use the following lemma to show that elements have factorizations in PIDs. We'll still have to show they're unique. The condition in the lemma is called the *ascending chain condition* (ACC) on ideals, and rings that satisfy it are called *Noetherian rings* in honor of Noether who studied such rings.

Lemma 3.41. In a principal ideal domain, there are no infinitely ascending chains of ideals. That is,

$$(a_1) \subsetneqq (a_2) \subsetneqq (a_3) \subsetneqq \cdots$$

does not exist.

Proof. Suppose there were such an infinitely ascending chain of ideals. Then the union $I = \bigcup_{i=1}^{\infty} (a_i)$ is an ideal, as you can easily check. It must be principal, so $I = (a)$ for some element a. But a is in the union, so it's in one of the ideals (a_i). Then

$$(a) \subseteq (a_i) \subsetneqq (a_{i+1}) \subseteq (a),$$

a contradiction. Q.E.D.

Theorem 3.42. In a principal ideal domain, every element has a factorization into irreducible elements.

Proof. Suppose that an element a_1 has no factorization into irreducible elements. Starting with the ideal (a_1), form any ascending chain of ideals generated by other elements with no factorizations, and extend the chain as far as possible. By the lemma, it stops somewhere, say at (a_n).

$$(a_1) \subsetneqq (a_2) \subsetneqq \cdots \subsetneqq (a_n).$$

We now have an element a_n which has no factorization into irreducible elements, but any ideal strictly containing (a_n) is generated by an element that does have such a factorization. Now, a_n is not irreducible itself, for that would be a factorization, so $a_n = bc$ where neither b nor c is a unit. Since both $(a_n) \subsetneqq (b)$ and $(a_n) \subsetneqq (c)$, therefore both b and c have factorizations, and the product of those factorizations gives a factorization for a_n, a contradiction. Q.E.D.

Theorem 3.43. Every principal ideal domain is a unique factorization domain.

Proof. The last theorem gave the existence of at least one factorization for an element a. We still have to show that there's at most one factorization. Suppose that a has two factorizations as products of irreducible elements.

$$a = p_1 \cdots p_n = q_1 \cdots q_m$$

Since the irreducible element p_1 is prime (in a PID), p_1 divides one of the q_i's, which we can renumber as q_1. Then $p_1 = u_1 q_1$ where u_1 is a unit. Substitute $u_1 q_1$ for p_1 and cancel q_1 to get the equation

$$u_1 p_2 \cdots p_n = q_2 \cdots p_n.$$

That completes the inductive step of mathematical induction. You can verify the base case to complete the proof. Q.E.D.

3.8.4 Euclidean domains

The third nice property that **Z** has is that there is a division algorithm that is the basis of the Euclidean algorithm.

For the integers, the division algorithm starts with an integer a (the dividend) and a nonzero integer b (the divisor) and delivers q (the quotient) and r (the remainder) such that

$$a = qb + r \quad \text{and} \quad 0 \le r < b.$$

This property allowed us to construct the Euclidean algorithm for finding GCDs and the extended Euclidean algorithm to show that the greatest common divisor of two numbers is a linear combination of them.

There are a few other integral domains that have the same kind of division algorithm where the remainder is somehow "smaller" than the divisor, but the concept of smaller and how to find q and r differs from domain to domain.

Definition 3.44. A *Euclidean valuation* on an integral domain D is a function $v : D - 0 \to \mathbf{Z}_{\ge 0}$ that satisfies the conditions

1. for nonzero elements a and b, $v(a) \le v(b)$, and
2. for each element a (the dividend) and nonzero element b (the divisor), there are elements q (the quotient) and r (the remainder) such that

$$a = qb + r \quad \text{where either } r = 0 \text{ or } v(r) < v(b).$$

An integral domain that admits a Euclidean valuation is called *Euclidean domain.*

Of course, **Z** is a Euclidean domain with the valuation being the absolute value $v(a) = |a|$. Another class of Euclidean domains are the rings of polynomials (in one variable) with coefficients in a given field.

Theorem 3.45 (The division algorithm for polynomials)**.** Let D be an integral domain, and $D[x]$ its polynomial ring in one variable. Let $f(x)$ be a polynomial (the dividend) and $g(x)$ a monic polynomial (the divisor). Then there exist unique polynomials $q(x)$ (the quotient) and $r(x)$ (the remainder) such that

$$f(x) = q(x)g(x) + r(x) \quad \text{where either } r(x) = 0 \text{ or } \deg(r(x)) < \deg(g(x)).$$

Proof. One case is when $f(x) = 0$ or $\deg f(x) < \deg g(x)$. Since the dividend already has a lower degree, the quotient $q(x) = 0$ and the remainder $r(x) = f(x)$.

That leaves the case when $\deg f(x) \geq \deg g(x)$. We'll prove it by induction on $n = \deg f(x)$ where the base case is $n = 0$. That's the case where $f(x)$ and $g(x)$ are both constants in the domain D, but $g(x)$ is monic, so $g(x) = 1$. Then $q(x) = f(x)$ and $r(x) = 0$.

Now for the inductive step. We'll assume the inductive hypothesis that the theorem is correct for all polynomials $f(x)$ of degree $< n$ and show it's true for those of degree n. Let

$$f(x) = a_0 + a_1x + \cdots + a_nx^n \text{ and } g(x) = b_0 + b_1x + \cdots + b_{m-1}x^{m-1} + x^m \text{ where } n \geq m.$$

The polynomial

$$f_1(x) = f(x) - a_nx^{n-m}g(x)$$

has a 0 coefficient for x^n, so its degree is less than n. By inductive hypothesis, there are polynomials $q_1(x)$ and $r_1(x)$ such that

$$f_1(x) = q_1(x)g(x) + r_1(x) \quad \text{where } r_1(x) = 0 \text{ or } \deg r_1(x) < \deg g(x).$$

We can eliminate the $f_1(x)$ from the last two equations to get

$$f(x) = (a_1(x) + a_nx^{n-m})g(x) + f_1(x).$$

That gives us the desired representation $f(x) = q(x)g(x)+r(x)$, finishing the inductive proof.

We now know the existence of the polynomials $q(x)$ and $r(x)$. We'll leave the uniqueness part to the reader since it's the existence we need. The uniqueness part doesn't need induction but there are a lot of details, Q.E.D.

You can easily make a slight modification of the proof to give a slightly better division algorithm for polynomials with coefficients in a field. The requirement that the divisor polynomial be monic can be dropped.

Theorem 3.46 (The division algorithm for polynomials over a field)**.** Let F be a field, and $F[x]$ its polynomial ring in one variable. Let $f(x)$ and $g(x)$ be polynomials. Then there exist unique polynomials $q(x)$ and $r(x)$ such that

$$f(x) = q(x)g(x) + r(x) \quad \text{where either } r(x) = 0 \text{ or } \deg(r(x)) < \deg(g(x)).$$

Corollary 3.47. The polynomial ring $F[x]$ with coefficients in a field F is a Euclidean domain where the valuation v assigns to a polynomial $f(x)$ the degree of f.

Soon we'll study polynomial rings in more detail.

There are other Euclidean domains, and if we have time we'll look at one. One that has already been mentioned is the ring of Gaussian integers $\mathbf{Z}[i] = \{a_1 + a_2 i \,|\, a_1, a_2 \in \mathbf{Z}\}$. Its valuation function, also called the norm, is $v(a_1 + a_2 i) = a_1^2 + a_2^2$. In order to divide one Gaussian integer $b_1 + b_2 i$ into another $a_1 + a_2 i$ to get a quotient $q_1 + q_2 i$ and remainder $r_1 + r_2 i$, you can perform the complex division $\dfrac{a_1 + a_2 i}{b_1 + b_2 i}$ to get an exact quotient, and choose $q_1 + q_2 i$ to be the closest Gaussian integer to that exact quotient. The remainder is then determined.

The Euclidean algorithm in Euclidean domains. First, we'll show that Euclidean domains are principal ideal domains, and therefore Unique Factorization Domains. Then we'll look at an example of the Euclidean algorithm in a Euclidean domain other than $\mathbf{Z}$.

Theorem 3.48. A Euclidean domain is a principal ideal domain.

Proof. Let I be an ideal in a Euclidean domain D with valuation v. We'll show I is a principal ideal. If I is the zero ideal (0), then it's principal of course. Assume now that I has a nonzero element, and let $S = \{v(x) \,|\, 0 \neq x \in I\}$. This is a nonempty subset of the nonnegative integers, so it has a least element, and let that be $v(a)$. Thus, a is a nonzero element of I, so $(a) \subseteq I$. Let x be any other nonzero element in I. Then $v(a) \leq v(x)$. Furthermore, there are elements q and r in D such that $x = aq + r$ and either $r = 0$ or $v(f) < v(a)$. But $r = x - aq \in I$, so if $r \neq 0$, then $v(r) > v(a)$ contradicts $v(a) \leq v(r)$. Therefore, $r = 0$, and hence $x = aq$, so $a|x$. Therefore, $I = (a)$. Thus, D is a PID. Q.E.D.

The Euclidean algorithm works in any Euclidean domain the same way it does for integers. It will compute the greatest common divisor (up to a unit), and the extended Euclidean algorithm will construct the greatest common divisor as a linear combination of the original two elements.

Let's take an example from the polynomial ring $\mathbf{Q}[x]$. Let's find the greatest common divisor of $f_1(x) = x^4 + 2x^3 - x - 2$ and $f_2(x) = x^4 - x^3 - 4x^2 - 5x - 3$. They have the same degree, so we can take either one of them as the divisor; let's take $f_2(x)$. Divide f_2 into f_1 to get a quotient of 1 and remainder of $f_3(x) = 3x^3 + 4x^2 + 4x + 1$. Then divide f_3 into f_2 to get a quotient and a remainder f_4, and continue until the remainder is 0 (which occurs on the next iteration.

$$\begin{array}{ll}
f_1(x) = x^4 + 2x^3 - x - 2 & f_1(x) = 1 \cdot f_2(x) + f_3(x) \\
f_2(x) = x^4 - x^3 - 4x^2 - 5x - 3 & f_2(x) = (\frac{1}{3}x - \frac{7}{9}) f_3(x) + f_4(x) \\
f_3(x) = 3x^3 + 4x^2 + 4x + 1 & f_3(x) = (\frac{27}{20}x - \frac{9}{20}) f_4(x) \\
f_4(x) = -\frac{20}{9}x^2 - \frac{20}{9}x - \frac{20}{9} &
\end{array}$$

Thus, a greatest common divisor is $f_4(x)$, which differs by a unit factor from the simpler greatest common divisor $x^2 + x + 1$. We can read the equations on the right in reverse to get f_4 as a linear combination of f_1 and f_2.

$$\begin{array}{rcl}
f_4(x) & = & f_2(x) - (\frac{1}{3}x - \frac{7}{9}) f_3(x) \\
& = & f_2(x) - (\frac{1}{3}x - \frac{7}{9})(f_1(x) - f_2(x)) \\
& = & (\frac{1}{3}x + \frac{2}{9}) f_2(x) - (\frac{1}{3}x - \frac{7}{9}) f_1(x)
\end{array}$$

3.9 Polynomial rings

We know a fair amount about $F[x]$, the ring of polynomials over a field F. It has a division algorithm, so it's a Euclidean domain with the Euclidean valuation being the degree of a polynomial, and it has division and Euclidean algorithms. Since it's Euclidean, it's also a principal ideal domain, and that means irreducible elements are prime, but we'll still use the term irreducible polynomial rather than prime polynomial. And since it's a PID, it's also a unique factorization domain, that is, every polynomial uniquely factors as a product of irreducible polynomials.

The nonzero prime ideals of $F[x]$ are just the principal ideals (f) generated by irreducible polynomials $f \in F[x]$, and, furthermore, they're maximal ideals, so $F[x]/(f)$ is a field. We've seen examples of this, for instance, $\mathbf{R}[x]/(x^2+1) \cong \mathbf{R}[i] = \mathbf{C}$, $\mathbf{Q}[x]/(x^2-2) \cong \mathbf{Q}(\sqrt{2})$, and $\mathbf{Z}_3[x]/(x^2+1) \cong \mathbf{Z}_3(i)$.

The main question for $F[x]$ is: what are the irreducible polynomials?

We'll study a few more properties for general polynomial rings, then look at $\mathbf{C}[x]$, then at $\mathbf{R}[x]$.

3.9.1 Polynomial rings with coefficients in a integral domain

We'll list some basic properties without proof. Assume that we're looking at polynomials with coefficients in an integral domain D.

- *The remainder theorem.* Dividing $f(x)$ by a linear polynomial $x-a$ gives a remainder equal to $f(a)$.

- *The factor theorem.* A linear polynomial $x-a$ divides $f(x)$ if and only if a is a root of $f(x)$, that is, $f(a)=0$.

- If $\deg f = n$ and $a_1, a_2, \ldots, a_n$ are n distinct roots of f, then

  $$f(x) = a(x-a_1)(x-a_2)\cdots(x-a_n)$$

  where a is the leading coefficient of f.

- A polynomial of degree n has at most n distinct roots.

- If two monic polynomials f and g both of degree n have the same value at n places, then they are equal.

3.9.2 $\mathbf{C}[x]$ and the Fundamental Theorem of Algebra

In the 16th century Cardano (1501–1576) and Tartaglia (1500–1557) and others found formulas for roots of cubic and quartic equations in terms of square roots and cube roots. At the time, only positive numbers were completely legitimate, negative numbers were still somewhat mysterious, and the first inkling of a complex number appeared. Incidentally, at this time symbolic algebra had not been developed, so all the equations were written in words instead of symbols!

Here's an illustration of how complex numbers arose. One of Cardano's cubic formulas gives the solution to the equation $x^3 = cx + d$ as

$$x = \sqrt[3]{d/2 + \sqrt{e}} + \sqrt[3]{d/2 - \sqrt{e}}$$

where $e = (d/2)^2 - (c/3)^3$. Bombelli used this to solve the equation $x^3 = 15x + 4$, which was known to have 4 as a solution, to get the solution

$$x = \sqrt[3]{2 + \sqrt{-121}} + \sqrt[3]{2 - \sqrt{-121}}.$$

Now, $\sqrt{-121}$ is not a real number; it's neither positive, negative, nor zero. Bombelli continued to work with this expression until he found equations that lead him to the solution 4. Assuming that the usual operations of arithmetic held for these "numbers," he determined that

$$\sqrt[3]{2 + \sqrt{-121}} = 2 + \sqrt{-1} \quad \text{and} \quad \sqrt[3]{2 - \sqrt{-121}} = 2 - \sqrt{-1}$$

and, therefore, the solution $x = 4$.

Cardano had noted that the sum of the three solutions of a cubic equation $x^3+bx^2+cx+d = 0$ is $-b$, the negation of the coefficient of x^2. By the 17th century the theory of equations had developed so far as to allow Girard (1595–1632) to state a principle of algebra, what we call now "the fundamental theorem of algebra." His formulation, which he didn't prove, also gives a general relation between the n solutions to an nth degree equation and its n coefficients.

For a generic equation

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = 0$$

Girard recognized that there could be n solutions, if you allow all roots and count roots with multiplicity. So, for example, the equation $x^2 + 1 = 0$ has the two solutions $\sqrt{-1}$ and $-\sqrt{-1}$, and the equation $x^2 - 2x + 1 = 0$ has the two solutions 1 and 1. Girard wasn't particularly clear what form his solutions were to have, just that there were n of them: $x_1, x_2, \ldots, x_n$.

Girard gave the relation between the n roots $x_1, x_2, \ldots, x_n$ and the n coefficients $a_1, \ldots, a_n$ that extended Cardano's remark. First, the sum of the roots $x_1 + x_2 + \cdots + x_n$ is $-a_1$ (Cardano's remark). Next, the sum of all products of pairs of solutions is a_2. Next, the sum of all products of triples of solutions is $-a_3$. And so on until the product of all n solutions is either a_n (when n is even) or $-a_n$ (when n is odd).

Here's an example. The 4th degree equation

$$x^4 - 6x^3 + 3x^2 + 26x - 24 = 0$$

has the four solutions -2, 1, 3, and 4. The sum of the solutions equals 6, that is $-2+1+3+4 = 6$. The sum of all products of pairs (six of them) is

$$(-2)(1) + (-2)(3) + (-2)(4) + (1)(3) + (1)(4) + (3)(4)$$

which is 3. The sum of all products of triples (four of them) is

$$(-2)(1)(3) + (-2)(1)(4) + (-2)(3)(4) + (1)(3)(4)$$

which is 26. And the product of all four solutions is -24.

Over the remainder of the 17th century, negative numbers rose in status to be full-fledged numbers. But complex numbers remained suspect through much of the 18th century. They weren't considered to be real numbers, but they were useful in the theory of equations and becoming more and more useful in analysis. It wasn't even clear what form the solutions to equations might take. Certainly "numbers" of the form $a + b\sqrt{-1}$ were sufficient to solve quadratic equations, even cubic and quartic equations.

Euler did a pretty good job of studying complex numbers. For instance, he studied the unit circle assigning the value $\cos\theta + i\sin\theta$ to the point on the unit circle at an angle θ clockwise from the positive real axis. (He didn't use the word 'radian'; that word was coined later.) In his study of this circle he developed what we call Euler's identity

$$e^{i\theta} = \cos\theta + i\sin\theta.$$

This was an especially useful observation in the solution of differential equations. Because of this and other uses of i, it became quite acceptable for use in mathematics. By the end of the 18th century numbers of the form $x + iy$ were in fairly common use by research mathematicians, and it became common to represent them as points in the plane.

Yet maybe some other form of "number" was needed for higher-degree equations. The part of the Fundamental Theorem of Algebra which stated there actually are n solutions of an nth degree equation was yet to be proved, pending, of course, some description of the possible forms that the solutions might take.

Still, at nearly the end of the 18th century, it wasn't yet certain what form all the solutions of a polynomial equation might take. Finally, in 1799, Gauss (1777–1855) published his first proof of the Fundamental Theorem of Algebra.

We won't look at his or any other proof of the theorem. We will, however, use the theorem.

Definition 3.49. A field F is *algebraically closed* if every polynomial $f \in F[x]$ factors as a product of linear factors. Equivalently, a polynomial f of degree n has n roots in F counting multiplicities.

A weaker definition could be made, and that's that every polynomial of degree at least 1 has at least one root in F. By induction, the remaining roots can be shown to exist.

Thus, the Fundamental Theorem of Algebra is a statement that $\mathbf{C}$ is an algebraically closed field. Therefore, the algebra of $\mathbf{C}[x]$ is particularly simple. The irreducible polynomials are the linear polynomials.

3.9.3 The polynomial ring $\mathbf{R}[x]$

Let's turn our attention now to polynomials with real coefficients. Much of what we can say about $\mathbf{R}[x]$ comes from the relation of $\mathbf{R}$ as a subfield $\mathbf{C}$, and consequently from the relation of $\mathbf{R}[x]$ as a subring of $\mathbf{C}[x]$. That is to say, we can interpret a polynomial f with real coefficients as a polynomial with complex coefficients.

Theorem 3.50. If a polynomial f with real coefficients has a complex root z, then its complex conjugate $\overline{z}$ is also a root.

Proof. Let $f(x) = a_nx^n + \cdots + a_1x + a_0$ where each $a_i \in \mathbf{R}$. If z is a root of f, then $f(z) = a_nz^n + \cdots + a_1z + a_0 = 0$. Take the complex conjugate of the equation, and note that $\overline{a}_i = a_i$. Then $f(\overline{z}) = a_n\overline{z}^n + \cdots + a_1\overline{z} + a_0 = 0$. Thus, $\overline{z}$ is also a root. Q.E.D.

This theorem tells us for a polynomial f with real coefficients, its roots either come in pairs of a complex number or singly as real numbers. We can name the $2k$ complex roots as

$$z_1, \overline{z}_1, z_2, \overline{z}_2, \ldots, z_k, \overline{z}_k$$

that is, as

$$x_1 + yi_1, x_1 - iy_1, x_2 + yi_2, x_2 - iy_2, \ldots, x_k + yi_k, x_k - iy_k$$

and the $n - 2k$ real roots as

$$r_{2k+1}, \ldots, r_n.$$

Using the fact that $\mathbf{C}$ is algebraically closed, we can write f as

$$\begin{aligned} f(x) &= a_n(x - z_1)(x - \overline{z}_1)\cdots(x - z_k)(x - \overline{z}_k)(x - r_{2k+1})\cdots(x - r_n) \\ &= a_n(x^2 - 2x_1x + x_1^2 + y_1^2)\cdots(x^2 - 2x_kx + x_k^2 + y_k^2)(x - r_{2k+1})\cdots(x - r_n) \end{aligned}$$

This last expression has factored f into quadratic and linear polynomials with real coefficients.

Theorem 3.51. The irreducible polynomials in $\mathbf{R}[x]$ are the linear polynomials and the quadratic polynomials with negative discriminant.

Proof. The remarks above show that only linear and quadratic polynomials can be irreducible. Linear polynomials are always irreducible. A quadratic polynomial will have no real roots when its discriminant is negative. Q.E.D.

3.10 Rational and integer polynomial rings

We've studied the irreducible polynomials in $\mathbf{C}[x]$ and $\mathbf{R}[x]$ with the help of the Fundamental Theorem of Algebra and found them to be easily classified. The irreducible polynomials in $\mathbf{C}[x]$ are the linear polynomials, and irreducible polynomials in $\mathbf{R}[x]$ are the linear polynomials and quadratic polynomials with negative discriminant. Determining which polynomials in $\mathbf{Q}[x]$ are irreducible is much harder. Of course, all the linear ones are, and we'll be able to tell which quadratic and cubic ones are irreducible fairly easily. After that it becomes difficult.

3.10.1 Roots of polynomials

The quadratic case. Let's look at a quadratic polynomial $f(x) = ax^2 + bx + c$. It will only be reducible over $\mathbf{Q}$ when it factors as two linear factors, that is, when it has rational roots. But we know that its complex roots are $\dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$. These are rational roots if and only if the discriminant $b^2 - 4ac$ is a perfect square. Thus, $f(x)$ is irreducible if and only if the discriminant is not a perfect square.

The cubic case. It is more difficult to determine when a cubic polynomial $f(x) = ax^3 + bx^2 + cx + d$ is irreducible, but not too difficult. Note that if f factors, then one of the factors has to be linear, so the question of reducibility reduces to the existence of a rational root of f.

Various solutions of a cubic equation $ax^3 + bx^2 + cx + d = 0$ have been developed. Here's one. First, we may assume that f is monic by dividing by the leading coefficient. Our equation now has the form $x^3 + bx^2 + cx + d = 0$. Second, we can eliminate the quadratic term by replacing x by $y - \frac{1}{3}b$. The new polynomial in y will have different roots, but they're only translations by $\frac{1}{3}b$. We now have the cubic equation

$$y^3 + (c - \tfrac{1}{3}b^2)y + (\tfrac{2}{27}b^3 - \tfrac{1}{3}bc + d) = 0$$

which we'll write as

$$y^3 + py + q = 0.$$

We'll follow Viète's method and replace y by $z - \dfrac{p}{3z}$. After simplifying and clearing the denominators we'll have the equation

$$z^6 + qz^3 - \frac{p^3}{27z} = 0$$

which is a quadratic equation in z^3. Its complex solutions are

$$z^3 = \frac{-q \pm \sqrt{q^2 + 4p^3/27}}{2} = -\frac{q}{2} \pm \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}.$$

Taking complex cube roots to get three values for z, then using $y = z - \dfrac{p}{3z}$ to determine y and $x = y - \frac{1}{3}b$ to determine x, we have the all three complex solutions to the original equation. At least one of these three complex solutions is real, and perhaps all three. But it's still a chore to determine when one of the roots is rational.

Some other way is needed to determine if there is a rational root, and there is one.

Rational roots of a polynomial. If we're looking for the roots of a polynomial with rational coefficients, we can simplify the job a little bit by clearing the denominators so that all the coefficients are integers. The following theorem helps in finding roots.

Theorem 3.52. Let $f(x) = a_nx^n + \cdots + a_1x + a_0$ be a polynomial with integral coefficients. If r/s is a rational root of f with r/s in lowest terms, then r divides the constant a_0 and s divides the leading coefficient a_n.

Proof. Since r/s is a root, therefore

$$f(x) = a_n(r/s)^n + a_n(r/s)^{n-1} + \cdots + a_1(r/s) + a_0 = 0,$$

and so, clearing the denominators, we have

$$a_nr^n + a_nr^{n-1}s + \cdots + a_1rs^{n-1} + a_0s^n = 0.$$

We can rewrite this equation as

$$(a_n r^{n-1} + a_n r^{n-2} s + \cdots + a_1 s^{n-1}) r = -a_0 s^n.$$

Now, since r divides $-a_0 s^n$, and r is relatively prime to s, and hence to s^n, therefore r divides a_0. In like manner, you can show s divides a_n. Q.E.D.

For example, to find the rational roots r/s of $f(x) = 27x^4 + 30x^3 + 26x^2 - x - 4$, r will have to divide 4, so the possibilities for r are $\pm 1, \pm 2, \pm 4$, and s will have to divide 27, so the possibilities for s are $1, 3, 9, 27$ (since we may assume s is positive). That gives 24 rational numbers to check, and among them will be found the two rational roots $\frac{1}{3}$ and $-\frac{4}{9}$. After one, $\frac{r}{s}$, is found f can be divided by $x - \frac{r}{s}$ to lower the degree of the polynomial and simplify the problem.

If a polynomial does have a rational root, then it's clearly reducible since that rational root determines a linear factor of the polynomial. But if a polynomial does not have a rational root, then it still may factor as quadratic and higher degree terms, that is, if its degree is at least 4. For example, $x^4 + x^2 + 1$ has no rational roots, but it factors as $(x^2 + x + 1)(x^2 - x + 1)$, so it is reducible.

3.10.2 Gauss's lemma and Eisenstein's criterion

Further study of $\mathbf{Q}[x]$ will require looking at $\mathbf{Z}[x]$. In other words, in order to study polynomials with rational coefficients, we'll have to look at polynomials with integral coefficients. We can take a polynomial with rational coefficients and multiply it by the least common multiple of the denominators of its coefficients to get another polynomial with the same roots but with integral coefficients. We can also divide by the greatest common divisor of the resulting coefficients to get yet another polynomial with the same roots, with integral coefficients, and the greatest common divisor of all its coefficients is 1. Such a polynomial is called primitive.

After that, we'll be able to prove Gauss's lemma which says that a primitive polynomial $f \in \mathbf{Z}[x]$ is reducible in $\mathbf{Q}[x]$ if and only if it's reducible in $\mathbf{Z}[x]$.

We can make more use of these results if, instead of considering just the case of the domain $\mathbf{Z}$ and its field of fractions $\mathbf{Q}$, we generalize to any unique factorization domain D and its field of fractions F. So, for the following discussion, fix a UFD D, and let F denote its field of fractions. Though, keep in mind the basic case when $D = \mathbf{Z}$, $F = \mathbf{Q}$, $D/(p) = \mathbf{Z}_p$, and $D/(p)[x] = \mathbf{Z}_p[x]$ to get a better idea of what's going on.

When we have a prime p in D, the projection $\gamma : D \to D/(p)$ induces a ring epimorphism $D[x] \to D/(p)[x]$ between polynomial rings where the coefficients of f are reduced modulo p giving a polynomial in $D/(p)[x]$. We'll denote the resulting polynomial in $D/(p)[x]$ by f_p.

Definition 3.53. The *content* of a polynomial in $D[x]$ is the greatest common divisor of all of its coefficients. If the content is 1, the polynomial is called *primitive.*

The content of a polynomial is only defined up to a unit.

Evidently, every polynomial in $D[x]$ equals a constant times a primitive polynomial, the constant being its content.

Lemma 3.54 (Gauss)**.** The product of two primitive polynomials in $D[x]$ is primitive, and the content of the product of any two polynomials in $D[x]$ is the product of their contents (up to a unit).

Proof. In order to show the first statement, we'll show if the product is not primitive, then one of the two polynomials is not primitive.

Let f and g be primitive polynomials and suppose that their product fg is not primitive. Then some prime p of D divides the content of fg, so p divides every coefficient of fg. Therefore, in $D/(p)[x]$, $(fg)_p = 0$, so $f_p g_p = 0$. But $D/(p)[x]$ is an integral domain (in fact, a UFD), so either $f_p = 0$ or $g_p = 0$. Therefore, p either divides all the coefficients of f or all the coefficients of g, hence one or the other is not primitive.

The second statement follows from the first just by using the fact that a polynomial equals its content times a primitive polynomial. Q.E.D.

Theorem 3.55 (Mod p irreducibility test.)**.** Let p be a prime integer, and let f be a polynomial whose leading coefficient is not divisible by p. If f is reducible in $F[x]$, then f_p is reducible in $D/(p)[x]$. If f_p is irreducible in $D/(p)[x]$, then f is irreducible in $F[x]$.

Proof. Suppose f is reducible in $F[x]$. Then there exist $g, h \in D[x]$ such that $f = gh$ where the degrees of g and h are at least 1. Since $f = gh$, therefore, $f_p = g_p h_p$. Since p does not divide the leading coefficient of f, neither does it divide the leading coefficients of g or h. Therefore $\deg g_p = \deg g \geq 1$ and $\deg h_p = \deg h \geq 1$. Thus, f_p is reducible. The last statement of the theorem is the contrapositive of the previous. Q.E.D.

Example 3.56. Consider any cubic polynomial f in $\mathbf{Z}[x]$ with an odd leading coefficient, an odd constant, and one of the other two coefficients odd, for instance, $f(x) = 77x^3 + 15x^2 + 8x + 105$. Reduce it modulo 2. For $f(x) = 77x^3 + 15x^2 + 8x + 105$, you'll get $f_2(x) = x^3 + x^2 + 1$. The resulting f_2 will have no roots in $\mathbf{Z}_2$ since it has three nonzero terms. A cubic polynomial with no roots is irreducible, so f_2 is irreducible in $\mathbf{Z}_2[x]$. Hence, by the mod p irreducibility test, f is irreducible in $\mathbf{Q}[x]$.

Another useful irreducibility test is Eisenstein's criterion.

Theorem 3.57 (Eisenstein's criterion)**.** Let $f \in D[x]$. If a prime p does not divide the leading coefficient of f, but it does divide all the other coefficients, and p^2 does not divide the constant of f, then f is irreducible in $F[x]$.

Proof. Suppose f is reducible. As in the previous theorem, there exist $g, h \in D[x]$ such that $f = gh$ where the degrees of g and h are at least 1. Reduce everything modulo p. Then $a_n x^n = f_p(x) = g_p(x) h_p(x)$ where a_n is the leading coefficient of f. Now $\mathbf{Z}_p[x]$ is a UFD, and since $f_p(x)$ is the unit a_n times the irreducible x raised to the nth power, therefore x divides both $g_p(x)$ and $h_p(x)$. Therefore $g_p(0) = h_p(0) = 0$. That means that p divides the constant terms of both g and h, which implies p^2 divides the constant term of f, contrary to the assumption. Q.E.D.

Example 3.58. Consider the polynomial $f(x) = x^n - a$. As long as a has a prime factor that appears to the first power, then Eisenstein's criterion implies f is irreducible.

Example 3.59 (Prime cyclotomic polynomials)**.** The polynomial $x^n - 1$ has as its roots the nth roots of unity (roots of 1) in $\mathbf{C}$. For instance when $n = 4$ its roots are $\pm 1, \pm i$. It is reducible for $n \geq 2$ since 1 is one of its roots. For a prime p, the p^{th} *cyclotomic polynomial* is

$$\Phi_p(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + \cdots + x + 1.$$

We'll use Eisenstein's criterion to show Φ_p is irreducible, but not directly. First, we'll use a translation. Let

$$f(x) = \Phi(x+1) = \frac{(x+1)^p - 1}{x} = x^{p-1} + \binom{p}{p-1} x^{p-2} + \cdots + \binom{p}{2} x + \binom{p}{1}.$$

Then Eisenstein's criterion applies to f. Since f is irreducible, so is Φ.

3.10.3 Polynomial rings with coefficients in a UFD, and polynomial rings in several variables.

Gauss's lemma has more uses than we've used it for. We can use it to show that if D is a UFD, then so is the polynomial ring $D[x]$. And we can apply that statement to conclude a polynomial ring $D[x, y]$ in two or $D[x_1, \ldots, x_n]$ more variables is also a UFD. Although these rings are UFDs, they're not PIDs.

Theorem 3.60. Let D be a unique factorization domain and F its ring of fractions. Then $D[x]$ is also a UFD. The irreducible polynomials in $D[x]$ are either irreducible elements of D or have content 1 and are irreducible polynomials in $F[x]$.

Proof. Let f be a nonzero polynomial in $D[x]$. It is equal to its content times a primitive polynomial. Its content is an element of D, and, since D is a UFD, its content uniquely factors (up to a unite) as a product of irreducible elements of D.

We're reduced to showing that that a primitive polynomial f in $D[x]$ of degree at least 1 uniquely factors as a product of irreducible polynomials.

Since f is a polynomial in $D[x]$, it's also a polynomial in $F[x]$, and we know $F[x]$ is a UFD being a polynomial ring with coefficients in a field F. Thus, f uniquely factors in F:

$$f(x) = f_1(x) f_2(x) \cdots f_k(x)$$

where each $f_i(x)$ is irreducible in $F[x]$. We only need to show that this factorization can be carried out in $D[x]$. Each polynomial $f_i(x)$ is a element a_i of F times a primitive polynomial $f_i'(x)$ in $D[x]$, so

$$f(x) = a_1 \cdots a_k f_1'(x) \cdots f_k'(x).$$

Since $f(x)$ is primitive and the product $f_1'(x) \cdots f_k'(x)$ is also primitive, therefore $a_1 \cdots a_k$ is a unit in D. Thus, $f(x)$ factors in $D[x]$. You can also show that it can factor in only one way in $D[x]$ since it only factors in one way in $F[x]$. Q.E.D.

Corollary 3.61. If D is a UFD, then a polynomial ring in several variables $D[x_1, x_2, \ldots, x_r]$ with coefficients in D is also a UFD.

Chapter 4

Groups

Recall that a group is a set equipped with one binary operation that is associative, has an identity element, and has inverse elements. If that binary operation is commutative, then the group is called an Abelian group.

4.1 Groups and subgroups

4.1.1 Definition and basic properties of groups

We'll look at basic properties of groups, and since we'll discuss groups in general, we'll use a multiplicative notation even though some of the example groups are Abelian.

Definition 4.1. The axioms for a group are very few. A group G has an underlying set, also denoted G, and a binary operation $G \times G \to G$ that satisfies three properties.

1. Associativity. $(xy)z = x(yz)$.

2. Identity. There is an element 1 such that $1x = x = x1$.

3. Inverses. For each element x there is an element x^{-1} such that $xx^{-1} = x^{-1}x = 1$.

Theorem 4.2. From these few axioms several properties of groups immediately follow.

1. Uniqueness of the identity. There is only one element e such that $ex = x = xe$, and it is $e = 1$. *Outline of proof.* The definition says that there is at least one such element. To show that it's the only one, suppose e also has the property of an identity and prove $e = 1$.

2. Uniqueness of inverses. For each element x there is only one element y such that $xy = yx = 1$. *Outline of proof.* The definition says that there is at least one such element. To show that it's the only one, suppose that y also has the property of an inverse of x and prove $y = x^{-1}$.

3. Inverse of an inverse. $(x^{-1})^{-1} = x$. *Outline of proof.* Show that x has the property of an inverse of x^{-1} and use the previous result.

4. Inverse of a product. $(xy)-1 = y^{-1}x^{-1}$. *Outline of proof.* Show that $y^{-1}x^{-1}$ has the property of an inverse of xy.

5. Cancellation. If $xy = xz$, then $y = z$, and if $xz = yz$, then $x = y$.

6. Solutions to equations. Given elements a and b there are unique solutions to each of the equations $ax = b$ and $ya = b$, namely, $x = a^{-1}b$ and $y = ba^{-1}$.

7. Generalized associativity. The value of a product $x_1x_2 \cdots x_n$ is not affected by the placement of parentheses. *Outline of proof.* The associativity in the definition of groups is for $n = 3$. Induction is needed for $n > 3$.

8. Powers of an element. You can define x^n for nonnegative values of n inductively. For the base case, define $x^0 = 1$, and for the inductive step, define $x^{n+1} = xx^n$. For negative values of n, define $x^n = (x^{-n})^{-1}$.

9. Properties of powers. Using the definition above, you can prove using induction the following properties of powers where m and n are any integers: $x^m x^n = x^{m+n}$, $(x^m)^n = x^{mn}$. (But note, $(xy)^n$ does not equal $x^n y^n$ in general, although it does for Abelian groups.)

4.1.2 Subgroups

A subgroup H of G is a group whose underlying set is a subset of the underlying set of G and has the same binary operation, that is, for $x, y \in H$, $x \cdot_H y = x \cdot_G y$.

An alternate description of a subgroup H is that it is a subset of G that is closed under multiplication, has 1, and is closed under inverses.

Of course, G is a subgroup of itself. All other subgroups of G, that is, those subgroups that don't have every element of G in them, are called *proper subgroups*.

Also, $\{1\}$ is a subgroup of G, usually simply denoted 1. It's called the *trivial subgroup* of G.

The intersection $H \cap K$ of two subgroups H and K is also a subgroup, as you can easily show. Indeed, the intersection of any number of subgroups is a subgroup.

The union of two subgroups is never a subgroup unless one of the two subgroups is contained in the other.

Example 4.3 (Subgroups of **Z**)**.** Consider the group **Z** under addition. A subgroup of **Z** has to be closed under addition, include 0, and be closed under negation. Besides 0 and **Z** itself, what are the subgroups of **Z**? If the subgroup is nontrivial, then it has a smallest positive element, n. But if n lies in a subgroup, then all multiples, both positive and negative, of n also must be in the subgroup. Thus, nZ is that subgroup of **Z**.

Example subgroups of a group. There are a number of other subgroups of a group that are important in studying nonabelian groups such as the center of a group and the centralizer of an element of a group.

Exercise 4.1. The *center* of a group G is $Z(G) = \{x \in G \mid ax = xa \text{ for all } a \in G\}$, Prove that $Z(G)$ is a subgroup of G.

Exercise 4.2. For $a \in G$, the *centralizer* of a is $Z_a(G) = \{x \in G \,|\, ax = xa\}$. Prove that $Z_a(G)$ is a subgroup of G.

Exercise 4.3. Prove that the center of G is the intersection of all the centralizer subgroups of G.

If S is a subset of G, then there is a smallest subgroup $\langle S\rangle$ of G containing S. It can be described as the intersection of all subgroups H containing S,

$$\langle S\rangle = \bigcap_{S\subseteq H} H.$$

Alternatively, it can be described as the subset of G of all products of powers of elements of S,

$$\langle S\rangle = \{x_1^{e_1}x_2^{e_2}\cdots x_n^{e_n} \,|\, n \geq 0, \text{each } x_i \in S, \text{and each } e_i \in \mathbf{Z}\}.$$

4.1.3 Cyclic groups and subgroups

If a is an element of a group G, then the subset of G generated by a

$$\langle a\rangle = \{a^n \,|\, n \in \mathbf{Z}\}$$

is a subgroup of G. It is called a *cyclic subgroup* of G, or the subgroup *generated* by a. If G is generated by some element a, then G is called a *cyclic group*.

The *order* of an element a in a group is the smallest positive integer n such that $a^n = 1$. It's denoted $\operatorname{ord} a$. If every positive power $a^n \neq 1$, then the order of n is ∞. So, for example, the order of 1 is 1 since $1^1 = 1$. An *involution* a is an element of a group which is its own inverse, $a^{-1} = a$. Clearly, the order of an involution other than 1, $a \neq 1$, is 2.

Exercise 4.4. Prove that the order of a is also equal to the order of the cyclic group (a) generated by a. That is, $\operatorname{ord} a = |\langle a\rangle|$.

An abstract cyclic group of order n is often denoted $C_n = \{1, a, a^2, \ldots, a^{n-1}\}$ when the operation is written multiplicatively. It is isomorphic to the underlying additive group of the ring $\mathbf{Z}_n$ where an isomorphism is $f : \mathbf{Z}_n \to C_n$ is defined by $f(k) = a^k$.

Exercise 4.5. Prove that any subgroup of a cyclic group is itself cyclic.

Exercise 4.6. Let G be a cyclic group of order n and a an element of G. Prove that a generates G, that is, $\langle a\rangle = G$, if and only if $\operatorname{ord} a = n$.

Cyclic groups are all Abelian, since $a^n a^m = a^{m+n} = a^m a^n$. The integers $\mathbf{Z}$ under addition is an infinite cyclic group, while $\mathbf{Z}_n$, the integers modulo n, is a finite cyclic group of order n.

Exercise 4.7. Prove that every cyclic group is isomorphic either to $\mathbf{Z}$ or to $\mathbf{Z}_n$ for some n.

Exercise 4.8. Prove that if k is relatively prime to n, then k generates $\mathbf{Z}_n$.

4.1.4 Products of groups

Just as products of rings are defined coordinatewise, so are products of groups. Using multiplicative notation, if G and H are two groups then $G \times H$ is a group where the product $(x_1, y_1)(x_2, y_2)$ is defined by (x_1x_2, y_1y_2). The identity element in $G \times H$ is $(1, 1)$, and the inverse $(x, y)^{-1}$ is (x^{-1}, y^{-1}). The projections $\pi_1 : G \times H \to G$ and $\pi_2 : G \times H \to H$ are group epimorphisms where $\pi_1(x, y) = x$ and $\pi_2(x, y) = y$.

Also, $\iota_1 : G \to G\times H$ and $\iota_2 : H \to G\times H$ are group monomorphisms where $\iota_1(x) = (x, 1)$ and $\iota_2(y) = (1, y)$. Thus, we can interpret G and H as subgroups of $G \times H$.

Note that G and H are both Abelian groups if and only if $G \times H$ is an Abelian group. The product of two Abelian groups is also called their *direct sum*, denoted $G \oplus H$.

The underlying additive group of a ring is an Abelian group, and some of the results we have for rings give us theorems for Abelian groups. In particular, the Chinese remainder theorem for cyclic rings $\mathbf{Z}_n$ gives us a theorem for cyclic groups C_n.

Theorem 4.4 (Chinese remainder theorem for groups)**.** Suppose that $n = km$ where k and m are relatively prime. Then the cyclic group C_n is isomorphic to $C_k \times C_n$. More generally, if n is the product $k_1 \cdots k_r$ where the factors are pairwise relatively prime, then

$$C_n \cong C_{k_1} \times \cdots \times C_{k_r} = \prod_{i=1}^{r} C_{k_i}.$$

In particular, if the prime factorization of n is $n = p_1^{e_1} \cdots p_r^{e_r}$. Then the cyclic group C_n factors as the product of the cyclic groups $C_{p_i^{e_i}}$, that is,

$$C_n \cong \prod_{i=1}^{r} C_{p_i^{e_i}}.$$

4.1.5 Cosets and Lagrange's theorem

Cosets are useful in developing the combinatorics of finite groups, that is, for counting subgroups and other things related to a finite group. They come in both left and right forms as you'll see in the definition below, but we'll only use left cosets. Our first combinatorial theorem is called Lagrange's theorem which says that the order of a subgroup divides the order of a group. Since the subgroup (a) generated by a single element has an order that divides the order of the group, therefore the order of an element divides the order of the group, too. We'll have our first classification theorem as a corollary, and that is that a group whose order is a prime number is cyclic. Thus, up to isomorphism, there is only one group of that order.

Definition 4.5. Let H be a subgroup of G. A *left coset* of H is a set of the form

$$aH = \{ah \mid h \in H\}$$

while a *right coset* is of the form $Ha = \{ha \mid h \in H\}$.

Theorem 4.6. Several properties of cosets follow from this definition.

1. The coset $1H$ is just the subgroup H itself. In fact, if $h \in H$ then $hH = H$.

2. More generally, $aH = bH$ if and only if $ab^{-1} \in H$. Thus, the same coset can be named in many different ways.

3. Cosets are disjoint. If $aH \neq bH$, then $aH \cap bH = \emptyset$. *Outline of proof.* It's probably easier to show the contrapositive: if $aH \cap bH \neq \emptyset$ then $aH \neq bH$. Suppose an element is in the intersection. Then it can be written as ah or as bh' where both h and h' are elements of H. The rest relies on the previous statement.

4. Cosets of H all have the same cardinality. *Outline of proof.* Check that the function $f(ah) = bh$ is a bijection $aH \to bH$.

5. Thus, the cosets of H partition G into subsets all having the same cardinality.

6. *Lagrange's theorem.* If G is a finite group, and H a subgroup of G, then $|H|$ divides $|G|$. Moreover, $|G|/|H|$ is the number of cosets of H.

Definition 4.7. The *index* of a subgroup H of a group G is the number of cosets of H. The index is denoted $[G : H]$. By Lagrange's theorem, $[G : H] = |G|/|H|$ when G is a finite group.

Corollary 4.8. If the order of a group is a prime number, then the group is cyclic.

Proof. Let $|G| = p$, a prime. Since p has no divisors except 1 and p, therefore, by Lagrange's theorem, G only has itself and the trivial subgroup as its subgroups. Let $a \neq 1$ be an element of G. It generates a cyclic subgroup (a) which isn't trivial, so $(a) = G$. Thus G is cyclic. Q.E.D.

Corollary 4.9. If a group is finite, then the order of every element divides the order of the group.

Proof. Let a be an element of a finite group G. Then the order of the subgroup (a) divides $|G|$. But $\operatorname{ord} a$ is the order of (a). Therefore $\operatorname{ord} a$ divides $|G|$. Q.E.D.

Products of subsets in a group. Occasionally we'll want to look at products HK of subsets H and K, especially when H and K are subgroups of a group G. This product is defined by

$$HK = \{xy \mid x \in H, y \in K\}.$$

Even when H and K are subgroups, it isn't necessary that HK is a subgroup, but there is a simple criterion to test if it is.

Theorem 4.10. Let H and K be subgroups of G. Then HK is also a subgroup of G if and only if $HK = KH$.

Proof. $\Longrightarrow$: Suppose that HK is a subgroup. First, we'll show that $KH \subseteq HK$. Let $xy \in KH$ with $x \in K$ and $y \in H$. Since $x = 1x \in HK$ and $y = y1 \in HK$, therefore their product xy is also in HK. Thus, $KH \subseteq HK$. Next, we'll show that $HK \subseteq KH$. Let $xy \in HK$ with $x \in H$ and $y \in K$. Then $(xy)^{-1}$ is also in HK, so $(xy)^{-1} = x_1y_1$ with $x_1 \in H$ and $y_1 \in K$. Therefore $xy = (x_1y_1)^{-1} = y_1^{-1}x_1^{-1} \in KH$. Thus $HK \subseteq KH$.

$\Longleftarrow$: Suppose that $HK = KH$. To show it's a subgroup, first note $1 \in HK$. Second, we'll show that HK is closed under multiplication. Let x_1y_1 and x_2y_2 be elements of HK with $x_1, x_2 \in H$ and $y_1, y_2 \in K$. Then $y_1x_2 \in KH = HK$, so $y_1x_2 = x_3y_3$ where $x_3 \in H$ and $y_3 \in K$. Therefore, $(x_1y_1)(x_2y_2) = (x_1x_3)(y_3y_2) \in HK$. Third, we'll show that HK is closed under inverses. Let $xy \in HK$ with $x \in H$ and $y \in K$. Then $(xy)^{-1} = y^{-1}x^{-1} \in KH = HK$. Q.E.D.

Corollary 4.11. If H and K are subgroups of an Abelian group G, then HK is also a subgroup of G.

4.2 Symmetric Groups S_n

We've looked at several examples of groups already. It's time to examine some in more detail.

4.2.1 Permutations and the symmetric group

Definition 4.12. A *permutation* of a set X is just a bijection $\rho : X \to X$ on that set. The permutations on X form a group called the *symmetric group*. We're primarily interested in permutations on a finite set. We'll call the elements of the finite set letters, but we'll denote them with numbers. The symmetric group on n elements $1, 2, \ldots, n$ is denoted S_n.

Note that the order of the symmetric group on n letters is $|S_n| = n!$.

A convenient and concise way to denote a permutation ρ is by what is called the cycle notation. Consider this permutation ρ in S_6

$$\begin{array}{rcccccc} n & 1 & 2 & 3 & 4 & 5 & 6 \\ \rho(n) & 4 & 3 & 2 & 6 & 5 & 1 \end{array}$$

Note that $\rho(1) = 4, \rho(4) = 6$, and $\rho(6) = 1$. These three letters form a 3-cycle $1 \overset{\rho}{\mapsto} 4 \overset{\rho}{\mapsto} 6 \overset{\rho}{\mapsto} 1$ of ρ denoted (146). Also note $2 \overset{\rho}{\mapsto} 3 \overset{\rho}{\mapsto} 2$, so (23) is a 2-cycle of ρ. Another name for a 2-cycle is *transposition*. Since $\rho(5) = 5$, therefore (5) by itself is a 1-cycle, also called a *fixed point*, of ρ. The cycle notation for this permutation is $\rho = (146)(23)$. Note that fixed points are not denoted in this notation. Alternatively, this permutation could be denoted (23)(146) or (461)(32).

Since fixed points aren't denoted in cycle notation, we'll need a special notation for the identity permutation since it fixes all points. We'll use 1 to denote the identity.

There's a bit of experience needed to quickly multiply two permutations together when they're in cycle notation. Let $\rho = (146)(23)$ and $\sigma = (15)(2643)$. By $\rho\sigma$ mean first perform the permutation ρ then perform σ (in other words, the composition $\sigma \circ \rho$ if we think of these permutations as functions). Then we need simplify the cycle notation

$$\rho\sigma = (146)(23)\ (15)(2643).$$

Note that first ρ sends 1 to 4, then σ sends 4 to 3, therefore $\rho\sigma$ sends 1 to 3. Next $3 \overset{\rho}{\mapsto} 2 \overset{\sigma}{\mapsto} 6$, so $3 \overset{\rho\sigma}{\mapsto} 6$, likewise $6 \overset{\rho}{\mapsto} 1 \overset{\sigma}{\mapsto} 5$, so $6 \overset{\rho\sigma}{\mapsto} 5$, and $5 \overset{\rho}{\mapsto} 5 \overset{\sigma}{\mapsto} 1$, so $5 \overset{\rho\sigma}{\mapsto} 1$. Thus, we have a cycle of $\rho\sigma$, namely, (1365). You can check that (2) and (4) are fixed points of $\rho\sigma$. Thus, we found the product. $(146)(23)\ (15)(2643) = (1365)$.

Incidentally, finding the inverse of a permutation in cycle notation is very easy—just reverse all the cycles. The inverse of $\rho = (146)(23)$ is $\rho^{-1} = (641)(32)$.

Small symmetric groups When $n = 0$ or $n = 1$, there's nothing in the symmetric group except the identity.

The symmetric group on two letters, S_2, has one nontrivial element, namely, the transposition (12). This is the smallest nontrivial group, and it's isomorphic to any group of order 2. It is, of course, an Abelian group.

The symmetric group on three letters, S_3, has order 6. We can name its elements using the cycle notation.

$$1, (12), (13), (23), (123), (132)$$

Besides the identity, there are three transpositions and two 3-cycles. This is not an Abelian group. For instance $(12)\ (13) = (123)$, but $(13)\ (12) = (132)$.

The symmetric group on four letters, S_4, has order 24. Besides the identity, there are $\binom{4}{2} = 6$ transpositions, $\binom{4}{3} \cdot 2 = 8$ 3-cycles, 6 4-cycles, and 3 products of two 2-cycles, like (12)(34).

4.2.2 Even and odd permutations

First we'll note that every cycle, and therefore every permutation, can be expressed as a product of transpositions. We'll soon see after that that a permutation can either be expressed as a product of an even number of transpositions or as a product of an odd number of transpositions, but not both. That will justify the definition of even and odd permutations.

Theorem 4.13. Any cycle can be expressed as a product of transpositions.

Proof. The cycle $(a_1a_2a_3 \cdots a_k)$ is the product $(a_1a_2)\,(a_1a_3)\,\ldots\,(a_1a_k)$. Q.E.D.

We'll look at an invariant that will help us distinguish even from odd permutations. It is P_n, the product of all differences of the form $i - j$ where $0 < i < j \leq n$.

$$\begin{aligned} P_n &= \prod_{0<i<j\leq n} (i-j) \\ &= (1-2)(1-3)\cdots(1-n) \\ &\ (2-3)\cdots(2-n) \\ &\ \qquad\cdots \\ &\ ((n-1)-n) \end{aligned}$$

Lemma 4.14. The effect of applying a transposition to the integers that make up P_n is to change the sign of P_n.

Proof. Let the transposition be (ab) where $0 < a < b \leq n$. The product P_n is made of three factors $P_n = P'P''P'''$ where $P' = (a - b)$, P'' is the product of factors that have either a or b but not both, and P''' is the product of factors that don't have either a or b. Now the transposition (ab) has no effect at all on P''' but negates P'. Its effect on P'' is more complicated. Suppose c is another letter. Case 1. $c < a < b$. The factors $(c - a)$ and $(c - b)$ of P'' are interchanged by the transposition (ab). Case 2. $a < c < b$. The factors $(a - c)$ and $(c - b)$ are interchanged and both negated. Case 3. $a < b < c$. Like case 1. Thus P'' does not change its value. Since only P' is negated, P_n is negated. Q.E.D.

Theorem 4.15. A permutation ρ is the product of an even number of transpositions if when ρ is applied to P_n above the value does not change; it's the product of an odd number if P_n is negated. (It can't be both since P_n is not 0.)

Definition 4.16. A permutation is *even* if it's the product of an even number of transposition, it's *odd* if it's the product of an odd number of transpositions. The identity 1 is an even permutation.

Note that a cycle is an even permutation if it has an odd length, but it's an odd permutation if it has an even length.

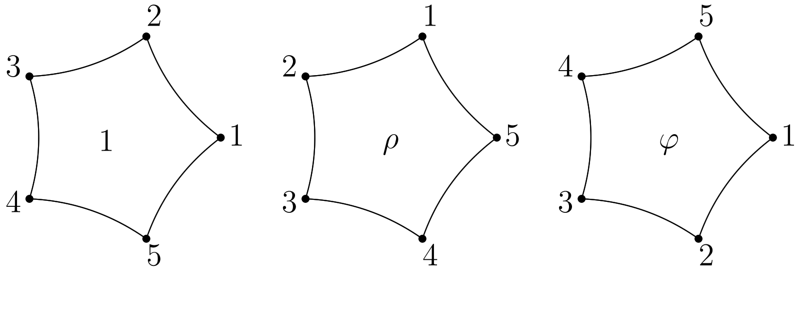
Also, the product of two even permutations is even, the product of two odds is even, and the product of an even and an odd is odd.

4.2.3 Alternating and dihedral groups

Example 4.17 (The alternating group A_n)**.** Since the product of even permutations is even, and the inverse of an even permutation is even, therefore the set of even permutations in the symmetric group S_n is a subgroup of S_n. It is called the *alternating group* on n letters, denoted A_n. For $n \geq 2$, the number of even permutations is the same as the number of odd permutations, since multiplying by the transposition (12) sets up the bijection. Therefore, the order of A_n is half the order of S_n.

Note that when we looked at S_4 above, we saw that there were 12 even permutations (the identity, eight 3-cycles, and three products of two 2-cycles) and there were 12 odd permutations (six transpositions and six 4-cycles).

Example 4.18 (The dihedral group D_n)**.** These are the symmetry groups of regular n-gons. We already looked at the case $n = 3$ of an equilateral triangle. Consider a regular polygon with n vertices.



We can label the vertices in order from 1 to n. A symmetry of a plane figure is a transformation of the plane that maps the figure to itself. We're only interested in isometries, transformations that preserve distance, right now, but other transformations have their applications, too.

The figure shows a pentagon. (The pentagon shown here is in the hyperbolic plane, but that doesn't matter.) One of its symmetries ρ is the one that rotates the pentagon 72° counterclockwise. It maps the vertex labelled 1 to 2, maps 2 to 3, and so forth. Knowing where the vertices are mapped is enough to determine the transformation, so we can identify ρ with the permutation it describes on the set of vertices. This ρ is the permutation (12345).

Another of the symmetries of the pentagon is a reflection like φ shown above, a reflection across a horizontal axis. In cycle notation $\varphi = (25)(34)$.

In fact, there are 10 symmetries of the regular pentagon. Besides the identity, there are four rotations and five reflections.

$$\begin{array}{ccccc} \text{identity} = 1 & \rho = (12345) & \rho^2 = (13524) & \rho^3 = (14253) & \rho^4 = (15432) \\ \varphi = (25)(34) & \varphi\rho = (12)(35) & \varphi\rho^2 = (13)(45) & \varphi\rho^3 = (14)(23) & \varphi\rho^4 = (15)(24) \end{array}$$

There are no more symmetries although we can write more expressions in terms of φ and ρ, for instance $\rho\varphi$. But $\rho\varphi = (15)(24)$ which is $\varphi\rho^4$.

Thus, we can see now how to represent the dihedral group, D_5, in the symmetric group S_5. In fact, it's represented the alternating group, A_5, since all the permutations are even permutations.

Presentations by generators and relations. Although it's nice to have a group represented in a symmetric group, sometimes it's more convenient to describe it more algebraically in terms of generators and relations. For D_5 we can see that ρ and φ are sufficient to generate the whole group in the sense that every element in the group can be written as some expression involving ρ and φ. But there are certain relations, actually equations, that ρ and φ satisfy in this group, namely $\rho^5 = 1$, $\phi^2 = 1$, and $\rho\varphi = \varphi\rho^{-1}$. Thus, we can present the group as

$$D_5 = \langle \rho, \varphi : \rho^5 = 1, \phi^2 = 1, \rho\varphi = \varphi\rho^{-1} \rangle.$$

The difficulty with a presentation of this type is knowing when you have enough generators and relations. If you don't have enough generators, you won't generate the whole group. If you don't have enough relations, you'll generate a larger group, but not the one you want. A proof needs to be supplied to be assured that this is the right presentation. Frequently, a diagram of some sort fills the bill.

4.3 Cayley's theorem and Cayley graphs

One of the reasons symmetric groups are so important is that every group is isomorphic to a subgroup of a symmetric group, a result of Cayley. This gives us another way to look at groups, especially small finite ones.

We'll prove Cayley's theorem, then look at a few Cayley graphs which depend on Cayley's theorem.

4.3.1 Cayley's theorem

Recall that a permutation of a set X is just a bijection $\rho : X \to X$ on that set and permutations on X form a group called the *symmetric group*. When the set is finite, we can write it as $\{1, 2, \ldots, n\}$, and S_n denotes the its symmetric group.

Cayley's theorem can be stated for infinite groups as well as finite groups.

Theorem 4.19 (Cayley)**.** Let G be a group, and let $S(G)$ be the symmetric group on G, that is, the group of permutations on the underlying set of G. The function $\varphi : G \to S(G)$ defined by $\varphi(a)(x) = ax$ is a group monomorphism. Therefore, G is isomorphic to a subgroup of $S(G)$.

Proof. $\varphi(a)$ is the permutation on G that maps x to ax. It's a bijection since its inverse sends x to $a^{-1}x$. To show that it's a group homomorphism, it is only necessary to show that $\varphi(ab) = \varphi(a)\varphi(b)$ for a and b in G. But $\varphi(ab)(x) = abx$, and $(\varphi(a)\varphi(b))(x) = \varphi(a)(\varphi(b)(x)) = \varphi(a)(bx) = abx$. Finally, φ is a monomorphism since if $\varphi(a) = \varphi(b)$, then evaluating the two permutations at 1 gives $a1 = b1$, so $a = b$. Q.E.D.

Although this representation theorem does show that every group is a subgroup of a symmetric group (up to isomorphism), it's practically not all that useful since if the group G has order n, it's being represented in a group of order $n!$, which is much too large to deal with if n is at all large.

Cayley graphs. With a Cayley graph we can represent a group G by a graph with vertices and labeled, directed edges. Each element of G is a vertex of the graph, and for each element a, we also have a directed edge labeled a from a vertex x to the vertex ax. In other words, the Cayley graph is a representation of G by the Cayley theorem to $S(G)$.

For a small example, let G be the cyclic group $G = \{1, a, b\}$ where $a^2 = b$ and $a^3 = 1$. The Cayley graph for G has three vertexes, labeled 1, a, and b. Each node has a loop on it labeled 1 since $1x = x$. There are three edges labelled a, $1 \xrightarrow{a} a \xrightarrow{a} b \xrightarrow{a} 1$, and three edges labelled b, $1 \xrightarrow{b} b \xrightarrow{b} a \xrightarrow{b} 1$. This is probably most conveniently drawn in a triangular figure.

There's a lot of redundancy in the graph in the sense that you don't need all the information to reconstruct the group. The loops labelled 1 might just as well be dropped since for any group $1x = x$. If we know the edges labelled a, then we can determine the edges labelled b since you just travel two a-edges to get a b-edge. That leaves just the triangle $1 \xrightarrow{a} a \xrightarrow{a} b \xrightarrow{a} 1$. More generally, if you know the edges for generators of a group, then all the other edges are determined.

Example 4.20 (D_5)**.** Recall that the dihedral group D_5 has 10 elements and the presentation

$$D_5 = (\rho, \varphi : \rho^5 = \varphi^2 = (\varphi\rho)^2 = 1).$$

The first relation, $\rho^5 = 1$ gives us a five cycle

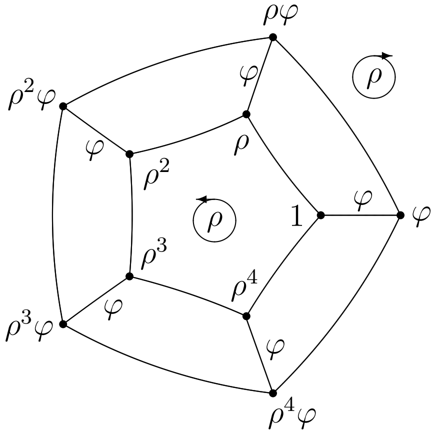
$$1 \xrightarrow{\rho} \rho \xrightarrow{\rho} \rho^2 \xrightarrow{\rho} \rho^3 \xrightarrow{\rho} \rho^4 \xrightarrow{\rho} 1$$

which we can draw as a pentagon, the center pentagon in the graph below. The second relation, $\varphi^2 = 1$, means we have the 2-cycle $1 \xrightarrow{\varphi} \varphi \xrightarrow{\varphi} 1$, and, more generally, for any element

a, we have a 2-cycle $a \xrightarrow{\varphi} a\varphi \xrightarrow{\varphi} a$. We'll draw these 2-cycles as undirected edges $a \overset{\varphi}{-} a$. We get five of these edges, one at each vertex of the center pentagon. The third relation, $(\varphi\rho)^2 = 1$, describes a square

$$a \xrightarrow{\varphi} a\varphi \xrightarrow{\rho} a\varphi\rho \xrightarrow{\varphi} a\varphi\rho\varphi \xrightarrow{\rho} a.$$

Starting at each of the new outer vertices of the graph, follow three edges to reach another outer vertex, and draw a ρ-edge back to where you started. When you finish, you have the Cayley graph for D_5.



Notice that the graph is completely symmetric. You could label any vertex 1 and fill in the names of the rest of the vertices by following the labelled arcs.

There is another presentation for D_5 that gives a different looking Cayley graph. Let $\psi = \rho\phi$. Then

$$D_5 = (\varphi, \psi : \varphi^{=}\psi^2 = (\varphi\psi)^5).$$

The Cayley graph has the same ten vertices, but the edges are all undirected and they form a cycle of length 10 with labels alternating between φ and ψ.

Example 4.21 (A_4)**.** Recall that the alternating group on $\{1,2,3,4\}$ has 12 elements. It's not cyclic, so at least two generators are required to generate it. In fact, two will do. Consider the three elements

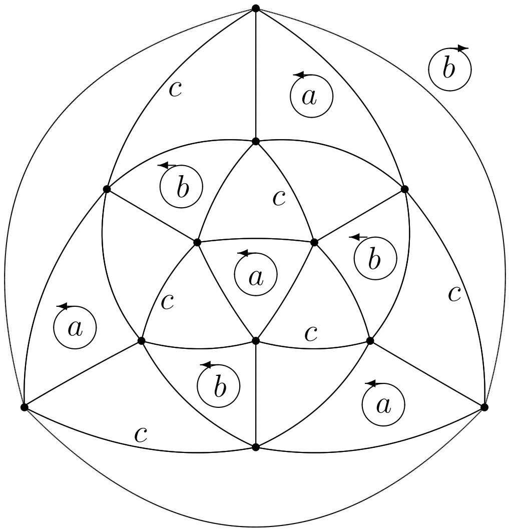
$$\begin{aligned} a &= (123) \\ b &= (124) \\ c = ab &= (14)(23) \end{aligned}$$

The two elements a and b are sufficient to generate A_4 as are the two elements a and c and many other pairs of elements (but not all pairs will do). In fact, A_4 can be represented in either of the following two ways:

$$(a, b : a^3 = b^3 = (ab)^2 = 1)$$
$$(a, c : a^3 = c^2 = (ac)^2 = 1)$$

So, if we have the Cayley graph with only a- and b-edges, then we have enough information to determine A_4, or if we have the graph with only a- and c-edges, then that's enough. Although these two graphs both have 12 vertices (since $|A_4| = 12$), they don't look very much alike. Let's look at the Cayley graph with all three kinds of edges, a-edges and b-edges and c-edges.

It's displayed here as a planar graph, but more of the symmetry would be apparent if it were displayed in three dimensions where the vertices and edges were those of an icosahedron. Some of the triangles in the figure are labelled a meaning that the three sides of the triangle are a-edges. The orientation on the circle surrounding a indicates the directions of the sides of the triangle. Likewise, some triangles are labelled b indicating that their sides are b-edges. Note that all the a- and b-triangles are oriented counterclockwise except the outer b-triangle. The remaining edges are the c-edges, and to save space, since c is an involution, rather than putting in two edges, one pointing one way and the other pointing the other way, just a single undirected edge is included. Each vertex in the graph has an a-edge coming in and one coming out, a b-edge coming in and one coming out, and an undirected c-edge meaning that it goes both in and out.



Since it only takes two of these three elements to generate A_4, this graph has superfluous information. All the edges labelled one of the letters can be removed making the graph simpler.

Exercise 4.9. Find a Cayley graph for the symmetric group S_4. There are various pairs or triples of generators you can use. One is the pair $a = (1234), b = (12)$.

4.3.2 Some small finite groups

We've seen a few families of finite groups including C_n the cyclic group of order n, D_n the dihedral group of order $2n$, S_n the symmetric group of order $n!$, and A_n the alternating group of order $n!/2$.

The classification of finite groups (up to isomorphism, of course) is extremely difficult, but we should look at a few more small finite groups. Later, we'll look at the classification of finite Abelian groups, and find that they're all products of cyclic groups.

Here's a table of the number of groups up to isomorphism of order up to 15.

order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
number of groups	1	1	1	2	1	2	1	5	2	2	1	5	1	2	1

We won't prove that these are all of them, but we will look at them all. There are combinatorial theorems, the most important being the Sylow theorems, that help in classifying finite groups.

We know nearly all of these 27 groups. The cyclic groups C_n account for 15 of them. There are 12 others. Some of them are products of smaller ones, for instance, the other group of order 4 is $C_2 \times C_2$, sometimes called the *Klein 4-group*.

The second group of order 6 is D_3, which is the same as S_3.

Two of the groups of order 8 are products, namely, $C_4 \times C_2$ and $C_2 \times C_2 \times \mathbf{C}_2$. Another is D_4 and the remaining one is called the quaternion group.

Example 4.22 (The quaternion group)**.** This group consists of eight of the units of the division ring $\mathbf{H}$, the quaternions. Let $G = \{\pm 1, \pm i, \pm j, \pm k\}$. Recall that the multiplication of quaternions has $i^2 = j^2 = k^2 = -1$, $ij = k$, $jk = i$, and $ki = j$, so this set of units is closed under multiplication and forms a group, called the *quaternion group*.

Exercise 4.10. Construct a Cayley graph for the quaternion group.

The second group of order 9 is $C_3 \times C_3$, and the second group of order 10 is D_5. We already know the other groups of order 12: D_6, $C_2 \times C_6$, $D_3 \times C_2$, and A_4, and the other group of order 14 is D_7.

4.4 Kernels, normal subgroups, and quotient groups

The kernel $\operatorname{Ker} f$ of a group homomorphism $f : G \to H$ plays the same role as the kernel of a ring homomorphism. It's defined as the the inverse image of the identity. It is a subgroup of the domain G, but a particular kind of subgroup called a normal subgroup. We'll see that every normal subgroup N of G is the kernel of some group homomorphism, in fact, of a projection $G \to G/N$ where G/N is a quotient group of G.

4.4.1 Kernels of group homomorphisms and normal subgroups

We'll use multiplicative notation.

Definition 4.23. Let $f : G \to H$ be a group homomorphism. Those elements of G that are sent to the identity 1 in H form the *kernel* of f.

$$\operatorname{Ker} f = f^{-1}(1) = \{x \in G \mid f(x) = 1\}.$$

Example 4.24. Let G be the symmetric group S_n and $f : G \to 1, -1$ map even permutations to 1 and odd permutations to -1. Then f is a group homomorphism, and $\operatorname{Ker} f = A_n$, the alternating subgroup of S_n.

Theorem 4.25. The kernel of a group homomorphism $f : G \to H$ is a subgroup $N = \operatorname{Ker} f$ of G such that for each $x \in G$, $xNx^{-1} \subseteq N$.

Proof. To show that N is a subgroup of G, note that (1) it's closed under multiplication, (2) it includes 1, and (3) it's closed under inverses. For (1), if $x, y \in N$, then $f(x) = f(y) = 1$, so $f(xy) = f(x)f(y) = 1$, therefore $xy \in N$. (2) is obvious. For (3), if $x \in N$, then $f(x) = 1$, so $f(x^{-1}) = f(x)^{-1} = 1^{-1} = 1$, therefore $x^{-1} \in N$.

Now to show that for $x \in G$, $xNx^{-1} \subseteq N$. Consider xyx^{-1} where $y \in N$. Then $f(y) = 1$, so $f(xyx^{-1}) = f(x)f(y)f(x)^{-1} = f(x)1f(x)^{-1} = f(x)f(x)^{-1} = 1$. Therefore, $xyx^{-1} \in N$. Thus, $xNx^{-1} \subseteq N$. Q.E.D.

Besides telling us what elements are sent to 1 by f, the kernel of f also tells us when two elements are sent to the same element. Since $f(x) = f(y)$ if and only if $f(xy^{-1}) = 1$, therefore, f will send x and y to the same element of S if and only if $xy^{-1} \in \operatorname{Ker} f$.

The properties of kernels of group homomorphisms that we just found determine the following definition.

Definition 4.26. A subgroup N of a group G is said to be a *normal subgroup* if for each $x \in G$, $xNx^{-1} \subseteq N$.

Exercise 4.11. Show that a subgroup N is normal in G if and only if for each $x \in G$, $xNx^{-1} = N$.

Exercise 4.12. Show that a subgroup N is normal in G if and only if for each $x \in G$, $xN = Nx$.

Both the trivial subgroup of G and G itself are always normal subgroups.

If G is an Abelian group, then every subgroup of G is a normal subgroup.

Theorem 4.27. Any subgroup of index 2 is a normal subgroup.

Proof. Let N be a subgroup of a group G of index 2. We'll show that $xN = Nx$ for each $x \in G$. In case $x \in N$, then $xN = N = Nx$. Now consider the case $x \notin N$. Then there are two left cosets of N, namely N itself and xN, and there are two right cosets, N and Nx. That gives us two partitions of G, but since N is a part of each partition, the other parts, namely xN and Nx must be equal. Q.E.D.

Even when a subgroup H is not normal in G, its conjugates are interesting.

Theorem 4.28. If H is a subgroup of G, and $x \in G$, then xHx^{-1} is also a subgroup of G, called a subgroup *conjugate* to H.

Proof. Clearly, $1 \in xHx^{-1}$. Given xyx^{-1} and xzx^{-1} with $y, z \in H$, then their product $xyx^{-1}xzx^{-1} = x(yz)x^{-1} \in xHx^{-1}$. Also, given xyx^{-1} with $y \in H$, then the inverse $(xyx^{-1})^{-1} = xy^{-1}x^{-1} \in xHx^{-1}$. Therefore, xHx^{-1} is a subgroup of G. Q.E.D.

Theorem 4.29. If no other subgroup of G has the same order as H, then H is normal.

Proof. Since any conjugate subgroup xHx^{-1} is in one-to-one correspondence with H, it has the same number of elements, so must equal H. Q.E.D.

Exercise 4.13. If H is a subgroup of G and N is a normal subgroup of G, prove that $H \cap N$ is a normal subgroup of H.

Exercise 4.14. If H is a subgroup of G and N is a normal subgroup of G, prove that HN is a subgroup of G. (Hint: show $HN = NH$.)

Exercise 4.15. Prove that the intersection of two normal subgroups is also a normal subgroup.

Exercise 4.16. Prove that if H and N are normal subgroups of G, then their product is also a normal subgroup of G, in fact, it's the subgroup generated by $H \cup N$.

4.4.2 Quandles and the operation of conjugation

Definition 4.30. If x and y are elements of a group, then $y^{-1}xy$ and yxy^{-1} are called *conjugates* of x. The set of all conjugates of an element x is called the *conjugacy class* of x.

Exercise 4.17. If x_1 is in the conjugacy class of x, prove that the conjugacy class of x_1 is the same as the conjugacy class of x.

Since a normal subgroup a group G is closed under conjugation, therefore a normal subgroup of G is the union of some of the conjugacy classes in G.

Example 4.31 (Conjugacy classes in symmetric groups)**.** Conjugation and conjugacy classes in symmetric groups are particularly easy to identify using cycle notation. Let $x = (13)(245)$ and $y = (142)$ be two elements in S_n. Then $y^{-1}xy = (124)(13)(245)(142) = (43)(125)$. Note how y conjugates the cycle (13) to the cycle (43), and it conjugates the cycle (245) to (125). The cycle structures for x and $y^{-1}xy$ are the same, but the elements in the cycles are permuted by y. This is generally the case for symmetric groups. It follows that a conjugacy class in S_n consists of all the elements in S_n with a given structure. Thus, for example, the conjugacy class of (13)(235) consists of all elements of the form $(ab)(cde)$ where a, b, c, d, and e are 5 distinct integers between 1 and n. For S_5 the size of that conjugacy class is $\binom{5}{2} \cdot 2 = 20$.

Exercise 4.18. Determine all the conjugacy classes of S_5 and their sizes. (The sum of their sizes will equal 120, of course.)

The operations of conjugation have certain properties. If we think of $y^{-1}xy$ as a binary operation $x \triangleright y$, and yx^{-1} as another operation $x \triangleright^{-1} y$, then these two operations satisfy the properties stated in the next definition.

Definition 4.32. A *quandle* is a set equipped with two operations, $\triangleright$ and $\triangleright^{-1}$ satisfying the following three conditions for all elements x, y, and z.

Q1. $x \triangleright x = x$.
Q2. $(x \triangleright y) \triangleright^{-1} y = x = (x \triangleright^{-1} y) \triangleright y$.
Q3. $(x \triangleright y) \triangleright z = (x \triangleright z) \triangleright (y \triangleright z)$.

The symbol $\triangleright$ is pronounced through, and $\triangleright^{-1}$ backthrough.

Exercise 4.19. Prove that if Q is a conjugacy class in a group G then Q is a quandle.

4.4.3 Quotients groups, and projections $\gamma : G \to G/N$

As mentioned above the kernel of a group homomorphism f tells us when two elements are sent to the same element: $f(x) = f(y)$ if and only if $xy^{-1} \in \operatorname{Ker} f$. We can use $\operatorname{Ker} f$ to construct a "quotient group" $G/\operatorname{Ker} f$ by identifying two elements x and y in G if xy^{-1} lies in $\operatorname{Ker} f$. In fact, we can do this not just for kernels of homomorphisms, but for any normal subgroup N. That is, we can use a normal subgroup N of G to determine when two elements x and y are to be identified, $x \equiv y$, and we'll end up with a group G/N.

Definition 4.33. A *congruence* $\equiv$ on a group G is an equivalence relation such that for all $x, x', y, y' \in G$,

$$x \equiv x' \text{ and } y \equiv y' \text{ imply } xy \equiv x'y'.$$

The equivalence classes for a congruence are called *congruence classes.*

Theorem 4.34. If $\equiv$ is a congruence on a group G, then the quotient set $G/_\equiv$, that is, the set of congruence classes, is a group where the binary operation is defined by $[x][y] = [xy]$.

Proof. First we need to show that the proposed definitions are actually well defined. That is, if a different representative x' is chosen from the congruence class $[x]$ and y' from $[y]$, then the same class $[x'y']$ results. That is

$$[x] = [x'] \text{ and } [y] = [y'] \text{ imply } [xy = xy'].$$

But that is the requirement in the definition of congruence.

Also, each of the axioms for a group need to be verified, but they're all automatic as they're inherited from the group G. Q.E.D.

Just as an ideal in a ring determines a congruence on the ring, a normal subgroup of a group determines a congruence on a group, and the proof is similar.

Theorem 4.35 (Congruence modulo a normal subgroup)**.** Let N be a normal subgroup of a group G. A congruence, called *congruence modulo* N, is defined by

$$x \equiv y \pmod{N} \text{ if and only if } xy^{-1} \in N.$$

The quotient group, $G/_\equiv$, is denoted G/N. The congruence classes are cosets of N, that is $[x] = xN$. The function $\gamma : G \to G/N$ defined by $\gamma(x) = [x] = xN$ is a group homomorphism, in fact, an epimorphism. It's called a *projection* or a *canonical homomorphism* to the quotient group. It's kernel is N.

Exercise 4.20. If $\equiv$ is a congruence on a group G, show that the congruence class of the identity, $[1] = N$, is a normal subgroup of G, and the congruence determined by N is the original congruence.

4.4.4 Isomorphism theorems

The image of a group homomorphism is isomorphic to the group modulo its kernel. Let $f : G \to H$ be a ring homomorphism. The image of f, denoted $f(G)$, is the set

$$f(G) = \{f(x) \in H \mid x \in G\}.$$

Exercise 4.21. Verify that the image $f(G)$ is a subgroup of H.

Exercise 4.22. Prove the following theorem. You'll need to show that the proposed function is well-defined, that it is a group homomorphism, and then that it's an isomorphism.

Theorem 4.36 (First isomorphism theorem)**.** If $f : G \to H$ is a group homomorphism then the quotient group $G/\operatorname{Ker} f$ is isomorphic to the image ring $f(G)$, the isomorphism being given by

$$\begin{array}{rcl} G/\operatorname{Ker} f & \to & f(G) \\ x \operatorname{Ker} f & \mapsto & f(x) \end{array}$$

This gives us two ways to look at the image, either as a quotient group of the domain G or as a subgroup of the codomain H.

Furthermore, we can now treat a group homomorphism $f : G \to H$ as a composition of three group homomorphisms.

$$G \xrightarrow{\gamma} G/\operatorname{Ker} f \cong f(G) \xrightarrow{\iota} H$$

The first is the projection from G onto its quotient ring $G/\operatorname{Ker} f$, the second is the isomorphism $G/\operatorname{Ker} f \cong f(G)$, and the third is the inclusion of the image $f(G)$ as a subgroup of H.

Theorem 4.37 (Second isomorphism theorem)**.** If H is a subgroup of G and N is a normal subgroup of G, then

$$H/(H \cap N) \cong (HN)/N.$$

Proof. Let $f : H \to (HN)/N$ be defined by $f(x) = xN$. This f is a group homomorphism since $f(xy) = xyN = xNyN = f(x)f(y)$.

Next, we'll show that f is an epimorphism. Let $xN \in (HN)/N$ where $x \in HN$. Then $x = yz$ for some $y \in H$ and $z \in N$. So $xN = yzN = yN = f(y)$. Thus, f is an epimorphism, that is, $f(H) = (HN)/N$. by the first isomorphism theorem, we have

$$H/\operatorname{Ker} f \cong (HN)/N.$$

Finally, we'll show that $\operatorname{Ker} f = H \cap K$ which will imply $H/(H \cap N) \cong (HN)/N$. Let x be an element of H which lies in $\operatorname{Ker} f$. Then xN is the identity element N in $(HN)/N$, so $x \in N$. But $x \in H$ also, so $x \in H \cap N$. Conversely, $x \in H \cap N$ implies $x \in \operatorname{Ker} f$. Q.E.D.

Theorem 4.38 (Third isomorphism theorem)**.** If H and K are both normal subgroups of G with $H \subseteq K$, then

$$(G/H)/(K/H) \cong G/K.$$

Exercise 4.23. Prove the third isomorphism theorem. Define $f : G/H \to G/K$ by $f(aH) = aK$. Check that this is a well-defined homomorphism. Show $\operatorname{Ker} f = H$. Show the image of f is all of G/K. Apply the first isomorphism theorem to finish the proof.

Theorem 4.39 (Correspondence theorem)**.** Let N be a normal subgroup of G. The subgroups of G containing N are in one-to-one correspondence with the subgroups of G/N. Thus, if H is a subgroup of G containing N, then H/N is a subgroup of G/N, and every subgroup of G/N so arises. Furthermore, H is normal in G if and only if H/N is normal in G/N.

Exercise 4.24. Prove the correspondence theorem. Show that for $H \supseteq N$ that H/N is, indeed, a subgroup of G/N. Show that if $\overline{H}$ is any subgroup of G/N that the set $H = \{x \in G \,|\, x/N \in \overline{H}\}$ is a subgroup of G containing N. Verify that these two operations are inverse to each other. Finally, verify the last statement.

4.4.5 Internal direct products

We can recognize when a group G is isomorphic to a product of two or more groups. Recall that if $G = M \times N$, then we can interpret M and N as subgroups of G. As such they are normal subgroups of G and their intersection is trivial. Furthermore, $G = MN$.

Definition 4.40. A group G is said to be an *internal direct product* of two subgroups M and N if $M \cap N = 1$, $MN = G$, and both M and N are normal subgroups of G.

We'll show in a moment that if G is the internal direct product of M and N, then G is isomorphic to the product group $M \times N$. But first, a lemma.

Lemma 4.41. If M and N are two normal subgroups of G whose intersection is trivial, then elements of M commute with elements of N.

Proof. Let $m \in M$ and $n \in N$. In order to show that $mn = nm$, we'll show the equivalent $mnm^{-1}n^{-1} = 1$. Let $x = mnm^{-1}n^{-1}$. Since $x = (mnm^{-1})n^{-1}$, and both mnm^{-1} and n^{-1} are elements of the normal subgroup N, therefore $x \in N$. But since $x = m(nm^{-1}n^{-1})$, and both m and $nm^{-1}n^{-1}$ are elements of the normal subgroup M, therefore $x \in M$. Since $x \in M \cap N = 1$, therefore $x = 1$. Q.E.D.

Theorem 4.42. If G is the internal direct product of M and N, then $M \times N \cong G$ where the isomorphism is given by $(m, n) \mapsto mn$.

Proof. Outline. Use the lemma to verify that the proposed isomorphism is a homomorphism. It's evidently a surjection since $MN = G$. To show that it's an injection, show that the kernel is trivial. Suppose $(m, n) \mapsto mn = 1$. Then $m = n^{-1}$ lies in both M and N, so it's trivial, that is, $m = n = 1$. Q.E.D.

Exercise 4.25. Prove that G is an internal direct product of two normal subgroups M and N if and only if every element $x \in G$ can be uniquely represented as a product mn with $m \in M$ and $n \in N$.

Although we've only looked at internal direct products of two subgroups, the definition can be generalized to more than two subgroups. We'll say that G is the *internal direct product* of r normal subgroups $N_1, N_2, \ldots, N_r$ if (1) they jointly generate G, that is, $N_1 N_2 \cdots N_r = G$, and (2) the intersection of any one N_i with the subgroup generated by the rest is trivial. It follows that $N_1 \times N_2 \times \cdots \times N_r \cong G$. Furthermore, an equivalent condition to being a internal direct product of the normal subgroups $N_1, N_2, \ldots, N_r$ is that every element $x \in G$ can be uniquely represented as a product $n_1 n_2 \cdots n_r$ with each $n_i \in N_i$.

4.5 Matrix rings and linear groups

The representation of rings and groups as subrings or subgroups of matrix rings is very helpful for a couple of reasons. One is that matrices describe linear transformations. That means that the elements of the ring or group can be interpreted as geometric transformations. A second is that matrix notation is so very convenient. Usually the coefficients are taken to be elements of a familiar field like $\mathbf{C}$, $\mathbf{R}$, or $\mathbf{Q}$, but for special purposes the coefficients may be taken in some other integral domain such as $\mathbf{Z}$.

For example, the field complex numbers $\mathbf{C}$ can be represented as a certain subring of $M_2(\mathbf{R})$, the ring of 2×2 matrices with coefficients in $\mathbf{R}$, and the division ring of quaternions $\mathbf{H}$ can be represented as a certain subring of $M_4(\mathbf{R})$.

Most of our examples have n equal to 2 or 3 and the coefficients are real.

4.5.1 Linear transformations

The ring of $n \times n$ matrices with real coefficients, $M_2(\mathbf{R})$, is a noncommutative ring when $n \geq 2$. We can interpret each matrix $A \in M_2(\mathbf{R})$ as a linear transformation $A : \mathbf{R}^n \to \mathbf{R}^n$ where a (column) n-vector $\mathbf{x} \in \mathbf{R}^n$ is mapped to another n-vector

$$A\mathbf{x} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{bmatrix}$$

The identity matrix

$$I = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}$$

corresponds to the identity transformation $I : \mathbf{R}^n \to \mathbf{R}^n$ where $I\mathbf{x} = \mathbf{x}$.

4.5.2 The general linear groups $GL_n(R)$

The invertible $n \times n$ matrices in $M_n(R)$, that is, the units in the ring $M_n(R)$, form the *general linear group* with coefficients in the commutative ring R, denoted $GL_n(R)$. They describe nonsingular transformations $R^n \to R^n$. Recall that a matrix A has an inverse if and only if its determinant $|A|$ is a unit in R.

Let's interpret some of these in the case when $n = 2$. The determinant of $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $|A| = ad - bc$, and when that's a unit in R, the inverse of A is $A = \dfrac{1}{|A|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

Let's let R be the field of real numbers $\mathbf{R}$. The real general linear group $GL_2(\mathbf{R})$ can be interpreted as the group of linear transformations of the plane $\mathbf{R}^2$ that leave the origin fixed. Here are a few linear transformations of the plane.

Rotation by an angle θ about the origin is described by the matrix

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

since a point $\begin{bmatrix} x \\ y \end{bmatrix}$ in $\mathbf{R}^2$ is sent to the point $\begin{bmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{bmatrix}$ The determinant of a rotation matrix is 1.

Reflection across a line through the origin at an angle θ to the x-axis is described by the matrix

$$\begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}.$$

The determinant is -1.

Expansions and contractions are described by scalar matrices $\begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}$ where r is the ratio. If $r > 1$, then it's an expansion (also called dilation), but if $0 < r < 1$, then it's a contraction.

There are numerous other kinds of transformations. Here's just one more example $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, an example of a shear parallel to the x-axis. Points above the x-axis are moved right, points below left, and points on the x-axis are fixed.

In three dimensions you can describe rotations, reflections, and so forth, as well.

4.5.3 Other linear groups

There are a number of interesting subgroups of $GL_n(R)$.

The special linear groups $SL_n(R)$. There are several subgroups of $GL_n(R)$, one of which is the special linear group $SL_n(R)$ which consists of matrices whose determinants equal 1, also called *unimodular* matrices. (There are other linear groups called "special" and in each case it means the determinant is 1.) Among the examples in $GL_2(\mathbf{R})$ mentioned above, the rotations and shears are members of $SL_2(\mathbf{R})$, but reflections have determinant -1 and expansions and contractions have determinants greater or less than 1, so none of them belong to the special linear group.

Since the absolute value of the determinant is the Jacobian of the transformation $\mathbf{R}^n \to \mathbf{R}^n$, therefore transformations in $SL_2(\mathbf{R})$ preserve area. Since the determinant is positive, these transformations preserve orientation. Thus, transformations in $SL_2(\mathbf{R})$ are the linear transformations that preserve orientation and area. More generally those in $SL_n(\mathbf{R})$ preserve orientation and n-dimensional content. Rotations and shears, and their products, are always in $SL_2(\mathbf{R})$.

The orthogonal groups $\mathcal{O}(n)$. These are subgroups of $GL_n(\mathbf{R})$, An *orthogonal* transformation is one that preserves inner products (also called dot products or scalar products). I'll use the notation

$$\langle \mathbf{a}, \mathbf{b} \rangle = a_1 b_1 + a_2 b_2 + \cdots a_n b_n$$

for the inner product of the vectors $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$. Other common notations are $(\mathbf{a}, \mathbf{b})$ or $\mathbf{a} \cdot \mathbf{b}$. For the transformation described by the matrix A to preserve inner products means that $\langle A\mathbf{a}, A\mathbf{b} \rangle = \langle \mathbf{a}, \mathbf{b} \rangle$. Since the length of a vector $|\mathbf{a}|$ is determined by the inner product, $|\mathbf{a}|^2 = \langle \mathbf{a}, \mathbf{a} \rangle$, therefore an orthogonal transformation preserves distance, too: $|A\mathbf{a}| = |\mathbf{a}|$. Conversely, if A preserves distance, it preserves inner products.

Note that since distance is preserved, so is area in dimension 2 or n-dimensional content in dimension n.

It's a theorem from linear algebra that a matrix A describes an orthogonal transformation if and only if its inverse equals its transform: $A^{-1} = A^T$; equivalently, $AA^T = 1$. These matrices, of course, are called *orthogonal* matrices. Note that the determinant of an orthogonal matrix is ± 1.

The orthogonal group $\mathcal{O}(n)$ is the subgroup of $GL_n(\mathbf{R})$ of orthogonal matrices. It's not a subgroup of $SL_n(\mathbf{R})$ since half the orthogonal matrices have determinant -1, meaning they reverse orientation. The special orthogonal group $S\mathcal{O}(n)$ is the subgroup of $\mathcal{O}(n)$ of matrices with determinant 1.

In two dimensions $\mathcal{O}(2)$ consists of rotations and reflections while $S\mathcal{O}(n)$ consists of only the rotations. In three dimensions $\mathcal{O}(3)$ consists of rotations (by some angle around some line through 0) and reflections (across some plane through 0). Again, $S\mathcal{O}(3)$ only has the rotations.

The unitary groups $\mathcal{U}(n)$. For matrices with complex coefficients, the most useful analogous group corresponding to the orthogonal group for real coefficients is something called a unitary group.

The inner product, also called the *Hermitian*, for the complex vector space $\mathbf{C}^n$ is defined as

$$\langle \mathbf{a}, \mathbf{b} \rangle = a_1 \overline{b}_1 + a_2 \overline{b}_2 + \cdots a_n \overline{b}_n$$

for the complex vectors $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ where the bar indicates complex conjugation. A matrix A, and the transformation $\mathbf{C}^n \to \mathbf{C}^n$ that it describes, are called *unitary* if it preserves the Hermitian. The collection of all unitary matrices in $GL_n(\mathbf{C})$ is called the unitary group $\mathcal{U}(n)$.

Another theorem from linear algebra is that a matrix A is unitary if and only if its inverse is the transform of its conjugate, $A^{-1} = \overline{A}^T$, equivalently, $A\overline{A}^T = I$.

There are many properties of complex unitary matrices that correspond to properties of real orthogonal matrices.

4.5.4 Projective space and the projective linear groups $PSL_n(F)$

Let F be a field, such as the field of real numbers. The projective linear group $PSL_n(F)$ is used to study projective space.

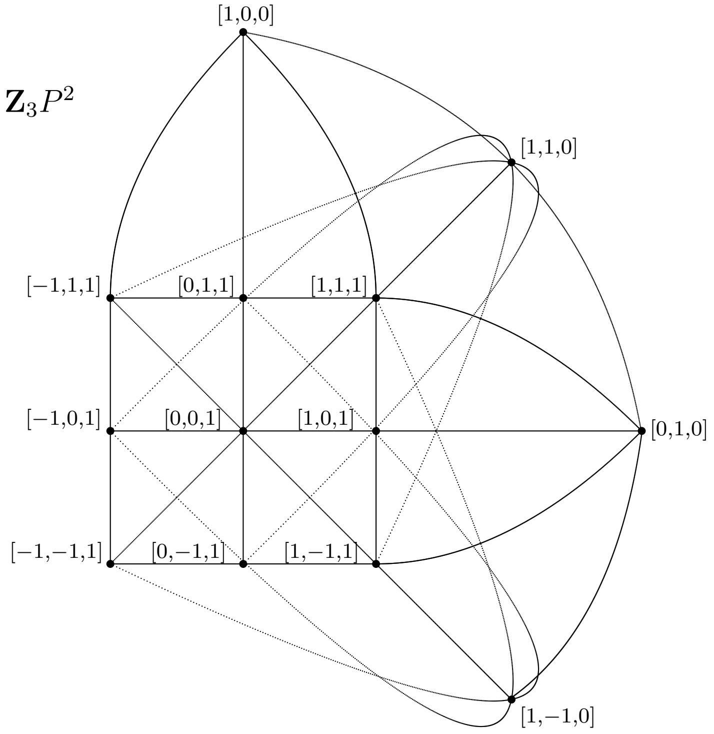
Projective space FP^n of dimension n is defined from affine space F^{n+1} of dimension $n+1$ as by means of an equivalence relation. Two points $\mathbf{a} = (a_0, a_1, \ldots, a_n)$ and $\mathbf{b} = (b_0, b_1, \ldots, b_n)$ of F^{n+1} name the same point of FP^n if their coordinates are proportional, that is, if there exists a nonzero element $\lambda \in F$ such that $b_i/a_i = \lambda$ for $i = 0, 1, \ldots, n$. We'll let $[a_0, a_1, \ldots, a_n]$ denote the point in FP^n named by $(a_0, a_1, \ldots, a_n) \in F^{n+1}$. Thus, $[a_0, a_1, \ldots, a_n] = [\lambda a_0, \lambda a_1, \ldots, \lambda a_n]$. The notation $[a_0, a_1, \ldots, a_n]$ is called *projective coordinates.*

Geometrically, this construction adds points at infinity to the affine plane, one point for each set of parallel lines.

Lines can also be named with projective coordinates $\mathbf{b} = [b_0, b_1, \ldots, b_n]$. If you do that, then a point $\mathbf{a} = [a_0, a_1, \ldots, a_n]$ lies on the line $\mathbf{b}$ if their inner product $\langle \mathbf{a}, \mathbf{b} \rangle$ is 0.

Here's one representation of the projective plane $\mathbf{Z}_3P^2$. There are 13 points and 13 lines, each line with 4 points, and each point on 4 lines.

We can name the 9 points in the affine plane $\mathbf{Z}_3^2$ with third coordinate 1, and the 4 points at infinity with third coordinate 0. The four points at infinity line on a line at infinity. Each of these points at infinity lie on all those line with a particular slope. For instance, the point $[1, -1, 0]$ lies on the three lines with slope -1 (and it lies on the line at infinity, too).



Similarly, we can take a quotient of $GL_{n+1}(F)$ as the projective linear group $PGL_n(F)$. Two matrices A and B in $GL_{n+1}(F)$ name the same element of $PGL_n(F)$ if each is a multiple of the other, that is, there exists $\lambda \neq 0 \in F$ such that $B = \lambda A$. Then $PGL_n(F)$ acts on FP^n, since $A\mathbf{a}$ and $\lambda A\mathbf{a}$ name the same element of FP^n.

The group $PGL_3(\mathbf{Z}_3)$ acts on the projective plane $\mathbf{Z}_3P^2$. It has $13 \cdot 12 \cdot 9 \cdot 4 = 5616$ elements.

The projective special linear group $PSL_n(F)$ is the subgroup of $PGL_n(F)$ named by unimodular matrices. It's $SL_n(F)$ modulo scalar matrices ωI where ω is an nth root of unity. Except for small values of n the projective special linear groups are all simple. Simplicity is defined in the next section.

The group $PSL_3(\mathbf{Z}_3)$ is actually the same as $PGL_3(\mathbf{Z}_3)$.

4.6 Structure of finite groups

The classification of finite groups is extremely difficult, but there are a tools we can use to see how that classification begins. In the next section we'll classify finite Abelian groups and see that they're isomorphic to products of cyclic groups, but the situation for general groups much more complicated.

4.6.1 Simple groups

The way we'll analyze groups is by their normal subgroups and quotients. In particular, if N is a maximal, proper normal subgroup of G, then G/N has no subgroups, for if it did, by the correspondence theorem, there would be a normal subgroup between N and G.

Definition 4.43. A nontrivial group is said to be *simple* if it has no proper, nontrivial, normal subgroups.

Exercise 4.26. Prove that the only Abelian simple groups are cyclic of prime order.

There are many nonabelian simple groups. There are several infinite families of them, and a few that aren't in infinite families, called *sporadic* simple groups. One infinite family of simple groups consists of alternating groups A_n with $n \geq 5$. Indeed, A_5 is the smallest nonabelian simple group. The projective special linear groups mentioned in the section above form another family of finite simple groups.

Exercise 4.27 (Nonsimplicity of A_4). Verify that there are five conjugacy classes in A_4 as shown in the following table.

Generator	Size	Order
1	1	1
(12)(34)	3	2
(123)	4	3
(132)	4	3

A normal subgroup of A_4 would be a union of some of these conjugacy classes including the identity conjugacy class of size 1, but its order would have to divide 12. Find all the proper nontrivial normal subgroups of A_4.

Exercise 4.28 (Simplicity of A_5). Verify that there are five conjugacy classes in A_5 as shown in the following table.

Generator	Size	Order
1	1	1
(12)(34)	15	2
(123)	20	3
(12345)	24	5
(12354)	24	5

A normal subgroup of A_5 would be a union of some of these conjugacy classes including the identity conjugacy class of size 1, but its order would have to divide 60. Verify that no combination of the numbers 1, 15, 20, 24, and 24, where 1 is included in the the combination, yields a sum that divides 60 except just 1 itself and the sum of all five numbers. Thus, there is no proper nontrivial normal subgroup of A_5.

4.6.2 The Jordan-Hölder theorem

Definition 4.44. A *composition series* for a group G is a finite chain of subgroups

$$1 = N_n \subseteq N_{n-1} \subseteq \cdots \subseteq N_1 \subseteq N_0 = G$$

such that each N_{i-1} is a maximal proper normal subgroup of N_i. The number n is called the *length* of the composition series, and the n quotient groups

$$N_{n-1}/1, \ldots, N_1/N_2, G/N_1$$

which are all a simple groups, are called *composition factors* determined by the composition series.

It is evident that any finite group G has at least one composition series. Just take N_1 to be a maximal proper normal subgroup of G, N_1 to bee a maximal proper normal subgroup of N_1, etc. Infinite groups may also have composition series, but not all infinite groups do.

Exercise 4.29. Find a composition series for the symmetric group S_4.

Exercise 4.30. Prove that an infinite cyclic group has no (finite) composition series.

Although a finite group may have more than one composition series, the length of the series is determined by the group as are composition factors at least up to isomorphism as we'll see in a moment. Thus, these are invariants of the group. They do not, however, completely determine the group.

Exercise 4.31. Show that the dihedral group D_5 and the cyclic group C_{10} have composition series with the same length and same factors.

Theorem 4.45 (Jordan-Hölder)**.** Any two composition series for a finite group have the same length and there is a one-to-one correspondence between the composition factors of the two composition series for which the corresponding composition factors are isomorphic.

Proof. We'll prove this by induction on the order of the group under question. The base case is for the trivial group which has only the trivial composition series.

Assume now that a group G has two composition series

$$1 = N_m \subseteq M_{m-1} \subseteq \cdots \subseteq M_1 \subseteq M_0 = G, \text{ and } 1 = N_n \subseteq N_{n-1} \subseteq \cdots \subseteq N_1 \subseteq N_0 = G$$

If $M_1 = N_1$, then by induction we conclude that the lengths of the rest of the composition are equal and the composition factors the rest of the rest of the series are the same, and of course, the factors G/M_1 and G/N_1 are equal, so the case $M_1 = N_1$ is finished.

Consider now the case $M_1 \neq N_1$. Since both M_1 and N_1 are normal subgroups of G, so is their intersection $K_2 = M_1 \cap N_1$. Let $1 = K_k \subseteq K_{k-1} \subseteq \cdots \subseteq K_3 \subseteq K_2$ be a composition series for their intersection. These subgroups of G are illustrated in the following diagram.

$$\begin{tikzcd}
 & M_{m-1} \arrow[r, dash] & \cdots \arrow[r, dash] & M_2 \arrow[r, dash] & M_1 \arrow[dr, dash] & \\
1 \arrow[ur, dash, bend left] \arrow[r, dash] \arrow[dr, dash, bend right] & K_{k-1} \arrow[r, dash] & \cdots \arrow[r, dash] & K_2 \arrow[ur, dash] \arrow[dr, dash] & & G \\
 & N_{n-1} \arrow[r, dash] & \cdots \arrow[r, dash] & N_2 \arrow[r, dash] & N_1 \arrow[ur, dash] &
\end{tikzcd}$$

By the second isomorphism theorem, we have $M_1/(M_1 \cap N_1) \cong G/N_1$. Therefore, K_2 is a maximal normal subgroup of M_1. Thus, we have two composition series for M_1, and by the inductive hypothesis, they have the same length, so $m = k$, and they have the same factors up to isomorphism in some order. Likewise we have two composition series for N_1, and they have the same length, so $k = n$, and the same factors up to isomorphism in some order. We now have four composition series for G, two including M_1 and two including N_1. They all have the same length, and since $G/M_1 \cong N_1/K_2$ and $G/N_1 \cong M_1/K_2$, they all have the same factors up to isomorphism in some order. Q.E.D.

There is a generalization of this theorem that applies to infinite groups that have composition series but its proof is considerably longer.

Solvable groups One of the applications of group theory is Galois' theory for algebraic fields. The groups of automorphisms of these fields are closely related to the solutions of algebraic equations. In particular, these groups can tell you if the equations have solutions that can be expressed in terms of radicals, that is square roots, cube roots, and higher roots. The condition for such solvability is none the factors in a composition series for a group are nonabelian simple groups, equivalently, that all the factors are cyclic groups of prime order.

Definition 4.46. A group is said to be *solvable* if it has a composition series all of whose factors are cyclic.

Exercise 4.32. Prove that if the order of a group is a power of a prime number, then that group is solvable.

Much more can be said about solvable groups than we have time for.

4.7 Abelian groups

We'll use additive notation throughout this section on Abelian groups. Also, we'll call the product of two Abelian groups A and B a direct sum and denote it $A \oplus B$.

We already know a fair amount about Abelian groups. We know about cyclic groups and the Chinese remainder theorem.

Every subgroup of an Abelian group is normal, so we'll just refer to them as subgroups and leave off the adjective "normal."

Our characterization of internal direct product looks a little different when the group is written additively. Here it is, rewritten for Abelian groups.

An Abelian group G is the *internal direct sum* of subgroups M and N if (1) they jointly generate G, that is, $M + N = G$, and (2) the intersection $M \cap N = 0$. If G is the internal direct sum of M and N, then $M \oplus N = G$. Furthermore, an equivalent condition to being a internal direct sum is that every element $x \in G$ can be uniquely represented as a sum $m + n$ with $m \in M$ and $n \in N$.

4.7.1 The category $\mathcal{A}$ of Abelian groups

The category of Abelian groups is a particularly nice category. Not only does it have products, but it also has coproducts, to be defined next, and the products are coproducts, and that's why we're calling them direct sums. It's not the only category with direct sums. The category of vector spaces over a fixed field has them too.

Coproducts in a category and their universal property When all the arrows in a diagram are reversed, a similar diagram, called the *dual* results. Recall that products in a category are characterized by a diagram.

The product $A \times B$ in a category along with the two projections $A \times B \xrightarrow{\pi_1} A$ and $A \times B \xrightarrow{\pi_2} B$ has the universal property that for each object X and morphisms $X \to A$ and $X \to B$, there is a unique morphism $X \to A \times B$, such that the diagram below commutes.

$$\begin{tikzcd}
& & A \\
X \arrow[urr] \arrow[r] \arrow[drr] & A \times B \arrow[ur, "\pi_1"] \arrow[dr, "\pi_2"] & \\
& & B
\end{tikzcd}$$

If we turn around all the arrows, we'll get the characterizing property for coproducts. The coproduct $A \coprod B$ in a category along with the two injections $A \xrightarrow{\gamma_1} A \coprod B$ and $B \xrightarrow{\gamma_1} A \coprod B$ has the universal property that for each object X and morphisms $A \to X$ and $B \to X$, there is a unique morphism $A \coprod B \to X$, such that the diagram below commutes.

$$\begin{tikzcd}
A \arrow[dr, "\gamma_1"'] \arrow[drr] & & \\
& A\coprod B \arrow[r] & X \\
B \arrow[ur, "\gamma_2"] \arrow[urr] & &
\end{tikzcd}$$

Exercise 4.33. In the category of Abelian groups, the coproduct object $A\coprod B$ is what we've called the direct sum $A \oplus B$, which is the same as the product $A \times B$. The injections $A \xrightarrow{\gamma_1} A\coprod B$ and $B \xrightarrow{\gamma_1} A\coprod B$ for Abelian groups are defined by $\gamma_1(x) = (x, 0)$ and $\gamma_1(y) = (0, y)$. Verify that the universal property holds.

4.7.2 Finite Abelian groups

The classification of finite groups is very difficult, but the classification of finite Abelian is not so difficult. It turns out, as we'll see, that a fine Abelian group is isomorphic to a product of cyclic groups, and there's a certain uniqueness to this representation. The theorem above on internal direct sums is essential in this classification.

Theorem 4.47. Let G be a finite Abelian group of order mn where m and n are relatively prime, both greater than 1. Let $M = \{x \in G \,|\, mx = 0\}$ and $N = \{x \in G \,|\, nx = 0\}$. Then M and N are subgroups of G, and G is the internal direct sum of M and N. Furthermore, $|M| = m$ and $|N| = n$.

Proof. Outline. That M and N are subgroups is quickly verified. Since m and n are relatively prime, therefore 1 is a linear combination of them, that is, there are integers s and t such that $1 = sm + tn$. Their intersection $M \cap N$ is trivial since if $x \in M \cap N$, then $mx = nx = 0$, hence $x = 1x = (sm + tn)x = smx + tnx = 0$. Together M and N generate G, since for $x \in G$, $x = smx + tnx$, but $smx \in N$ since $nsmx = (nm)sx = 0$, likewise $tnx \in M$. Thus $M + N = G$. Therefore, G is the internal direct sum of M and N. Q.E.D.

Let G be a Abelian group and p a prime number. The set

$$G(p) = \{x \,|\, p^k x = 0 \text{ for some } k \geq 0\}$$

is a subgroup of G. It is called the *p-primary component* of G.

As a corollary to the above theorem consider the case when $|G|$ is factored as a power of primes.

Corollary 4.48 (Primary decomposition theorem)**.** Let G be a finite Abelian group whose order has prime factorization $p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$. Then G is a direct sum of the p_i-primary components

$$G \cong G(p_1) \oplus G(p_2) \oplus \cdots \oplus G(p_r)$$

and $|G(p_i)| = p_i^{e_i}$ for each i.

We've reduced the problem of classifying finite Abelian groups to classifying those whose orders are powers of a prime p. Such groups are called *p-primary groups* or simply *p-groups.* If the power is greater than 1, then there are different groups of that order. For example, there are three distinct Abelian groups of order 125, namely, $\mathbf{Z}_{125}$, $\mathbf{Z}_{25} \oplus \mathbf{Z}_5$ and $\mathbf{Z}_5 \oplus \mathbf{Z}_5 \oplus \mathbf{Z}_5$. The first has an element of order 125, but the other two don't, while the second has an element of order 25, but the third doesn't. Hence, they are not isomorphic.

Our strategy for a p-primary group will be to pick off direct summands containing elements of maximal orders, one at a time. That will show that a p-primary group is a direct sum of cyclic groups whose orders are nonincreasing powers of p. We'll then show those powers of p are determined by the p-primary group.

A difficulty in the proof is that there are many choices to be made resulting in different direct sums, but we'll see that the orders of the cyclic subgroups turns out to be the same no matter how we make the choices.

The proof of the theorem is particularly technical, so we'll separate parts of the proof as lemmas.

Lemma 4.49. Let G be a noncyclic p-primary group and a an element of G of maximal order. Then there is an element b in the complement of $\langle a\rangle$ of order p.

Proof. Let c be an element in the complement of $\langle a\rangle$ of smallest order. Since the order of pc is $\frac{1}{p}$ times the order of c, which is a smaller order than the order of c, therefore pc lies in $\langle a\rangle$. So $pc = ka$ for some integer k. Let p^m denote the ord a, the largest order of any element in G. Then $\operatorname{ord}(ka) \leq p^{m-1}$ since $p^{m-1}(ka) = p^{m-1}pc = p^m c = 0$. Therefore, ka is not a generator of the cyclic group $\langle a\rangle$ since that group has p^m elements. Hence, $\text{GCD}(p^m, k) \neq 1$, and so p divides k. Let $k = pj$. Then $pb = ka = pji$. Let $b = c - ja$. Then $pb = 0$, but $b \notin \langle a\rangle$ as $c = b + ka \notin \langle a\rangle$. Q.E.D.

Proof. Let $|G| = p^n$ and $\operatorname{ord} a = p^m$ with $m < n$.

We'll prove the lemma by induction. Assume it is valid for all groups of order less than p^n. Let b be an element in the complement of $\langle a\rangle$ of order p shown to exist in the previous lemma. Since $\operatorname{ord} b = p$ and $\notin \langle a\rangle$, therefore $\notin \langle a\rangle \cap \notin \langle b\rangle = 0$.

We'll reduce modulo $\langle b\rangle$ to a smaller p-primary group $G/\langle b\rangle$ where we can use the inductive hypothesis, then bring the results back up to G.

First, we'll show that $a + \langle b\rangle$, which is the image of a in $G/\langle b\rangle$, has the same order that a does in G, namely p^m, which implies that $a + \langle b\rangle$ is an element of maximal order in the group $G/\langle b\rangle$. Suppose $\operatorname{ord}(a + \langle b\rangle) < p^m$. Then $p^{m-1}(a + \langle b\rangle)$ is the 0 element of $G/\langle b\rangle$, in other words, $p^{m-1}a \in \langle b\rangle$. But $p^{m-1}a \in \langle a\rangle$, and the intersection of $\langle a\rangle$ and $\langle b\rangle$ is trivial. Therefore, $p^{m-1}a = 0$ which contradicts $\operatorname{ord} a = p^m$.

We now know $a + \langle b\rangle$ is an element of maximal order in the group $G/\langle b\rangle$, so we can apply the inductive hypothesis to conclude that $G/\langle b\rangle$ is the direct sum of the cyclic subgroup generated by $a + \langle b\rangle$ and another subgroup $K/\langle b\rangle$. Note that by the correspondence theorem, every subgroup of a quotient group $G/\langle b\rangle$ is the image of a group in G, so we may take K to be a subgroup of G.

We'll show that $G = \langle a\rangle \oplus K$ by showing that (1) $\langle a\rangle \cap K = 0$, and (2) $\langle a\rangle K = G$.

(1). If $x \in \langle a \rangle \cap K$, then its image $x + \langle b \rangle$ in the quotient group $G/\langle b \rangle$ lies in both the cyclic subgroup generated by $a + \langle b \rangle$ and $K/\langle b \rangle$. But their intersection is the 0 element in $G/\langle b \rangle$, therefore $x \in \langle b \rangle$. Since $x \in \langle a \rangle$ also, and $x \in \langle a \rangle \cap \langle b \rangle$ is trivial, therefore $x = 0$.

(2). We can show $\langle a \rangle K$ is all of G by a counting argument. We know that the order of $G/\langle b \rangle$ is the product of the order of the cyclic subgroup generated by $a + \langle b \rangle$ and the order of $K/\langle b \rangle$, the order of G is p times the order of $G/\langle b \rangle$, the order of $\langle a \rangle$ is the same as the order of the cyclic subgroup generated by $a + \langle b \rangle$, and the order of K is p times the order of $K \langle b \rangle$. Therefore, the order of G equals the product of the order of $\langle a \rangle$ and the order of K. Thus $\langle a \rangle K = G$. Q.E.D.

You can prove the first statement of following theorem by induction using the lemma we just proved, then apply the primary decomposition theorem for the second statement. This is the existence half of the theorem we want. We'll still need some kind of uniqueness of the terms in the direct sum.

Theorem 4.50. *A p-primary group is a direct sum of cyclic groups whose orders are powers of p. A finite Abelian group is the direct sum of cyclic groups.*

There are a couple of ways to describe the uniqueness of the terms. Since we've been using cyclic groups whose orders are prime powers, let's stick to that.

There's a concept we'll need in the following lemma. If G is an Abelian group and p an integer, then the subset $G^p = \{x \mid px = 0\}$ is a subgroup of G. In fact, it's just the kernel of the group homomorphism $G \to G$ that maps x to px.

Exercise 4.34. Show that it is, indeed, a group homomorphism.

Lemma 4.51. *Suppose that G is a p-primary group that can be written as a direct sum of nontrivial cyclic subgroups in two ways*

$$G = H_1 \oplus H_2 \oplus \cdots \oplus H_m = K_1 \oplus K_2 \oplus \cdots \oplus K_n$$

where $|H_1| \geq |H_1| \geq \cdots \geq |H_m|$ and $|K_1| \geq |K_1| \geq \cdots \geq |K_n|$. Then $m = n$ and for each i, $|H_i| = |K_i|$.

Proof. Outline. By induction on the order of G. First verify that

$$G^p = H_1^p \oplus H_2^p \oplus \cdots \oplus H_m^p = K_1^p \oplus K_2^p \oplus \cdots \oplus K_n^p.$$

If any of the groups H_i^p or K_j^p are trivial, then drop them to get

$$G^p = H_1^p \oplus H_2^p \oplus \cdots \oplus H_{m'}^p = K_1^p \oplus K_2^p \oplus \cdots \oplus K_{n'}^p$$

to get two direct sums of nontrivial cyclic subgroups. By induction, $m' = n'$ and for each $i \leq m'$, $|H_i^p| = |K_i^p|$. Since $|H_i| = p|H_i^p|$ and $|K_i| = p|K_i^p|$, therefore $|H_i| = |K_i|$ for each $i \leq m'$. Finish with a counting argument to show that the number of trivial groups that were dropped is the same for the H's as for the K's. They're the subgroups H_i and K_i of order n. Q.E.D.

Putting the last theorem and lemma together, we have the following theorem.

Theorem 4.52 (Fundamental theorem of finite Abelian groups)**.** *A finite Abelian group is the direct sum of cyclic groups whose orders are prime powers. The number of terms in the direct sum and the orders of the cyclic groups are determined by the group.*

Index