

Algebraic Structures

Andrew D. Hwang

May 2013

To the Instructor

This book is based upon the one-semester course *Algebraic Structures* taught at the College of the Holy Cross, a “transition-to-proofs” course roughly equivalent to an introductory discrete mathematics course. The target readership is sophomore-level mathematics majors. The prerequisites are minimal: a solid precalculus background, the ability to read actively, and a capacity for abstract reasoning.

Algebraic Structures is loosely divided into three intertwining portions: background topics including logic, sets, induction, mappings, matrices, and permutations; axioms for the integers and Euclidean number theory; and the basics of (mostly finite) group theory, ending with quotient groups and the first homomorphism theorem. This book also includes a final chapter on groups of plane transformations, including Euclidean motions and a very brief introduction to smooth flows.

Explicit mention of rings and fields is notably absent.

The book is structured linearly, but contains more material than can be covered comfortably in one semester. Optional destinations include the binomial theorem, one-sided inverses of mappings, the geometry of residue class multiplication, linear congruences, symmetries of Platonic solids, and groups of plane transformations.

The instructor of a transitional course faces the challenge of balancing examples and theory, comprehensibility and logical care, all within the constraint of students’ disparate backgrounds. This book makes a special effort to forge linguistic and conceptual links between formal precision and the underlying intuition. The exposition is mostly concrete, but aims continually to extend students’ “comfort level” with abstraction, experimentation, and “realistically complicated” computation.

Complex numbers are informally introduced early on, from both algebraic and geometric perspectives, and are used repeatedly as a source

of examples. Euler's formula plays an important role throughout, connecting the algebraic simplicity of the law of exponents with the geometry of rotations and polar coordinates. (The power series justification of Euler's formula is given in an appendix.) Important structures, particularly finite cyclic groups, are revisited in numerous guises, including complex roots of unity, residue classes of integers, cycle permutations, and rotational symmetries of regular polygons.

Examples and exercises have been selected for their likely conceptual familiarity to students, intrinsic interest, and connections to other parts of the book or to more advanced parts of mathematics and physics. The exercises range in difficulty from routine to very challenging, with these terms calibrated to the target readership. Hints are normally provided when a problem requires a novel idiom or other "sneaky" idea.

To the Student

Mathematics is unique among human intellectual endeavors; it is not art, philosophy, or science, but it shares certain features with each. The example of digital data storage will help convey the nature and uses of mathematics and the flavor of the material covered in this book.

Computers store, manipulate, and transmit data as *bits* or *binary digits*. Physically, bits have been represented and conveyed in a vast array of schemes from historical to modern, including

- Shaking or nodding one's head.
- Dots and dashes in Morse code.
- Holes and “no-holes” in a paper strip: punch cards, ticker tape.
- Magnetic domains: floppy and ZIP disks, PC hard drives.
- Light and dark spots or bands: compact disks, UPC symbols, QR codes.
- Charged and uncharged capacitors: flash memory, RAM.

A mathematician or theoretical computer scientist sees no essential difference between these schemes: The central mathematical “object” is a *pair of contrasting states*. Depending on context, the states might be called (and, in actual practice, *are* called) “zero and one”, “false and true”, “white and black”, “no and yes”, “open and closed”, “low and high”, or “off and on”.

Mathematical abstraction extends beyond data, however, encompassing the operations performed on objects.

Binary arithmetic. Think of 0 as representing an arbitrary even integer, and 1 as representing an arbitrary odd integer; namely, identify an integer with its remainder on division by 2. The sum of two odd integers is even (“ $1 + 1 = 0$ ”), the product of an even and an odd

integer is even ($0 \cdot 1 = 0$), and so forth. These relationships may be tabulated as

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

Boolean logic. Think of F as representing an arbitrary “false” assertion (such as $2 + 2 = 5$) and T as representing an arbitrary “true” sentence (such as $1 + 1 = 2$). Since, e.g., “ $2 + 2 = 5$ or $1 + 1 = 2$, but not both” is true, we write “F xor T = T”. (“xor” stands for “exclusive or”: one statement is true, but not both.) Since “ $2 + 2 = 5$ and $1 + 1 = 2$ ” is false, we write “F and T = F”. Tabulating the “truth value” of the statements made by conjoining two statements, according to whether or not each statement is true or false,

xor	F	T	and	F	T
F	F	T	F	F	F
T	T	F	T	F	T

Abstract implementation. The truly mathematical observation is *the entries of these pairs of tables correspond*: Under the correspondence even-False and odd-True, “addition (mod 2)” corresponds to “xor”, and “multiplication (mod 2)” corresponds to “and”. The tables above are different “implementations” of the same abstract structure, which we might denote by

\vee	○	●	\wedge	○	●
○	○	●	○	○	○
●	●	○	●	○	●

These three pairs of tables exemplify an abstract relationship, known as “isomorphism”, between operations on pairs of contrasting states. Any logical consequence that holds for one implementation necessarily holds for other implementations. For example, if we let variables x and y stand for either of two contrasting states, and we denote by x' the state unequal to x , then the identities

$$(x')' = x, \quad x \vee y = (x \wedge y') \vee (x' \wedge y) = x' \vee y'$$

hold regardless of what values are assigned to x and y , and, more significantly, *no matter which implementation is used*. In this fashion,

mathematical structures can be studied and organized with their extraneous details stripped away. More subtle patterns can sometimes be discerned in the abstraction, leading to a deeper understanding of the original structures.

Among of the most remarkable internal features of mathematics is its absoluteness: the perfect and intricate logical meshing of truth even when drawn from widely separated sub-disciplines, the universality of theorems across cultures, the sense among mathematicians that their work describes an objective (if non-physical) reality.

As a language, mathematics is unparalleled in its ability to express features of the natural world, often with astounding accuracy. At the same time, mathematics has no known *a priori* connection to the real world. The objects of mathematics are idealized concepts (such as “a pair of contrasting states”), and do not have physical existence in the same way stars, molecules, or people do. Conversely, stars, molecules, and people are not mathematical objects, though they do possess attributes that can be modeled by mathematical concepts.

This book was written to help you bridge the gap between informal intuition and the more formal language and framework of modern mathematics. Learning mathematics requires active preparation and participation from you, but offers continual rewards, including deepened comprehension of the natural world and the sheer enjoyment of mathematical beauty.

Practice reading actively, with a pencil and scratch paper. When you encounter a new definition, try to construct examples and non-examples before reading further, and ask yourself how you might test an object to see if it satisfies the definition.

Develop the habit of filling in the missing steps of calculations and omitted “standard” steps of proofs. When you first read the statement of a theorem, pause to think about what the theorem claims, and whether or not you *believe* the assertion. Try to sketch out an argument on your own before reading the book’s proof.

Situate new general concepts and examples in the context of your existing mathematical knowledge. Pay attention to the overall structure of proofs, not merely to the details. Look for commonalities in arguments, and be sure you are able to use these strategies yourself. Your repertoire of proof techniques and other mathematical idioms will grow steadily.

Work on mathematics outside of class every day, rather than in one

or two long “marathon sessions” per week. Don’t become discouraged if new ideas don’t immediately “click”. Re-read confusing passages after a day or two. Speak with classmates and your instructor for clarification as necessary. At the same time, develop intellectual self-reliance. The more mathematics you have made your own, the easier learning new mathematics becomes.

Above all, cultivate the enjoyment of thinking about new ideas, solving problems, and finding meaningful connections between seemingly disparate concepts. The greatest reward of your mathematical studies will, ideally, be a deeper experience of life itself.

Contents

1	Logic and Proofs	1
1.1	Statements and Negations	1
1.2	Negation and Logical Connectives	2
1.3	Quantification	5
1.4	Truth Tables and Applications	7
2	An Introduction to Sets	13
2.1	Specifying Sets	13
2.2	Complex Numbers	15
2.3	Sets and Logic	22
2.4	Advice on Writing Proofs	24
3	The Integers	31
3.1	Counting and Arithmetic Operations	31
3.2	Axioms for \mathbf{Z}	32
3.3	Consequences of the Axioms	37
3.4	The Division Algorithm	40
4	Mappings and Relations	45
4.1	Mappings, Images, and Preimages	46
4.2	Surjectivity and Injectivity	49
4.3	Composition and Inversion	54
4.4	Equivalence Relations	58
5	Induction and Recursion	71
5.1	Mathematical Induction	71
5.2	Applications	79
5.3	The Binomial Theorem	86

6	Binary Operations	93
6.1	Properties of Binary Operations	95
7	Groups	107
7.1	The Law of Exponents	110
7.2	Subgroups	112
7.3	Groups of Complex Numbers	118
8	Divisibility and Congruences	125
8.1	Residue Classes of Integers	126
8.2	Greatest Common Divisors	129
9	Primes	135
9.1	Coprimality	136
9.2	Prime Factorizations	138
10	Multiplicative Inverses	145
10.1	Invertibility	146
10.2	The Geometry of Multiplication	150
10.3	Linear Congruences	152
11	Linear Transformations	155
11.1	The Cartesian Vector Space	155
11.2	Plane Transformations	159
11.3	Cartesian Transformations	164
12	Isomorphisms	179
12.1	Classification of Cyclic Groups	184
13	The Symmetric Group	191
13.1	Structure of the Symmetric Group	192
13.2	Cycle Multiplication	195
13.3	Parity and the Alternating Group	201
14	Examples of Finite Groups	207
14.1	Cayley's Theorem	207
14.2	The Dihedral Groups	209
14.3	The Quaternion Group	213
14.4	Subgroups of S_4	214
14.5	Symmetries of Polyhedra	216
14.6	Rubik's Cube	219

15 Cosets	225
15.1 Normal Subgroups	230
16 Homomorphisms	235
16.1 Definition and Properties	235
16.2 Homomorphisms and Cyclic Groups	239
16.3 Quotient Groups	240
16.4 The Homomorphism Theorem	243
17 Continuous Symmetries	247
17.1 Euclidean Planar Motions	248
17.2 Time-Invariant Planar Flows	259
Euler's Formula	265
Index	267

Chapter 1

Logic and Proofs

Mathematics admits no “absolute truth”. Instead, most mathematicians work within the axiom system known as Zermelo-Fraenkel with choice, or ZFC for short. ZFC formalizes the concept of a *set*, an abstraction of a collection of objects, called *elements*. For now, the details of ZFC are unimportant. This chapter describes the basic rules of logic. Chapter 2 provides an informal introduction to ZFC.

ZFC is believed to be logically consistent, and the “correctness” of mathematical statements is evaluated according to “provability” and “logical consistency” with respect to ZFC. Theorems proved in ZFC are said colloquially to be “true”. Strictly speaking, however, mathematicians do not find metaphysical truths, but instead deduce logical *conclusions* starting from assumptions called *hypotheses*.

1.1 Statements and Negations

A *statement* is a sentence having a *truth value*, T (True) or F (False). Contact with the external world can be made via experience, but in mathematics *true* and *false* may be viewed as undefined terms.

As noted earlier, the basic objects of ZFC are sets, collections of elements. The examples below refer to the set of *integers*, or whole numbers: 0, 1, -1 , 2, -2 , and so forth.

Example 1.1. -4 is an even integer.

The decimal expansion of π is non-repeating and contains the string ‘999999’. (True)

$2 + 2 = 5$. (False)

Example 1.2. Sentences that are *not* statements include “ n is an even integer” (whose truth value depends on n) “ 10^{1000} is a large number” (“large” has not been given mathematical meaning), and the self-referential examples, “This sentence is true” (whose truth value must be specified as an axiom) and “This sentence is false” (which cannot be consistently assigned a truth value).

1.2 Negation and Logical Connectives

Conventionally, abstract statements are denoted P and Q .

Not. The *negation* of a statement P is its logical opposite $\neg P$. You may regard the negation as P preceded by the clause “It is not the case that...”, but usually a more pleasant wording can be found.

Example 1.3. P : $2 + 2 = 4$. $\neg P$: $2 + 2 \neq 4$.

Let P and Q be statements. New statements can be constructed using the “logical connectives” *and*, *or*, and *implies*.

And. The statement “ P and Q ” has its ordinary meaning: The compound statement is true provided both P and Q are true, and is false otherwise.

Example 1.4. $2 + 2 = 4$ and $0 < 1$. (True)

$2 + 2 = 5$ and $0 < 1$. (False)

$2 + 2 = 5$ and $1 < 0$. (False)

Or. The statement “ P or Q ” always has the “inclusive” meaning in mathematics: P is true, or Q is true, *or both*.

Example 1.5. $2 + 2 = 4$ or $0 < 1$. (True)

$2 + 2 = 5$ or $0 < 1$. (True)

$2 + 2 = 5$ or $1 < 0$. (False)

Remark 1.6. In colloquial English, “or” is frequently used in the “exclusive” sense. The sentence “You will earn a 70% on the final exam or you will not pass the course” is conventionally interpreted to mean “If you earn a 70% on the final exam, then you will pass the course, and if you do not earn a 70%, then you will not pass.”

Mathematicians and computer scientists denote “exclusive or” by “xor” to distinguish it from “or”. The statement “ P xor Q ” means

P is true, or Q is true, *but not both*. When needed, “ P xor Q ” can be expressed as “ $(P \text{ or } Q) \text{ and } \neg(P \text{ and } Q)$ ”. In this book, xor does not appear again.

Implies. A statement of the form “If P then Q ”, also read “ P *implies* Q ”, is called a *logical implication* and plays a central role in mathematics. P is called the *hypothesis* of the implication, Q the *conclusion*.

By definition, a logical implication is *valid* provided Q is true whenever P is true. In other words, “ P implies Q ” is a valid deduction unless P is true and Q is false.

Example 1.7. If $1 \neq 0$, then $1^2 \neq 0$. (True)

If $1 \neq 0$, then $1^2 = 0$. (False)

If $1 = 0$, then $0 = 0$. (True)

If $1 = 0$, then $1^2 = 0$. (True)

If “ P implies Q ” is valid, we think of Q as being *deduced* or *derived* from P . The definition of “valid implication” ensures that by starting with true hypotheses and making valid deductions, we obtain only true conclusions, not falsehoods. There are two noteworthy and potentially confusing consequences of this convention, however.

First, it is valid (not logically erroneous) to deduce an arbitrary conclusion from a false hypothesis. An implication with false hypothesis is said to be *vacuously true*. Humorous examples abound: “If $1 = 0$, then money grows on trees.”

In particular, the third and fourth implications of the preceding example are vacuously true. It may be helpful to point out that in each case, we can give a proof. If $1 = 0$, then subtracting this equation from itself gives $0 = 0$, which proves the third statement, while squaring gives $1^2 = 0^2 = 0$, proving the fourth statement.

Second, a valid implication need not connect causally related statements. The implication “If $0 = 0$, then 2 is an even integer” is valid because both the hypothesis and conclusion are true, but is effectively a *non sequitur*; the conclusion does not “follow” from the hypothesis in any obvious sense. A valid implication does not, of itself, constitute a proof. In the example at hand, we know the implication is valid only because there exists a separate proof, consisting of implications whose validity can be checked directly.

In these two senses, mathematicians are liberal in deeming an implication to be valid. Again, “validity” is the weakest criterion that

excludes the act of drawing a false conclusion from a true hypothesis.

Remark 1.8. If, in some axiom system, some statement P and its negation $\neg P$ are both true, then *every* statement Q is provable, since either “ P implies Q ” or “ $\neg P$ implies Q ” is vacuously true. The pair $\{P, \neg P\}$ is called a *logical contradiction*. An axiom system is *inconsistent* if a contradiction can be derived in it.

Work of K. Gödel in the 1930s showed ZFC cannot be proved consistent without using some other (“more powerful”) axiom system whose consistency is unknown. However, if there is a contradiction in ZFC, there is a contradiction in ordinary arithmetic.

Belief in the consistency of ZFC is about as close as mathematics gets to an “article of faith”.

In this book, and throughout mathematics in practice, valid deductions do actually link causally related statements. Most implications involve classes of objects, and assert that every object satisfying some condition must also satisfy some other condition.

Negation and Conjunctions

If P and Q are statements, then the statement “ P and Q ” is false if *at least one* of P and Q is false. If someone assures you two statements are both true, only one has to be false for the assurance to be unfounded. Formally, the compound statements

$$\neg(P \text{ and } Q), \quad (\neg P) \text{ or } (\neg Q)$$

express the same logical condition.

Analogously, if someone assures you at least one statement of two is true, then both must be false for the assurance to be unfounded. Formally, the compound statements

$$\neg(P \text{ or } Q), \quad (\neg P) \text{ and } (\neg Q)$$

express the same logical condition.

Together, the two preceding relationships are known as *De Morgan’s laws*, after the 19th Century English logician A. De Morgan. Loosely, the conjunctions “and” and “or” are interchanged by negation, perhaps contrary to first impression.

Consequently, the order of negation and a connective matters:

Example 1.9. The integers 1 and 0 are **not both** zero. (True.)

The integers 1 and 0 are **both not** zero. (False.)

Remark 1.10. All too frequently, one sees humorous ambiguities of the type “While driving, teens should not use cell phones and obey traffic laws”. To avoid confusion, this sentence should be phrased “While driving, teens should obey traffic laws and not use cell phones” (placing the negation where it clearly applies only to one clause) or “While driving, teens should not use cell phones, and should obey traffic laws” (explicitly delimiting the negation).

In formal logic, “ $\neg P$ and Q ” means “ $(\neg P)$ and Q ”.

1.3 Quantification

To accommodate classes of objects in the framework of statements, we allow statements to contain *variables* standing for elements of a set, so long as each variable is “quantified”, accompanied by the phrase “for every” or “there exists”. The quantifiers are crucial; pay close attention to them while reading, and *do not omit them when thinking and writing*.

Example 1.11. For every integer n , $n^2 - n$ is an even integer. (True)

For every integer n , $n^2 \geq 0$. (True)

For every integer n , $n^2 = 1$. (False)

The preceding “for every” statements involve *universal quantification*. Each statement encapsulates multiple statements. For example, the first statement of the preceding example encapsulates an infinite collection of statements, one for each integer: $0^2 - 0$ is an even integer; $1^2 - 1$ is an even integer; $(-1)^2 - (-1)$ is an even integer; and so forth.

Example 1.12. There exists an integer n such that $n^2 = 1$. (True)

There exists an integer n such that $n^2 = 2$. (False)

There exists an n such that both n and $n + 1$ are even. (False)

The preceding “there exists” statements involve *existential quantification*. Again, each encapsulates multiple statements. For example, the third expresses that at least one truism is found among the statements: 0 and 1 are both even; 1 and 2 are both even; -1 and 0 are both even; and so forth. The compound statement is false because *every* individual statement is false.

Remark 1.13. The statements of the preceding examples contain only “bound” (i.e., quantified) variables.

Sentences containing “free” or “unbound” variables (such as “ n is an even integer” or “ $x^2 + x - 2 = 0$ ”) are not statements. However, sentences containing unbound variables play the useful role of *conditions* in mathematics, selecting objects (perhaps integers n or real numbers x) for which the resulting statement is true.

Many mathematical theorems take the universally quantified form “For every x satisfying $P(x)$, condition $Q(x)$ is true”. For stylistic variety, such statements may be worded as implications involving “arbitrary” values of variables.

Example 1.14. If x is a real number such that $x^2 + x - 2 = 0$, then $x = 1$ or $x = -2$. (True)

If n is an integer, then there exist unique integers q and r such that $n = 4q + r$ and $0 \leq r < 4$. (True)

If a , b , and c are positive integers, then $a^3 + b^3 \neq c^3$. (True)

Quantifiers and Negation

The universal quantifier “for every” may be viewed as an enhancement of the “and” conjunction: “For every integer n , the condition $P(n)$ is true” means that the infinitely many statements $P(0)$, $P(1)$, $P(-1)$, and so forth, are *all* true.

The existential quantifier “there exists” may be viewed similarly as an enhancement of “or”: “There exists an integer n such that the condition $P(n)$ is true” means that among the infinitely many statements $P(0)$, $P(1)$, $P(-1)$, \dots , *at least one* is true.

Example 1.15. Logical negation “converts” a “for every” statement into a “there exists” statement of negations, and converts a “there exists” statement into a “for every” statement of negations:

P : For every integer n , $n^2 \geq 0$.

$\neg P$: There exists an integer n such that $n^2 < 0$.

P : There exist integers m and n such that $m^2 + n^2 = 8$.

$\neg P$: For all integers m and n , $m^2 + n^2 \neq 8$.

Remark 1.16. This type of linguistic transformation needs to become second nature. Particularly, a positive assertion regarding a class of objects can be disproved by finding a counterexample, but cannot be proved by finding an example.

Remark 1.17. When the hypothesis of a logical implication contains a variable but no quantifier is explicitly present, the convention is to read “for every”. For example, “If $x > 0$ then $x^2 > 0$ ” should be read “For every real number x , if $x > 0$ then $x^2 > 0$ ” (assuming the context dictates real numbers as opposed to, say, integers).

If an implicitly-quantified statement is negated, the existential quantifier must be added explicitly: “There exists a real number $x > 0$ such that $x^2 \leq 0$ ”.

To avoid confusion, including your own, include logical quantifiers explicitly. This book makes a special effort to set a good example.

Implications, and Multiple Quantifiers

Among the most subtle conditions in mathematics are those containing multiple quantifiers. Elementary algebra seldom ventures into this territory, but analysis, the mathematics underlying and extending differential and integral calculus, is suffused with definitions and theorems of this type. When you encounter multiply-quantified statements, slow down and read several times to ensure you thoroughly understand the dependencies implicit in the ordering.

Example 1.18. For every integer n , there exists an integer M such that $n \leq M$. (True; every integer n is smaller than some other integer M .)

There exists an integer M such that for every integer n , $n \leq M$. (False; there is no largest integer M , i.e., no integer that is greater than every other integer n .)

1.4 Truth Tables and Applications

The logical operators (“not”, “and”, “or”, and “implies”) introduced above are neatly summarized by *truth tables*:

P	Q	$\neg P$	P and Q	P or Q	P implies Q
T	T	F	T	T	T
T	F	F	F	T	F
F	T	T	F	T	T
F	F	T	F	F	T

Truth tables furnish a useful tool for studying sentences built of other statements and logical connectives. This section gives a few applications.

Logical Equivalence. Two statements P and Q are *logically equivalent* if each implies the other: P implies Q and Q implies P . For brevity, we may write P iff Q , “iff” being short for “if and only if”. A truth table calculation shows P and Q are equivalent precisely when they have the same truth value:

P	Q	P implies Q	Q implies P	P iff Q
T	T	T	T	T
T	F	F	T	F
F	T	T	F	F
F	F	T	T	T

The Converse. The implications “ P implies Q ” and “ Q implies P ” are said to be *converse* to each other. The preceding table shows these implications are not equivalent.

The Contrapositive. The implications “ P implies Q ” and “ $\neg Q$ implies $\neg P$ ” are said to be *contrapositive* to each other. An implication and its contrapositive are logically equivalent:

P	Q	P implies Q	$\neg Q$	$\neg P$	$\neg Q$ implies $\neg P$
T	T	T	F	F	T
T	F	F	T	F	F
F	T	T	F	T	T
F	F	T	T	T	T

This fact of logic should become second nature to you. Many implications are easier to understand and prove in contrapositive form.

Example 1.19. In each statement, x stands for a real number. Let P be the statement “ $x^2 - 1 \neq 0$ ” and Q be the statement “ $x \neq 1$ ”.

The implication P implies Q is true, but may require a few seconds’ thought to see.

The converse implication, “If $x \neq 1$, then $x^2 - 1 \neq 0$ ” is an invalid deduction. The number $x = -1$ is a *counterexample*: It satisfies the converse hypothesis Q , but not the converse conclusion P .

The contrapositive reads, “If $x = 1$, then $x^2 - 1 = 0$.” This implication is obviously true, and on general grounds its truth implies the truth of P implies Q .

Example 1.20. In each statement below, n is a positive integer. A positive integer n is said to be *prime* if $n > 1$, and if n has no positive divisors other than 1 and n .

Direct implication: If n is a prime, then $n = 2$ or n is odd. (True.)

Converse: If $n = 2$ or n is odd, then n is a prime. (False: $n = 1$ and $n = 9$ are the two smallest of infinitely many counterexamples.)

Contrapositive: If $n \neq 2$ and n is not odd, then n is not prime. (True. Every such integer has the form $n = 2k$ for some integer $k > 1$.)

One final example, drawn from analysis rather than from algebra, will illustrate the power of the contrapositive.

Example 1.21. In each statement, $x \geq 0$ is a real number, and n is a positive integer.

Direct implication: If $x < 1/n$ for every n , then $x = 0$.

Converse: If $x = 0$, then $x < 1/n$ for every n .

Contrapositive: If $x > 0$, then there exists an n such that $1/n \leq x$.

It turns out that all three statements are true. The second is easily seen, even though the conclusion consists of infinitely many statements: $0 < 1$, $0 < 1/2$, $0 < 1/3$, etc.

The third statement is true, and not difficult to see; informally, $1/k \rightarrow 0$ as $k \rightarrow \infty$, so if $x > 0$, there is some positive integer n such that $1/n \leq x$.

The direct implication is therefore true, since its contrapositive is true. However, the direct implication exhibits a new phenomenon: The hypothesis consists of infinitely many statements, $x < 1$, $x < 1/2$, $x < 1/3$, etc., but *no finite number of these statements implies the conclusion*. Indeed, if we assume only finitely many inequalities of the form $x < 1/n$, there is a largest denominator, say N , and our collection of inequalities is equivalent to the single inequality $x < 1/N$, which does not imply $x = 0$.

Exercises

Exercise 1.1. In each pair P, Q of conditions, n represents an integer.

(i) Give the negations of P and Q , and (ii) Form the implication P implies Q , its converse, and its contrapositive, and determine whether each is true.

(a) P : $n^2 - 4 = 0$. Q : $n = 2$.

(b) P : n is even. Q : n is an integer multiple of 4.

(c) P : n is even. Q : n is the square of an even integer.

Exercise 1.2. Let P and Q be arbitrary statements. Use a truth table to prove that “ P implies Q ” is logically equivalent to “ $\neg P$ or Q ”.

Exercise 1.3. Let P , Q , and R be arbitrary statements. Use a truth table to prove the following pairs of statements are logically equivalent:

- (a) “ $\neg(P \text{ or } Q)$ ” and “ $\neg P$ and $\neg Q$ ”.
- (b) “ $\neg(P \text{ and } Q)$ ” and “ $\neg P$ or $\neg Q$ ”.
- (c) “ $(P \text{ or } Q) \text{ and } R$ ” and “ $(P \text{ and } R) \text{ or } (Q \text{ and } R)$ ”.
- (d) “ $(P \text{ and } Q) \text{ or } R$ ” and “ $(P \text{ or } R) \text{ and } (Q \text{ or } R)$ ”.

Exercise 1.4. A game-show host presents the contestant with the equation “ $a^2 + b^2 = c^2$ ”. The contestant replies, “What is the Pythagorean theorem?”

Why is the contestant’s reply logically deficient? Modify it to give a mathematically satisfactory question.

Exercise 1.5. The President, a law-abiding citizen who always tells the truth, has time for one more Yes/No question at a press conference. In an attempt to embarrass the President, a reporter asks, “Have you stopped offering illegal drugs to visiting Heads of State?”

- (a) Which answer (“Yes” or “No”) is logically truthful?
- (b) Suppose the President answers “Yes”. Can the public conclude that the President has offered illegal drugs to visiting Heads of State? What if the answer is “No”?
- (c) Explain why both answers are embarrassing.

If the President were a Zen Buddhist she might reply “mu” (pronounced “moo”), meaning “Your question is too flawed in its hypotheses to answer meaningfully.”

Exercise 1.6. Each of the following quotes has a logically humorous aspect. Explain why each statement is awkward, and either find a more natural alternative that conveys the same meaning or determine what the speaker probably intended to say.

- (a) Is anybody here Pope? (Stand-up comedian Jim Gaffigan.)

- (b) We will have the best-educated Americans in the world. (Then Vice-President Dan Quayle.)

Exercise 1.7. Grasping the correct usage of the phrases “for every”, “there exists”, and “such that” can be tricky. Explain why each of the following is anomalous, and determine the presumed meaning.

- (a) There exists a real number x such that $2 + 2 = 4$.
- (b) If $\delta > 0$ for every δ such that $\delta > 1$, then $0 < 1 < \delta^2$.
- (c) If $y = x^2$ for every $x > 0$, then $y > 0$.

Exercise 1.8. The human brain has evolved to detect “cheating”—behavior violating established rules. These rules may have logical formulations, but the “cheating” interpretation can be remarkably easier to “see”.

- (a) Each card in a deck is printed with a letter “D” or “N” on one side and a number between 16 and 70 on the other. Your job is to assess whether or not cards satisfy the criterion: “Every ‘D’ card has a number greater than or equal to 21 printed on the reverse.” You are also to separate cards that satisfy this criterion from those that do not.

Write the criterion as an “If . . . , then . . .” statement, and determine which of the following cards satisfy the criterion:

20	46	16	25
D	D	N	N
(i)	(ii)	(iii)	(iv)

- (b) You are shown four cards:

18	35	D	N
(i)	(ii)	(iii)	(iv)

Which cards must be turned over to determine whether or not they satisfy the criterion of part (a)?

- (c) The legal drinking age in a certain state is 21. Your job at a gathering is to ensure that no one under 21 years of age is drinking alcohol, and to report those that are. A group of four people consists of a 20 year old who is drinking, a 46 year old who is drinking, a 16 year old who is not drinking, and a 25 year old who is not drinking. Which of these people is/are violating the law?

After reporting this incident, you find four people at the bar: An 18 year old and a 35 year old with their backs to you, and two people of unknown age, one of whom is drinking. From which people do you need further information to see whether or not they are violating the law?

- (d) Explain why the card question is logically equivalent to the drinking question. Which did you find easier to answer correctly?

Chapter 2

An Introduction to Sets

Modern mathematics is built on the concept of a “set”, a collection of “elements”. These primitive notions will serve in lieu of definitions. This chapter informally introduces the set of complex numbers, connects sets with the basics of logic, and gives advice on constructing and writing mathematical proofs.

2.1 Specifying Sets

Example 2.1. The collection of all integers (whole numbers) is a set. Its elements are 0, 1, -1 , 2, -2 , and so forth. The set of integers is denoted \mathbf{Z} , from the German *Zahl* (number). Formal axioms for the integers are given in Chapter 3.

Example 2.2. The collection of “prime numbers”, integers p greater than 1 that have no divisors other than 1 and p , is a set. The numbers 2, 5, and $2^{13466917} - 1$ are elements, while 4 and $2^{13466917} = 2 \cdot 2^{13466916}$ are not.

Example 2.3. The set of periodic table entries in the year 1960 has 102 elements. “Hydrogen”, “promethium”, and “astatine” are elements of this set, while “Massachusetts”, “ammonia”, and “surprise” are not.

Abstract sets will be denoted with capital letters, such as A or B . Elements are normally denoted with lower case letters, such as a and b . We write “ $a \in A$ ” as shorthand for the statement “ a is an element of (the set) A ”, and “ $b \notin A$ ” for the logical negation “ b is not an element of A ”. For example, $0 \in \mathbf{Z}$, $-7 \in \mathbf{Z}$, and $\frac{1}{2} \notin \mathbf{Z}$.

Definition 2.4. Let A and B be sets. We say A is a *subset* of B , and

write “ $A \subseteq B$ ”, if $x \in A$ implies $x \in B$, that is, if every element of A is an element of B . Two sets A and B are *equal* if $A \subseteq B$ and $B \subseteq A$, namely if they have exactly the same elements: $x \in A$ if and only if $x \in B$.

The most basic and explicit way of describing a set is to list its elements. Curly braces are used to denote a list of elements comprising a set. Sets do not “keep track of” what order the elements are listed, or whether their elements are multiply-listed.

Example 2.5. Each of the sets $A = \{-1, 0, 1\}$, $B = \{0, 1, -1\}$, and $C = \{0, 1, 0, -1, 1\}$, contains three elements, and in fact $A = B = C$.

Example 2.6. Let A be a set. For each element a in A , there is a *singleton* set $\{a\}$ contained in A . Take care to distinguish a and $\{a\}$; a is an object, while $\{a\}$ is a “package” that contains exactly one object.

Example 2.7. There exists an *empty set* \emptyset containing *no* elements. For all x , the statement $x \in \emptyset$ is false. In particular, for every set A the logical implication “ $x \in \emptyset$ implies $x \in A$ ” is vacuous (has false hypothesis). Consequently, $\emptyset \subseteq A$ is true for all A .

Remark 2.8. The empty set is unique: If \emptyset and \emptyset' are sets having no elements, then $\emptyset \subseteq \emptyset'$ and $\emptyset' \subseteq \emptyset$ are both true, so $\emptyset = \emptyset'$.

In mathematics, we always restrict attention to sets contained in a fixed set \mathcal{U} , called a *universe*. Specific subsets of \mathcal{U} are conveniently described using *set-builder notation*, in which elements are selected according to logical conditions formally known as a *predicates*. The expression $\{x \text{ in } \mathcal{U} : P(x)\}$ is read “the set of all x in \mathcal{U} such that $P(x)$ ”.

Example 2.9. The expression $\{x \text{ in } \mathbf{Z} : x > 0\}$, read “the set of all x in \mathbf{Z} such that $x > 0$ ”, specifies the set \mathbf{Z}^+ of *positive integers*.

To personify, if \mathcal{U} is a population whose elements are individuals, then a subset A of \mathcal{U} is a club or organization, and the predicate defining A is a membership card. We screen individuals x for membership in A by checking whether or not x carries the membership card for A , namely whether or not $P(x)$ is true.

Example 2.10. Thanks to *Russell’s paradox*, named for the English logician B. Russell, there is no “set \mathcal{U} of all sets”. If there were, the set $R = \{x \text{ in } \mathcal{U} : x \notin x\}$ of all sets that are not elements of themselves would have the property that $R \in R$ if and only if $R \notin R$, a contradiction.

Example 2.11. The expression $\{x \text{ in } \mathbf{Z} : x = 2n \text{ for some } n \text{ in } \mathbf{Z}\}$ is the set of *even integers*. We often denote this set $2\mathbf{Z}$, the idea being that the general even integer arises from multiplying some integer by 2.

Similarly, the set of *odd integers* could be expressed as

$$2\mathbf{Z} + 1 = \{x \text{ in } \mathbf{Z} : x = 2n + 1 \text{ for some } n \text{ in } \mathbf{Z}\}.$$

Remark 2.12. For brevity, we sometimes write, e.g., the set of even integers as $\{2n : n \in \mathbf{Z}\}$, read “the set of $2n$ such that n is an element of \mathbf{Z} ”. This way of writing a set is convenient, and the meaning is generally clear, but it isn’t technically proper, compare Example 2.10. To define a set formally, first give the universe, then specify the predicate.

Remark 2.13. The elements of a set may be other sets. For example, the set $A = \{2\mathbf{Z}, 2\mathbf{Z} + 1\}$ has two elements, $2\mathbf{Z}$ and $2\mathbf{Z} + 1$. Note carefully that A is not a subset of \mathbf{Z} : The elements of A are not themselves integers, but sets of integers.

2.2 Complex Numbers

Points in the Cartesian plane may be viewed as numerical entities in a way that extends the familiar real number line. The resulting “complex number system” illustrates many of the algebraic and geometric concepts introduced later.

Definition 2.14. A *complex number* is an expression $\alpha = a + ib$ in which a and b are real numbers and i is a symbol satisfying $i^2 = -1$. The real numbers a and b are, respectively, the *real part* and *imaginary part* of α . We say α is *real* if $b = 0$, *non-real* if $b \neq 0$, and *imaginary* if $a = 0$.

Viewing the real and imaginary parts of a complex number $\alpha = a + bi$ as Cartesian coordinates, we identify α with the point (a, b) , Figure 2.1.

Definition 2.15. The set of complex numbers is the *complex plane* \mathbf{C} . Each real number a is identified with the complex number $a + 0 \cdot i$. The set of all such points is the *real axis*. The set of all imaginary numbers $0 + b \cdot i$ is the *imaginary axis*. The *conjugate* of α is the complex number $\bar{\alpha} = a - bi$ obtained by reflecting α across the real axis.

Remark 2.16. Imaginary numbers may seem tainted with suspicion, as if they don’t really exist but it’s mathematically convenient to pretend

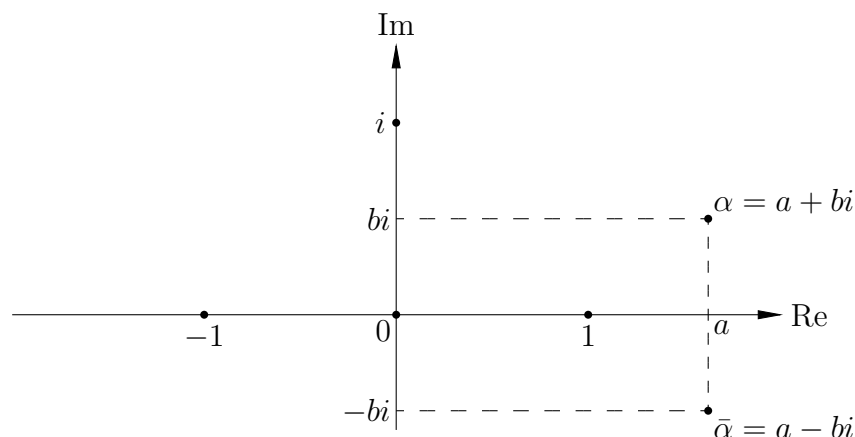


Figure 2.1: The complex plane.

they do. This sentiment traces back to the Ancient Greeks, who viewed numbers as lengths, what we now call “real numbers”. Indeed, no real number has square equal to -1 .

As noted above, however, i has a perfectly concrete existence as the point $(0, 1)$ in the Cartesian plane. Even the mysterious equation $i^2 = -1$ turns out to have a natural interpretation: Multiplication by i corresponds to a counterclockwise quarter-turn of the complex plane about the origin. Performing this operation twice, namely squaring, amounts to a half-turn, which multiplies each complex number by -1 .

From a modern perspective, the complex numbers earn their status as “numbers” by admitting operations of addition, subtraction, multiplication, and division that generalize the familiar algebraic properties of real numbers. We turn next to the algebraic and geometric descriptions of these operations.

Definition 2.17. Let $\alpha_1 = a_1 + ib_1$ and $\alpha_2 = a_2 + ib_2$ be complex numbers. Their *sum* is defined by the formula

$$\alpha_1 + \alpha_2 = (a_1 + ib_1) + (a_2 + ib_2) = (a_1 + a_2) + i(b_1 + b_2).$$

The formula for subtraction is similar, and left to you to work out. Adding two complex numbers corresponds to the parallelogram law for vector addition in the plane, see Figure 2.2.

Definition 2.18. A set A contained in \mathbf{C} is *closed under addition* if for all α_1 and α_2 in A , the sum $\alpha_1 + \alpha_2$ is in A .

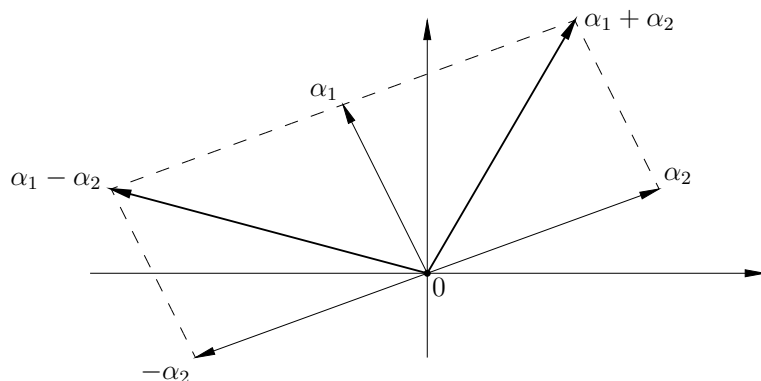


Figure 2.2: Adding and subtracting complex numbers.

Example 2.19. The set $\{0\}$ is closed under addition, since $0 + 0 = 0$.

Example 2.20. Suppose A is closed under addition and $1 \in A$. Of necessity, $2 = 1 + 1$, $3 = 2 + 1$, $4 = 3 + 1$, and so forth, are in A . That is, A contains the set of positive integers. Since the set of positive integers is closed under addition, our hypotheses imply nothing further.

Similarly, if A is closed under addition and $\alpha \neq 0$ is an element of A , then every positive integer multiple of α is an element of A . Since these multiples are distinct, the set A must be infinite.

If A is closed under addition, it does not follow that A is “generated” by one element as in the previous examples.

Example 2.21. The set \mathbf{Z} of integers is closed under addition in \mathbf{C} , as are the set \mathbf{Q} of rational numbers (ratios of integers) and the set \mathbf{R} of real numbers. None of these sets is obtained by adding a single element to itself repeatedly.

Example 2.22. The set $\mathbf{Z} + i\mathbf{Z} = \{m + in : m, n \in \mathbf{Z}\}$ of *Gaussian integers*, Figure 2.3, is closed under addition: If $\alpha_1 = m_1 + in_1$ and $\alpha_2 = m_2 + in_2$ are Gaussian integers, the addition formula for complex numbers gives $\alpha_1 + \alpha_2 = (m_1 + m_2) + i(n_1 + n_2)$. Since a sum of integers is an integer, the real and imaginary parts of $\alpha_1 + \alpha_2$ are integers. That is, $\alpha_1 + \alpha_2 \in \mathbf{Z} + i\mathbf{Z}$. Since α_1 and α_2 were arbitrary, $\mathbf{Z} + i\mathbf{Z}$ is closed under addition.

Example 2.23. The set A of complex numbers that are either real or imaginary, i.e., the union of the real and imaginary axes, is *not*

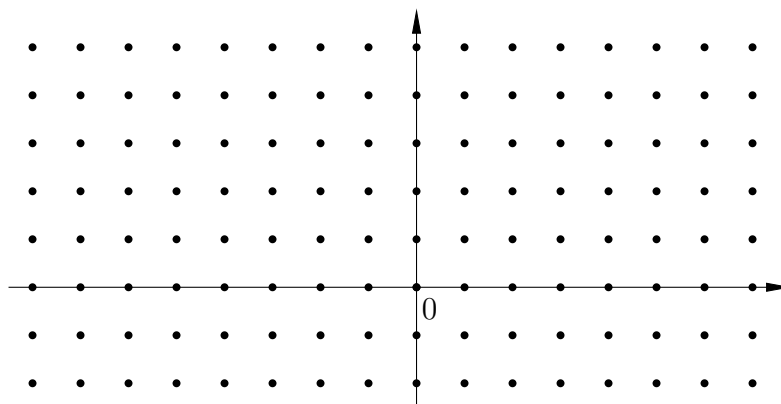


Figure 2.3: The Gaussian integers.

closed under addition. Since “closed under addition” is a “for every” condition, its negation is a “there exists” condition; that is, it suffices to find a *single counterexample*. For instance, $1 \in A$ (since 1 is real) and $i \in A$ (since i is imaginary) but $1 + i \notin A$ (the sum is neither real nor imaginary), so A is not closed under addition.

To define multiplication of complex numbers, we treat i as a symbol distributing over addition of real numbers, commuting with multiplication of real numbers, and satisfying $i^2 = -1$. A short calculation using familiar laws of algebra leads us to

$$\begin{aligned}(a_1 + ib_1)(a_2 + ib_2) &= a_1a_2 + i(a_1b_2 + a_2b_1) + i^2b_1b_2 \\ &= (a_1a_2 - b_1b_2) + i(a_1b_2 + a_2b_1).\end{aligned}$$

Definition 2.24. Let $\alpha_1 = a_1 + ib_1$ and $\alpha_2 = a_2 + ib_2$ be complex numbers. Their *product* is defined by the formula

$$\alpha_1\alpha_2 = (a_1a_2 - b_1b_2) + i(a_1b_2 + a_2b_1).$$

Example 2.25. If $\alpha = a + bi$, then $i\alpha = i(a + bi) = -b + ai$. As expected, the vector $(-b, a)$ is obtained by rotating the vector (a, b) through a quarter turn.

As a consistency check, $i(i\alpha) = i(-b + ai) = -a - bi = -\alpha$.

Example 2.26. If $\alpha = a + bi$, then

$$(2.1) \quad \alpha\bar{\alpha} = (a + bi)(a - bi) = a^2 - (bi)^2 = a^2 + b^2.$$

By the Pythagorean theorem, $\alpha\bar{\alpha} = (\text{distance from } 0 \text{ to } \alpha)^2$.

Complex multiplication is *commutative*: For all complex numbers α_1 and α_2 , we have $\alpha_2\alpha_1 = \alpha_1\alpha_2$. We may therefore attempt to define division by declaring $\beta = \alpha_1/\alpha_2$ if and only if $\beta\alpha_2 = \alpha_1 = \alpha_2\beta$.

Remark 2.27. If multiplication were not commutative, the equations $\alpha_1 = \beta\alpha_2$ and $\alpha_1 = \alpha_2\beta$ might well be incompatible conditions for α_1 .

To define complex division, let α_1 and α_2 be complex numbers with $\alpha_2 \neq 0$. We wish to write $\alpha_1/\alpha_2 = c_1 + ic_2$, namely, to find formulas for c_1 and c_2 in terms of the real and imaginary parts of the numerator and denominator.

The trick is analogous to rationalizing the denominator in high school algebra: Here we “realify” the denominator, multiplying top and bottom by the conjugate number $\bar{\alpha}_2 = a_2 - ib_2$ and using (2.1):

$$\frac{a_1 + ib_1}{a_2 + ib_2} = \frac{a_1 + ib_1}{a_2 + ib_2} \cdot \frac{a_2 - ib_2}{a_2 - ib_2} = \frac{(a_1a_2 + b_1b_2) + i(-a_1b_2 + a_2b_1)}{a_2^2 + b_2^2}.$$

Example 2.28. To divide $\alpha_1 = 2 - i$ by $\alpha_2 = 4 + 3i$, calculate as follows:

$$\begin{aligned} \frac{2 - i}{4 + 3i} &= \frac{2 - i}{4 + 3i} \cdot \frac{4 - 3i}{4 - 3i} = \frac{(8 - 3) + (-6 - 4)i}{4^2 + 3^2} \\ &= \frac{5 - 10i}{25} = \frac{1 - 2i}{5}. \end{aligned}$$

In practice, direct calculation is easier than memorizing the formula.

Example 2.29. If $\alpha = a + bi \neq 0$, then

$$\frac{1}{\alpha} = \frac{1}{a + ib} = \frac{a - ib}{a^2 + b^2} = \frac{a}{a^2 + b^2} - i \frac{b}{a^2 + b^2}.$$

That is, every non-zero complex number has a reciprocal.

The arithmetic operations on complex numbers satisfy familiar rules of algebra.

Example 2.30. For all complex α and β , the *difference of squares* identity holds: $\alpha^2 - \beta^2 = (\alpha + \beta)(\alpha - \beta)$.

Example 2.31. If $\alpha x^2 + \beta x + \gamma = 0$ with α , β , and γ complex and $\alpha \neq 0$, then

$$x = \frac{-\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha},$$

by the same completing-the-square proof you have seen for real coefficients. There are no “exceptional” cases; every quadratic has exactly two complex solutions, counting multiplicity.

Complex multiplication has a beautiful and useful geometric interpretation, most easily expressed in terms of *polar coordinates*. Recall that every point (a, b) in the plane can be written $(r \cos \theta, r \sin \theta)$ for some radius $r \geq 0$ and some angle θ , measured counterclockwise from the positive x axis and unique up to an added integer multiple of 2π .

Definition 2.32. Let $\alpha = a + bi = r \cos \theta + ir \sin \theta$ be a complex number. The radius r is called the *magnitude* of α , and the polar angle is the *argument* of α . If $-\pi < \theta < \pi$, we say θ is the *principal argument* of α .

Remark 2.33. The magnitude of $\alpha = a + ib$, denoted $|\alpha|$, is given by (2.1):

$$|\alpha| = r = \sqrt{a^2 + b^2} = \sqrt{\alpha \bar{\alpha}}.$$

Example 2.34. Since $i = 0 + 1 \cdot i = \cos \frac{\pi}{2} + i \sin \frac{\pi}{2}$, the magnitude of i is 1 and the principal argument of i is $\frac{\pi}{2}$.

Example 2.35. Let θ be a real number. By *Euler's formula* (see appendix), we have $\cos \theta + i \sin \theta = e^{i\theta}$. The magnitude of $e^{i\theta}$ is 1, and the argument is θ .

Generally, $\alpha = |\alpha|(\cos \theta + ir \sin \theta) = |\alpha|e^{i\theta}$.

If $e^{i\theta_1} = (\cos \theta_1 + i \sin \theta_1)$ and $e^{i\theta_2} = (\cos \theta_2 + i \sin \theta_2)$ are complex numbers of unit magnitude, the sum formulas for the cosine and sine functions allow us to write their product as

$$\begin{aligned} e^{i\theta_1} \cdot e^{i\theta_2} &= (\cos \theta_1 + i \sin \theta_1)(\cos \theta_2 + i \sin \theta_2) \\ &= (\cos \theta_1 \cos \theta_2 - \sin \theta_1 \sin \theta_2) + i(\cos \theta_1 \sin \theta_2 + \cos \theta_2 \sin \theta_1) \\ &= \cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2) = e^{i(\theta_1 + \theta_2)}. \end{aligned}$$

That is, *the law of exponents holds for imaginary exponents*. Since every complex number has polar form $\alpha = |\alpha|e^{i\theta}$, complex multiplication satisfies

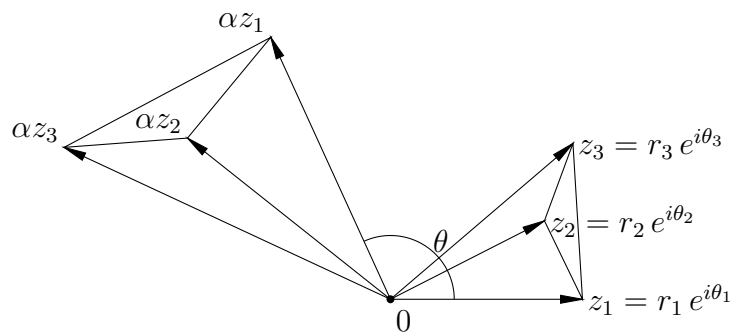
$$\alpha_1 \alpha_2 = (|\alpha_1| e^{i\theta_1})(|\alpha_2| e^{i\theta_2}) = (|\alpha_1| |\alpha_2|) e^{i(\theta_1 + \theta_2)}.$$

Geometrically, we multiply two complex numbers by multiplying their magnitudes and adding their arguments (polar angles).

Example 2.36. Since $i = \cos \frac{\pi}{2} + i \sin \frac{\pi}{2} = e^{i\frac{\pi}{2}}$, we have

$$i\alpha = i|\alpha|e^{i\theta} = |\alpha|e^{i(\theta + \frac{\pi}{2})};$$

again we see that multiplication by i rotates the plane about the origin by a quarter turn counterclockwise.

Figure 2.4: Complex multiplication by $\alpha = |\alpha|e^{i\theta}$.

Definition 2.37. A set A contained in \mathbf{C} is *closed under multiplication* if, for all α_1 and α_2 in A , the product $\alpha_1 \cdot \alpha_2$ is an element of A .

Example 2.38. The set of complex numbers of magnitude 1 is the *unit circle*

$$U(1) = \{z \text{ in } \mathbf{C} : |z| = 1\} = \{z \text{ in } \mathbf{C} : z = e^{i\theta} \text{ for some real } \theta\}.$$

The set $U(1)$ is closed under multiplication: If $|\alpha_1| = 1$ and $|\alpha_2| = 1$, i.e., $\alpha_1, \alpha_2 \in U(1)$, then $|\alpha_1\alpha_2| = |\alpha_1||\alpha_2| = 1$, so $\alpha_1\alpha_2 \in U(1)$.

Example 2.39. The *finite* subsets $\{1\}$ and $\{-1, 1\}$ of $U(1)$ are also closed under multiplication. More generally, for each positive integer n there exists a subset U_n of $U(1)$ that contains exactly n elements and is closed under multiplication:

$$\begin{aligned} U_n &= \{1 = e^0, e^{i2\pi/n}, e^{i4\pi/n}, \dots, e^{i2\pi(n-1)/n}\} \\ &= \{e^{i2\pi k/n} : k = 0, \dots, n-1\}. \end{aligned}$$

Figure 2.5 depicts the cases $n = 4$ and $n = 6$.

The elements of U_n are precisely the complex numbers $\zeta = re^{i\theta}$ satisfying the equation $\zeta^n = 1$, namely the so-called *n th roots of unity*. To see why, note that $1 = \zeta^n = r^n e^{in\theta}$ precisely when $r = 1$ and $n\theta$ is an integer multiple of 2π . Assuming without loss of generality that $0 \leq \theta < 2\pi$, we have $0 \leq n\theta < 2n\pi$, so that $n\theta = 0, 2\pi, 4\pi, \dots, 2(n-1)\pi$, or $n\theta = 2k\pi$ for some integer k with $0 \leq k < n$.

To see that the set of n th roots of unity is closed under multiplication, note that if $\zeta_1^n = 1$ and $\zeta_2^n = 1$, then $(\zeta_1\zeta_2)^n = \zeta_1^n \zeta_2^n = 1$, which means $\zeta_1\zeta_2$ is an n th root of unity.

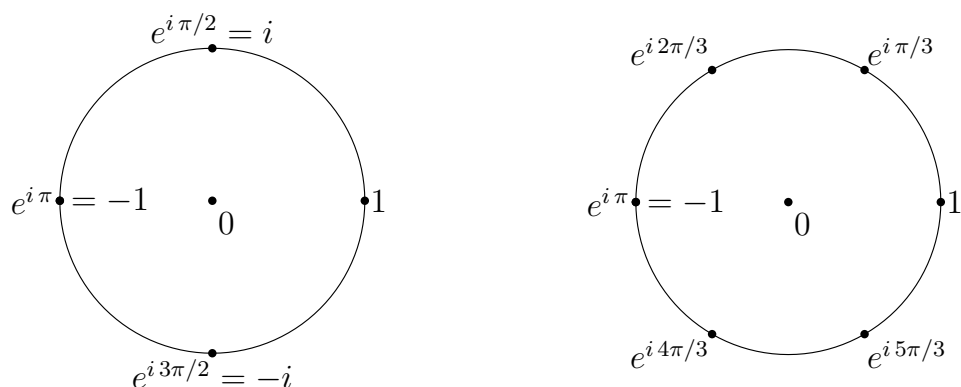


Figure 2.5: The unit circle, $U(1)$, and two finite subsets, U_4 and U_6 , that are closed under multiplication.

2.3 Sets and Logic

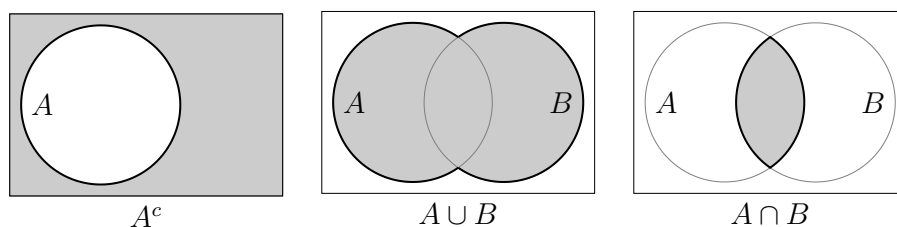
Let \mathcal{U} be a universe, and let A and B be subsets of \mathcal{U} . The statements $x \in A$ and $x \in B$ may be viewed as predicates P and Q on elements of \mathcal{U} . By definition, the logical implication “ $x \in A$ implies $x \in B$ ” corresponds to the set relation “ $A \subseteq B$ ”. Logical negation, disjunction (or), and conjunction (and) similarly have natural interpretations in terms of A and B .

The *complement* of A : $A^c = \{x \text{ in } \mathcal{U} : x \notin A\}$.

The *union* of A and B : $A \cup B = \{x \text{ in } \mathcal{U} : x \in A \text{ or } x \in B\}$.

The *intersection* of A and B : $A \cap B = \{x \text{ in } \mathcal{U} : x \in A \text{ and } x \in B\}$.

A *Venn diagram* represents subsets of a universe \mathcal{U} pictorially. The universe is depicted as a rectangle, and subsets are disks or, if necessary, more complicated shapes. The complement of A , or the union and intersection of two sets A and B , might be drawn as indicated:



Two sets A and B are *disjoint* if $A \cap B = \emptyset$, namely if A and B

have no elements in common. A Venn diagram of disjoint sets might be drawn as a pair of non-overlapping disks.

Example 2.40. The sets $2\mathbf{Z}$ and $2\mathbf{Z} + 1$ of even and odd integers are disjoint: No integer is both even and odd. The sets $A = 2\mathbf{Z}$ and $B = \mathbf{Z}^+$ are *not* disjoint: For example, 2, 4, and 84 are elements of $A \cap B$, since each is both positive and a multiple of 2.

Definition 2.41. Let A be a set. The *power set* of A , $\mathcal{P}(A)$, is the set of all subsets of A .

Example 2.42. If $A = \{0, 1\}$ has two elements, the power set $\mathcal{P}(A)$ has four elements:

$$\mathcal{P}(A) = \{\emptyset, \{0\}, \{1\}, A\}.$$

The empty set and A itself are always subsets of A , so a power set is never empty. Indeed, $\mathcal{P}(\emptyset) = \{\emptyset\}$ has a single element.

Definition 2.43. Let A be a set, and I a set of indices. A family of subsets $\{A_i\}_{i \in I}$ of A constitutes a *partition* of A if each element of A is an element of *exactly one* of the sets A_i .

In other words, $\{A_i\}_{i \in I}$ is a partition of A if $A_i \cap A_j = \emptyset$ for $i \neq j$ (each pair of sets is disjoint), and A is the union of the sets A_i .

Example 2.44. The sets $A_0 = 2\mathbf{Z}$ and $A_1 = 2\mathbf{Z} + 1$ are a partition of $A = \mathbf{Z}$; every integer is either even or odd, and no integer is both. Here the index set is $I = \{0, 1\}$.

The sets $A_0 = 3\mathbf{Z}$, $A_1 = 3\mathbf{Z} + 1$, $A_2 = 3\mathbf{Z} + 2$ are another partition of \mathbf{Z} , since every integer leaves a unique remainder of 0, 1, or 2 upon division by 3:

\mathbf{Z}	\cdots	-4	-3	-2	-1	0	1	2	3	4	5	6	\cdots
A_0	\cdots		-3			0			3			6	\cdots
A_1	\cdots			-2			1			4			\cdots
A_2	\cdots	-4			-1			2			5		\cdots

Example 2.45. We will prove in Chapter 3 (Theorem 3.15) that if $n > 1$ is an integer, there is a partition of \mathbf{Z} into n subsets, $A_k = n\mathbf{Z} + k$ with $k = 0, \dots, n - 1$ an integer. An integer x is an element of A_k if and only if x leaves a remainder of k on division by n .

In Chapter 8, we will write $[k]_n = n\mathbf{Z} + k$, and form a set \mathbf{Z}_n having n elements: $\mathbf{Z}_n = \{[0]_n, [1]_n, \dots, [n-1]_n\}$. Note that $\mathbf{Z}_n \subseteq \mathcal{P}(\mathbf{Z})$: The *elements* of \mathbf{Z}_n are *subsets* of \mathbf{Z} .

2.4 Advice on Writing Proofs

Discovering and writing proofs are nearly opposite activities. You'll find most of the writing you do in discovering mathematics does not need to be written up; it's just "scaffolding".

Example 2.46. Assume $\alpha \in \mathbf{C}$. Prove: $|\alpha| = |\bar{\alpha}|$.

(Preliminary Work). When proving an identity such as this we have an obvious strategy: Express each side in terms of simpler information and see if the answers agree. Here, set $\alpha = a + ib$ with a and b real. Then $\bar{\alpha} = a - ib$, so we have

$$|\alpha| = \sqrt{a^2 + b^2}, \quad |\bar{\alpha}| = \sqrt{a^2 + (-b)^2} = \sqrt{a^2 + b^2}.$$

These are indeed equal.

(The Written Solution). Assume $\alpha \in \mathbf{C}$. Prove: $|\alpha| = |\bar{\alpha}|$.

Proof: Let α be an arbitrary complex number, and write $\alpha = a + ib$ with a and b real. We have $\bar{\alpha} = a - ib$, and therefore

$$|\bar{\alpha}| = \sqrt{a^2 + (-b)^2} = \sqrt{a^2 + b^2} = |\alpha|,$$

as was to be shown.

Remark 2.47. When writing up a formal proof of an algebraic identity Q , the preferred style is to build a chain of equalities from one side to the other. *Do not* write down the desired conclusion Q , then manipulate each side until you have an identity P . At best, this "two-column" argument establishes the converse, Q implies P , which is not equivalent to P implies Q , and does not even imply the truth of Q . See Exercises 2.15 and 2.16 for pitfalls of the "two-column" style of proof.

Example 2.48. Prove or disprove: $2\mathbf{Z} + 1 = 2\mathbf{Z} - 1$.

(Preliminary Work). By definition of equality of sets, we are to determine whether each set is a subset of the other. Some initial formalization can be performed mechanically. Give each set a name, write down its definition, and express the question in terms of this framework.

Here, we have two sets of integers,

$$\begin{aligned} A = 2\mathbf{Z} + 1 &= \{x \text{ in } \mathbf{Z} : x = 2u + 1 \text{ for some } u \text{ in } \mathbf{Z}\}, \\ B = 2\mathbf{Z} - 1 &= \{y \text{ in } \mathbf{Z} : y = 2v - 1 \text{ for some } v \text{ in } \mathbf{Z}\}. \end{aligned}$$

We wish to show either that $A \subseteq B$ and $B \subseteq A$ (which by definition means $A = B$ as sets), or that at least one of these inclusions is false.

Next, try to determine intuitively whether or not the statement is false (which can be shown by exhibiting a counterexample, an element of one set that is not an element of the other set) or true. To get an element of $2\mathbf{Z}+1$, add 1 to an even integer: $1 = 0+1$, $3 = 2+1$, $5 = 4+1$, $-1 = -2+1$, and so forth, are elements. Similarly, subtracting 1 from an even integer gives an element of $2\mathbf{Z} - 1$: $-1 = 0 - 1$, $1 = 2 - 1$, $3 = 4 - 1$, $-3 = -2 - 1$, and so forth, are elements.

This evidence doesn't merely suggest the two sets *are* equal, it even points to a strategy of proof: Any integer one greater than an even integer is one less than the next largest even integer. We'll sketch out an informal proof to settle notation and iron out any unforeseen logical wrinkles.

The statement " $A \subseteq B$ " may be phrased "if $x \in A$, then $x \in B$ ". If $x \in A$, then by the definition of A there exists an integer u such that $x = 2u + 1 = 2(u + 1) - 1$. Setting $v = u + 1$ (an integer because u is), we see x has the form $2v - 1$ for some integer v , which by definition means $x \in B$. This shows $A \subseteq B$.

The inclusion $B \subseteq A$ is entirely similar, so at this stage we can write up a formal proof. The considerations above that led to the proof are customarily omitted from the formal write-up. Note, however, that the proof involves choices not easily known ahead of time; the scratch work is important!

(The Written Solution). Show $2\mathbf{Z} + 1 = 2\mathbf{Z} - 1$.

Proof: By definition, $A = \{x \text{ in } \mathbf{Z} : x = 2u + 1 \text{ for some } u \text{ in } \mathbf{Z}\}$ and $B = \{y \text{ in } \mathbf{Z} : y = 2v - 1 \text{ for some } v \text{ in } \mathbf{Z}\}$. Assume $x \in A$. By hypothesis, there exists an integer u such that $x = 2u + 1$. Let $v = u + 1$, so $u = v - 1$, and note v is an integer. Since

$$x = 2u + 1 = 2(v - 1) + 1 = 2v - 2 + 1 = 2v - 1,$$

$x \in B$. Since x was arbitrary (i.e., x could have been any element of A), we have shown $A \subseteq B$.

Conversely, suppose $y = 2v - 1 \in B$ for some integer v . Let $u = v - 1$, so that $v = u + 1$. Then

$$y = 2v - 1 = 2(u + 1) - 1 = 2u + 1,$$

so $y \in A$. Since y was arbitrary, we have shown $B \subseteq A$.

Since $A \subseteq B$ and $B \subseteq A$, we have $A = B$.

Writing proofs requires practice. The final result should be a coherent, logical, step-by-step argument starting with the given hypotheses and leading to the conclusion.

Example 2.49. Let A and B be subsets of \mathcal{U} . Find the most general conditions on A and B under which $A \cap B = A$.

(Examples). If you're comfortable with sets and operations, go for the frontal assault ("reducing to the definitions", below). Otherwise, proceed by writing out examples on scratch paper or a blackboard. If Venn diagrams are more natural, use those. If concrete sets are easier to think about, use those. At this stage it's all right to let $\mathcal{U} = \mathbf{Z}$, the set of integers, but in the final proof, do not make any assumptions on the nature of \mathcal{U} , A , or B .

(Simpler cases). Since the target condition involves two sets, we can reduce to a simpler question by "fixing" one set and letting the other set vary.

If $A = \emptyset$, then $A \cap B = \emptyset \cap B = \emptyset = A$ regardless of B . If $A = \mathcal{U}$, then $A \cap B = \mathcal{U} \cap B = B$, which is not equal to A unless $B = \mathcal{U}$.

These examples show the condition $A \cap B = A$ *can* be true, but is *not always* true. The guiding task is to discover what common aspect these examples possess. If you're still not sure, draw a Venn diagram with a circle representing A , and ask: What condition on B guarantees that $A \subseteq A \cap B$? Draw circles that are disjoint from A , that are contained in A , that partially overlap A , or that contain A . The evidence of this "experiment" should point toward the desired condition.

(Reducing to the definitions). The condition $A \cap B = A$ encapsulates two set inclusions, $A \cap B \subseteq A$ and $A \subseteq A \cap B$. The first inclusion is true for all pairs of sets: If $a \in A \cap B$, then $a \in A$ and $a \in B$, so perforce $a \in A$. Since a is an arbitrary element of $A \cap B$, this argument shows $A \cap B \subseteq A$.

We are therefore seeking the most general conditions under which $A \subseteq A \cap B$, namely, " $a \in A$ implies ' $a \in A$ and $a \in B$ '". Clearly, this is equivalent to " $a \in A$ implies $a \in B$," which may be rephrased as $A \subseteq B$, our putative answer.

As a consistency check, recall $A = \emptyset$ and $A = \mathcal{U} = B$ satisfied the condition. In each case, $A \subseteq B$ holds. If the purported abstract condition is violated by examples, it's definitely wrong.

(Putative conclusion). As the result of considerations above, we claim that $A \cap B = A$ if and only if $A \subseteq B$. To *prove* this formally, it

suffices to establish two logical implications:

$$A \cap B = A \text{ implies } A \subseteq B, \quad A \subseteq B \text{ implies } A \cap B = A.$$

Here, approximately, is what you'd normally write up:

(The Written Solution). $A \cap B = A$ if and only if $A \subseteq B$.

Proof: ($A \cap B = A$ implies $A \subseteq B$) Assume $A \cap B = A$, namely $A \cap B \subseteq A$ and $A \subseteq A \cap B$. Since the first inclusion holds for all sets, our initial hypothesis is equivalent to $A \subseteq A \cap B$.

Let a be an arbitrary element of A . Since $A \subseteq A \cap B$ by hypothesis, $a \in A \cap B$, so $a \in A$ and $a \in B$. In particular, $a \in B$. We have shown that if $a \in A$, then $a \in B$; this means that $A \subseteq B$, as was to be shown.

($A \subseteq B$ implies $A \cap B = A$) By hypothesis, if $a \in A$, then $a \in B$, so if $a \in A$, then $a \in A$ and $a \in B$. Since a is arbitrary we have $A \subseteq A \cap B$. The reverse inclusion $A \cap B \subseteq A$ holds for all sets A and B . We have shown that if $A \subseteq B$, then $A \cap B = A$. This completes the proof.

Find your own writing style. *Do write accurately and precisely*, but don't be pedantic or excessively wordy.

Avoid pronouns, especially "it". In the middle of even a simple proof, two or three objects tend to be under consideration, and "it" can often refer to any of them. If you're unable to decide exactly what "it" refers to, you've located something you don't fully understand.

Exercises

Exercise 2.1. Let A be a set and assume $a \in A$. Determine whether each statement is always true, sometimes true, or never true. If the statement is sometimes true, give examples of A and/or a for which the statement is true or is false.

- (a) $a \in \{a\}$ (b) $a \subseteq A$ (c) $\{a\} \subseteq \emptyset$ (d) $\emptyset \in A$ (e) $\{a\} \in A$

Exercise 2.2. Let $A = 2\mathbf{Z}$ and $B = 3\mathbf{Z}$.

- (a) Find $A \cap B$; that is, determine which integers are in $A \cap B$.
 (b) List the elements of $A \cup B$ between -12 and 12 .

Exercise 2.3. In each part, let $A = \mathbf{Z} + i\mathbf{Z}$.

- (a) Let $B = \{z \text{ in } \mathbf{C} : |z| \leq 2\}$. List the elements of $A \cap B$, and illustrate with a sketch.

- (b) How many elements of A satisfy $|z| \leq 5$?
 Suggestion: Listing them all may be a bit tedious, but by using symmetry you can cut your work by a factor of four.
- (c) Let $n > 0$ be a positive integer, and let C_n be the number of elements of A satisfying $|z| \leq n$. Prove $C_n < (2n + 1)^2$.
 Hint: Find a provably larger set containing $(2n + 1)^2$ elements.
- (d) Modify the idea of part (c) to prove $(n + 1)^2 < C_n$.

Exercise 2.4. In each part, let $A = \mathbf{Z} + i\mathbf{Z}$.

- (a) Show A is closed under multiplication.
- (b) Which elements of A have a reciprocal (multiplicative inverse) in A ?

Exercise 2.5. Each part refers to the set

$$\mathbf{Q}[\sqrt{2}] = \mathbf{Q} + \mathbf{Q}\sqrt{2} = \{m + n\sqrt{2} : m, n \in \mathbf{Q}\}.$$

- (a) Show $\mathbf{Q}[\sqrt{2}]$ is closed under addition by giving a formula for the sum of two elements. Suggestion: Compare Example 2.22.
- (b) Show $\mathbf{Q}[\sqrt{2}]$ is closed under multiplication by giving a formula for the product of two elements.
- (c) Show if $\alpha = m + n\sqrt{2}$ is a non-zero element of $\mathbf{Q}[\sqrt{2}]$, there exists a unique $\alpha' = m' + n'\sqrt{2}$ in $\mathbf{Q}[\sqrt{2}]$ such that $\alpha\alpha' = 1$.

Exercise 2.6. Let $\zeta = e^{2\pi i/3} = \frac{1}{2}(-1 + \sqrt{3}i)$, and $A = \mathbf{Z} + \zeta\mathbf{Z}$.

- (a) Give a formal definition of the set A .
- (b) Prove A is closed under addition.
- (c) Prove A is closed under multiplication. (This fact depends on the precise form of ζ .)
- (d) Show that $U_6 = U(1) \cap A$, and illustrate with a sketch.
- (e) Which elements of A have a reciprocal in A ? Explain.

Exercise 2.7. (a) Let $A = \{a, b, c\}$ be a set with three distinct elements. List the elements of the power set $\mathcal{P}(A)$.

- (b) How would your answer to part (a) differ if $A = \{0, 1, 2\}$?
- (c) Describe how you could use your answer to part (a) to list the elements of the power set of $A' = \{a, b, c, d\}$. Suggestion: There are two types of subset of A' , those having d as an element, and those not having d as an element.

Exercise 2.8. Let A and B be subsets of \mathcal{U} .

- (a) Suppose $A \subseteq B$. Prove $\mathcal{P}(A) \subseteq \mathcal{P}(B)$ as subsets of $\mathcal{P}(\mathcal{U})$.
- (b) Suppose that $\mathcal{P}(A) = \mathcal{P}(B)$ as subsets of $\mathcal{P}(\mathcal{U})$. Prove $A = B$.

Exercise 2.9. Let A and B be arbitrary subsets of a universe \mathcal{U} .

- (a) Prove $A \cup B = A$ if and only if $B \subseteq A$.
- (b) Prove $A \cap B = B$ if and only if $B \subseteq A$.

Exercise 2.10. Let A and B be subsets of \mathcal{U} .

- (a) Prove $A \subseteq B$ if and only if $B^c \subseteq A^c$, and illustrate with a Venn diagram.
- (b) How is part (a) related to contrapositives?

Exercise 2.11. Let A , B , and C be subsets of a universe \mathcal{U} , and let P , Q , and R be the predicates $x \in A$, $x \in B$, and $x \in C$. Use truth tables to establish the indicated identities.

- (a) $(A \cup B) \cup C = A \cup (B \cup C)$.
- (b) $(A \cap B) \cap C = A \cap (B \cap C)$.

Exercise 2.12. Let A , B and C be subsets of a universe \mathcal{U} . As in Exercise 2.11, use truth tables to establish *De Morgan's laws* (a) and (b) and the *distributive laws* (c) and (d).

- (a) $(A \cup B)^c = A^c \cap B^c$.
- (b) $(A \cap B)^c = A^c \cup B^c$.
- (c) $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$.
- (d) $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.

Exercise 2.13. Draw Venn diagrams illustrating each part of the preceding exercise, and compare with Exercise 1.3.

Exercise 2.14. Let A and B be subsets of \mathcal{U} . Their *difference* is defined to be $A \setminus B = \{x \text{ in } A : x \notin B\}$.

- (a) Prove $A \setminus B = A \cap B^c$, and illustrate with a Venn diagram.
- (b) List the elements of $\mathbf{Z} \setminus \mathbf{Z}^+$ between -5 and 5 .
- (c) List the elements of $2\mathbf{Z} \setminus 3\mathbf{Z}$ between -12 and 12 .
- (d) List the elements of $3\mathbf{Z} \setminus 2\mathbf{Z}$ between -12 and 12 .

Exercise 2.15. Explain in detail what is wrong with this two-column “proof” that $-1 = 1$.

$-1 = 1$	to be shown,
$(-1)^2 = 1^2$	square both sides,
$1 = 1$	true statement.

Therefore $-1 = 1$.

Exercise 2.16. Let a and b denote real numbers, and assume $a = b$.

- (a) What is wrong with the following “proof” that $2 = 1$?

$b^2 = ab$	$a = b$,
$b^2 - a^2 = ab - a^2$	subtract a^2 ,
$(b + a)(b - a) = a(b - a)$	factor each side,
$(b + a) = a$	cancel common factor,
$2a = a$	$a = b$,
$2 = 1$	cancel common factor.

- (b) If the proof is read from bottom to top, is each step valid?

Chapter 3

The Integers

The *integers* or *whole numbers* play a foundational role in this book. Despite their familiarity and seeming simplicity, the integers required millennia to discover, and even today are the source of deep open problems (and wry quips) in mathematics.

3.1 Counting and Arithmetic Operations

Imagine a shepherd of three thousand years ago. Each morning he lets his sheep out to pasture, and each evening brings them in. How can he be sure he hasn't lost any sheep during the day?

One can imagine a scheme: The shepherd gathers a supply of small stones. In the morning, as each sheep passes, he places one of the stones aside. Once all the sheep have left, he puts these stones into a pouch for safe keeping. That evening, he takes a stone from the pouch for each returning sheep. If he has stones left over, he has lost sheep.

The apocryphal shepherd has made a fundamental abstraction about the physical world: There is a meaningful notion of *counting*, or *a number of things*—sheep, or stones, or sunrises, or nicks on a tree branch. By keeping track of small stones (or *calculi* as they will be known in Latin many centuries in the shepherd's future), the shepherd can keep track of his flock.

The number of integers is infinite. No matter how many sheep or stones one has, even if the number be like unto the grains of sand on all the beaches of the world, one can imagine having one more.

This abstraction comes with unexpected benefits. Two shepherds can combine their flocks, and *calculate* how many sheep they have be-

tween them simply by agglomerating their piles of stones. They quickly discover it does not matter whose stones are appended to whose, the result is the same either way. They have discovered the operation of *addition*, and the *commutative law*.

Two shepherds can compare their flocks to see who has more sheep. No matter whose flocks are compared, one flock is larger than the other, or the flocks are the same size; absolute comparison is always possible. Moreover, if Aleph has more sheep than Beth, who has more sheep than Gimel, then Aleph has a larger flock than Gimel. They have discovered the *ordering* of the counting numbers, and the *transitive law*.

Debts can be accounted and paid by *subtracting* stones from a pile:

Aleph: You owe me • • • • • sheep.

Beth: But I possess only • • • sheep.

Aleph: Very well. Give me those, and you only owe me • • sheep.

Example 3.1. Negative numbers cannot be used heedlessly. An apocryphal medieval philosopher saw two people go into a house. When three later came out, the philosopher reasoned, “If one person goes in, the house will be empty again.”

3.2 Axioms for \mathbb{Z}

Millennia later, in the 19th Century, the start of the modern era of abstraction, mathematicians set out axioms for the integers, terms of a mathematical contract to formalize our intuition and experience of counting and arithmetic. For future reference, these axioms are collected in Table 3.1 on page 34.

Though the axioms are numerous, we must still prove properties a modern school student regards as obvious: 1 (the multiplicative identity element) is the smallest positive integer; $a < b$ if and only if $b > a$; $a \cdot 0 = 0$ for all a ; and so forth.

Remark 3.2. The integers can be constructed in ZFC from a much more spartan set of hypotheses: an element 0 and a notion of “successorship”. The set of positive integers can be defined as the set of repeated successors of 0, addition is defined by repeated successorship, and multiplication by repeated addition (for example, $2 \cdot 3 = 2 + 2 + 2$). In this context, the axioms acquire the status of theorems.

The axioms fall loosely into three groups of four, dealing respectively with addition, multiplication, and ordering or comparison. (Our numbering is not in any sense universal; concentrate on the meaning of each axiom rather than on its number.) Explanatory comments follow.

Addition

Addition is a “binary operation”, a way of “combining” two integers a and b to get an integer $a + b$. The addition operation satisfies the following conditions, given in both “plain English” and axiomatic form.

- **Associativity:** Regrouping a three-fold sum does not affect the final value.

For all integers a , b , and c , we have $a + (b + c) = (a + b) + c$.

- **Identity Element:** There exists an integer 0 that has no effect upon addition.

For all integers a , we have $a + 0 = 0 + a = a$.

- **Inverses:** Addition by an arbitrary element can be cancelled.

For each a in \mathbf{Z} , there exists an integer b (depending on a ; note the order of the quantifiers) with the property that $a + b = b + a = 0$.

- **Commutativity:** The order of summands is immaterial.

For all integers a and b , $a + b = b + a$.

Remark 3.3. The additive inverse of an integer a turns out to be unique, and is denoted $-a$.

By definition, *subtraction* means adding the additive inverse: $a - b = a + (-b)$. As a binary operation, subtraction is neither associative nor commutative, and there is no identity element for subtraction.

Multiplication

From the modern viewpoint, multiplication of positive integers is merely iterated addition. “Five times three” means the total number of sheep possessed by three shepherds, each owning five sheep. “Negative five times three” means the total *debt* of three shepherds, each having a debt of five sheep.

There exists a set \mathbf{Z} , a subset \mathbf{Z}^+ of \mathbf{Z} , and two operations, $+$ and \cdot , satisfying the following axioms.

- A1.** (Associativity of addition) For all elements a , b , and c in \mathbf{Z} , we have $a + (b + c) = (a + b) + c$.
- A2.** (Additive identity element) There exists an integer 0 such that $a + 0 = 0 + a = a$ for all a in \mathbf{Z} .
- A3.** (Additive inverses) For every a in \mathbf{Z} , there exists $-a$ in \mathbf{Z} such that $a + (-a) = (-a) + a = 0$.
- A4.** (Commutativity of addition) For all a and b in \mathbf{Z} , $a + b = b + a$.

- M1.** (Associativity of multiplication) For all a , b , and c in \mathbf{Z} , we have $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.
- M2.** (Multiplicative identity element) There exists an integer $1 \neq 0$ such that $a \cdot 1 = 1 \cdot a = a$ for all a in \mathbf{Z} .
- M3.** (Commutativity of multiplication) For all a and b in \mathbf{Z} , $a \cdot b = b \cdot a$.
- M4.** (Distributivity of multiplication over addition) For all a , b , and c in \mathbf{Z} , $a \cdot (b + c) = a \cdot b + a \cdot c$.

- O1.** (Law of Trichotomy) If $-\mathbf{Z}^+ = \{b \text{ in } \mathbf{Z} : -b \in \mathbf{Z}^+\}$, then the sets \mathbf{Z}^+ , $\{0\}$, and $-\mathbf{Z}^+$ are a partition of \mathbf{Z} .
- O2.** (Sum of positive numbers) If a and b are elements of \mathbf{Z}^+ , then $a + b \in \mathbf{Z}^+$.
- O3.** (Product of positive numbers) If a and b are elements of \mathbf{Z}^+ , then $a \cdot b \in \mathbf{Z}^+$.
- O4.** (Well-ordering) If $A \subseteq \{0\} \cup \mathbf{Z}^+$ is non-empty, then there is a “smallest element” in A , i.e., there exists an a_0 in A such that $a + (-a_0) \in \{0\} \cup \mathbf{Z}^+$ for every a in A .

Table 3.1: Axioms for the integers.

Attempting to reason along these lines is treacherous, however, as medieval philosophers discovered. What could “negative three shepherds” possibly mean? And what if each had a debt of five sheep? The mind reels. Even in the 21st Century, one can start fruitless arguments in mathematical web forums by asking what multiplying one negative number by another *really means*.

In the end, abstraction cuts the Gordian knot and lights the way forward. The operation of multiplication has the expected meaning when the operands are non-negative integers, and the meaning is dictated by conformity to the axioms when one or both operands are negative.

Remark 3.4. With the benefit of hindsight, we can view integers as displacements along a number line. If we view integers as complex numbers in the standard way (the integer a “is” the complex number $a + 0 \cdot i$ on the real axis), then positive integers represent rightward displacements and negative numbers represent leftward displacements. Multiplying a displacement by a negative number reverses the direction (and possibly scales the magnitude).

Like addition, multiplication is a “binary operation”, satisfying the following conditions.

- **Associativity:** Regrouping a three-fold product does not affect the final value.

For all integers a , b , and c , we have $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.

- **Identity Element:** There exists an integer $1 \neq 0$ that has no effect upon multiplication.

For all integers a , we have $a \cdot 1 = 1 \cdot a = a$.

- **Commutativity:** The order of factors is immaterial.

For all integers a and b , $a \cdot b = b \cdot a$.

- **Distributivity:** Multiplication distributes over sums.

For all integers a , b , and c , $a \cdot (b + c) = a \cdot b + a \cdot c$.

Remark 3.5. The fourth axiom is sometimes called “left-distributivity”, meaning multiplication *on the left* distributes over addition. Since integer multiplication is commutative, we get “right-distributivity” as an easy consequence:

For all integers a , b , and c , $(b + c) \cdot a = b \cdot a + c \cdot a$. (Show this.)

Remark 3.6. Though addition and multiplication of integers are close cousins, their relationship is asymmetrical. First, not every integer has a multiplicative inverse; the equation $2x = 1$ has no integer solutions, for instance. Second, addition does not distribute over multiplication; the equation $a + (bc) = (a + b)(a + c)$ (the distributive law with the roles of addition and multiplication exchanged) is not an identity. Indeed, this equation holds if and only if $a = 0$ or $a + b + c = 1$, a claim you may enjoy proving with elementary algebra.

Ordering and Well-Ordering

The last major clause in the integer contract concerns ordering of integers, which our shepherds first encountered in regard to the relative sizes of their flocks. Our formulation of ordering abstractly characterizes a certain subset \mathbf{Z}^+ of \mathbf{Z} . Before giving the axioms for \mathbf{Z}^+ , we formally introduce some familiar names.

Definition 3.7. The set \mathbf{Z}^+ is the set of *positive* integers.

The set $-\mathbf{Z}^+ = \{b \text{ in } \mathbf{Z} : -b \in \mathbf{Z}^+\}$ is the set of *negative* integers.

The set $\mathbf{N} = \{0\} \cup \mathbf{Z}^+$ is the set of *non-negative* integers, or *natural numbers*.*

- Trichotomy: Every integer is positive, negative, or zero, and these categories are mutually exclusive.

For every integer a , *exactly one* of the following holds: (i) $a \in \mathbf{Z}^+$; (ii) $-a \in \mathbf{Z}^+$; (iii) $a = 0$.

- Sums: A sum of positive integers is positive.

If a and b are elements of \mathbf{Z}^+ , then $a + b \in \mathbf{Z}^+$.

- Products: A product of positive integers is positive.

If a and b are elements of \mathbf{Z}^+ , then $a \cdot b \in \mathbf{Z}^+$.

- Well-Ordering: Every non-empty set of \mathbf{N} has a smallest element.

The well-ordering property plays a central role throughout algebra. Some auxiliary definitions will help bridge the gap between our intuitive idea of the integers as a “ordered list” and the formal axioms.

Definition 3.8. Let a and b be arbitrary integers, and consider the integer $c = a + (-b) = a - b$. By trichotomy, exactly one of the following is true: $c \in \mathbf{Z}^+$, $-c \in \mathbf{Z}^+$, or $c = 0$.

*Many authors define $\mathbf{N} = \mathbf{Z}^+$, and do not consider 0 to be a “natural number”.

If $c = a - b \in \mathbf{Z}^+$, we say a is *greater than* b , and write $a > b$.

If $-c = -(a - b) \in \mathbf{Z}^+$, we say a is *less than* b , and write $a < b$.

If $a < b$ or $a = b$, namely if $-(a - b) \in \mathbf{N}$, we say a is *less than or equal to* b , and write $a \leq b$.

We can now give the fourth axiom precise meaning: If $A \subseteq \mathbf{N}$ is non-empty, there exists an a_0 in A such that $a_0 \leq a$ for every a in A . This meaning of “smallest” mirrors its meaning in common English.

3.3 Consequences of the Axioms

We first show that the identity elements for addition and multiplication, and additive inverses, are uniquely defined. To accomplish this, we use a mathematical idiom you should absorb: Assume two integers satisfy some property, and prove they are equal.

Theorem 3.9. *The additive identity element and the additive inverse of an arbitrary integer a are unique. Precisely:*

- (i) *The integer 0 is uniquely defined by A1. and A2.*
- (ii) *The integer 1 is uniquely defined by M1. and M2.*
- (iii) *The integer $-a$ is uniquely defined by A1., A2. and A3.*
- (iv) *For every integer a , $-(-a) = a$. In particular, $-(-1) = 1$.*

Proof. (i) Suppose integers 0 and $0'$ satisfy A3. By A3. with $a = 0'$, we have $0 + 0' = 0'$. However, by A3. with $a = 0$, we have $0 + 0' = 0$. Combining, $0 = 0 + 0' = 0'$. The proof of (ii) is entirely analogous, and is left to you.

(iii) Assume $a \in \mathbf{Z}$. If $a + b = b + a = 0$ and $a + c = c + a = 0$, then

$$\begin{array}{ll}
 b = b + 0 & \text{A2. with } a = b \\
 = b + (a + c) & a + c = 0 \text{ by hypothesis} \\
 = (b + a) + c & \text{A1.} \\
 = 0 + c & b + a = 0 \text{ by hypothesis} \\
 = c & \text{A2. with } a = c.
 \end{array}$$

(iv) Part (iii) constitutes a useful principle: Let a and b be integers. To check whether $b = -a$, it suffices to show $a + b = 0$. But the equation $(-a) + a = 0$ may therefore be interpreted as saying the additive inverse of $(-a)$ is a , which is (iv). \square

Next we turn to some “obvious” properties of multiplication. Keep in mind that while axiom A3. guarantees existence of the integer $-a$ (the additive inverse of a), there is no assumption that $-a = -1 \cdot a$.

Theorem 3.10. *If a is an arbitrary integer, then*

- (i) $a \cdot 0 = 0 \cdot a = 0$.
- (ii) $-1 \cdot a = -a$. In particular, $(-1) \cdot (-1) = -(-1) = 1$.

Proof. (i) Assume a is an arbitrary integer, and let $b = a \cdot 0$. By A2. with $a = 0$, $0 + 0 = 0$. Multiplying both sides by a and using the distributive law M4. gives

$$b + b = a \cdot 0 + a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 = b.$$

Adding $-b$ to each side,

$$b = 0 + b = ((-b) + b) + b = (-b) + (b + b) = (-b) + b = 0,$$

as claimed.

(ii) By part (ii) of Theorem 3.9, the additive inverse of each integer is unique, so to prove $-1 \cdot a = -a$ it suffices to prove $a + (-1 \cdot a) = 0$. However,

$$a + (-1 \cdot a) = (1 \cdot a) + (-1 \cdot a) = (1 + (-1)) \cdot a = 0 \cdot a = 0.$$

Be sure you are able to justify each equality using the axioms and/or properties established earlier. \square

Theorem 3.11. *Let a , b , c , and d be arbitrary integers.*

- (i) *If $a < b$ and $b < c$, then $a < c$.*
- (ii) *$0 < a$ if and only if $-a < 0$.*
- (iii) *If $a < b$ and $0 < c$, then $ac < bc$.*
- (iv) *If $a < b$ and $c < 0$, then $bc < ac$.*

Proof. (i) By definition, $a < b$ means $b + (-a) \in \mathbf{Z}^+$. Similarly, $b < c$ means $c + (-b) \in \mathbf{Z}^+$. By Axiom O2., a sum of elements of \mathbf{Z}^+ is an element of \mathbf{Z}^+ , so

$$c + (-a) = c + (-b + b) + (-a) = (c + (-b)) + (b + (-a)) \in \mathbf{Z}^+,$$

proving $a < c$.

(ii) Suppose $0 < a$, i.e., $a \in \mathbf{Z}^+$, and consider the integer $-a$. By the trichotomy property, $a \neq 0$, and exactly one of the following is true: $-a = 0$, $-a \in \mathbf{Z}^+$, or $-a \in -\mathbf{Z}^+$. We will show that the first two of these statements are false.

If $-a = 0$, then $a = a + 0 = a + (-a) = 0$. Contrapositively, $a \neq 0$ implies $-a \neq 0$.

If $-a \in \mathbf{Z}^+$, then $0 = a + (-a) \in \mathbf{Z}^+$ by Axiom O2., since $a \in \mathbf{Z}^+$. But $0 \notin \mathbf{Z}^+$ by the trichotomy property, so $-a \notin \mathbf{Z}^+$.

The third condition, $-a \in -\mathbf{Z}^+$, must therefore hold.

Conversely, suppose $a < 0$, and let $b = -a$. Since $-b = a$ by Theorem 3.9 (iv), $b \in \mathbf{Z}^+$ by definition of \mathbf{Z}^+ .

(iii) By hypothesis, $b + (-a) \in \mathbf{Z}^+$ and $c = c + (-0) \in \mathbf{Z}^+$. Axiom O3. guarantees their product is an element of \mathbf{Z}^+ :

$$b \cdot c + (-a) \cdot c = (b + (-a)) \cdot c \in \mathbf{Z}^+.$$

Since $(-a) \cdot c = (-1 \cdot a) \cdot c = -1 \cdot (ac) = -ac$, we have $ac < bc$. The proof of (iv) is entirely analogous: Use the same argument with c replaced by $-c$ in \mathbf{Z}^+ . \square

As an application, we prove $a < b$ if and only if $b > a$. Let a and b be arbitrary distinct integers, and let $c = b + (-a)$, so $c \neq 0$. Now, $-c = a + (-b)$ by Theorem 3.9 (iii), since

$$\begin{aligned} (a + (-b)) + ((-a) + b) &= a + ((-b) + (-a)) + b && \text{A1.} \\ &= a + ((-a) + (-b)) + b && \text{A4.} \\ &= (a + (-a)) + ((-b) + b) && \text{A1.} \\ &= 0 + 0 = 0. && \text{A3. and A2.} \end{aligned}$$

Using part (ii) of the preceding theorem, $a < b$ if and only if $0 < c$, if and only if $0 > -c$, if and only if $b > a$, as was to be shown.

Theorem 3.12. *Let a , b , and c be arbitrary integers.*

- (i) *If $a \neq 0$, then $0 < a^2$. In particular, $0 < 1$.*
- (ii) *If $ab = 0$, then $a = 0$ or $b = 0$.*
- (iii) *If $a \neq 0$ and $ab = ac$, then $b = c$.*
- (iv) *If $0 < a$, then $1 \leq a$. In words, 1 is the smallest positive integer.*

Proof. (i) If $a \neq 0$, then by trichotomy and part (ii) of the preceding theorem, either $a \in \mathbf{Z}^+$ or $a \in -\mathbf{Z}^+$. In the first case, $a^2 = a \cdot a \in \mathbf{Z}^+$ by Axiom O3. In the second case, $-a \in \mathbf{Z}^+$, so

$$a^2 = ((-1) \cdot (-1)) \cdot (a \cdot a) = ((-1) \cdot a)^2 = (-a)^2 \in \mathbf{Z}^+.$$

In either case, $a^2 \in \mathbf{Z}^+$, or $0 < a^2$. Since $1 = 1^2$ by M2., $0 < 1$.

(ii) We prove the contrapositive: If $a \neq 0$ and $b \neq 0$, then $ab \neq 0$. By trichotomy, it suffices to consider four cases: $0 < a$ and $0 < b$; $0 < a$ and $b < 0$; $a < 0$ and $0 < b$; $a < 0$ and $b < 0$. The proofs are similar, so the details of one case will convey the idea.

If $0 < a$ and $b < 0$, then $0 < -b$ by part (ii) of Theorem 3.11. Since a product of positive numbers is positive, $0 < a(-b) = -(ab)$. Invoking part (ii) of Theorem 3.11 again, $ab < 0$; in particular, $ab \neq 0$.

(iii) Since $ab = ac$, we have $0 = ab - ac = a \cdot (b - c)$ by the distributive axiom. The preceding part of this theorem implies $a = 0$ or $b - c = 0$. Since $a \neq 0$ by hypothesis, $b - c = 0$, namely $b = c$.

(iv) Since the set $\mathbf{Z}^+ = \mathbf{N} \setminus \{0\}$ of positive integers is non-empty, there exists a smallest positive integer a_0 by the well-ordering axiom. Since 1 is positive by part (i) of this theorem, $a_0 \leq 1$.

By Axiom O3., $0 < a_0^2$. Since a_0 is the smallest positive integer, we have $a_0 \leq a_0^2$, or $0 \leq a_0^2 - a_0 = a_0(a_0 - 1)$. By parts (iii) and (iv) of Theorem 3.11, we have $a_0 - 1 \geq 0$, or $1 \leq a_0$.

Since $a_0 \leq 1$ and $1 \leq a_0$, we have $a_0 = 1$. □

Remark 3.13. The proof of part (iii) illustrates a useful idiom: To prove two integers are equal, show their difference is equal to zero.

Example 3.14. Suppose a is an integer such that $a^2 = a$. Subtracting and factoring, $a(a - 1) = a^2 - a = 0$. By part (ii), either $a = 0$, or $a - 1 = 0$, i.e., $a = 1$. This conclusion is no surprise, but naive manipulation of the axioms is unlikely to yield as concise a proof.

3.4 The Division Algorithm

Suppose N objects, such as jelly beans or playing cards, are to be divided among n people. Every child learns the algorithm: Put the people into some fixed order (such as left to right around a circle). Following the order cyclically, give one object to each person in succession until

none remain. This process “minimizes unfairness” in that either everyone receives the same number q of objects, or else everyone receives at least q objects, but some number r (with $0 < r < n$) receive one extra.

Mathematically, this process is known as the *division algorithm*, Figure 3.1. The numbers q and r are the *quotient* and *remainder* of N on division by n .

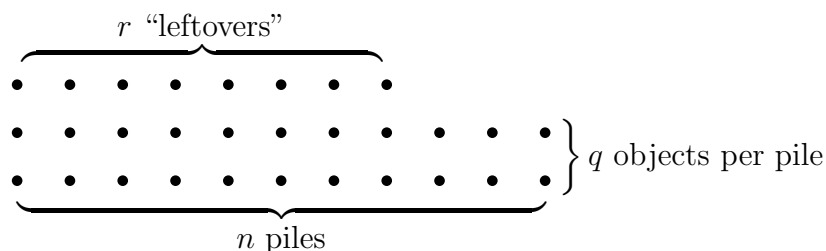


Figure 3.1: Dividing N objects among n piles.

Theorem 3.15. Assume $N \in \mathbf{Z}$ and $n \in \mathbf{Z}^+$. There exist unique integers q and r , with $0 \leq r < n$, such that $N = nq + r$.

The proof formalizes the naive algorithm. If $N > 0$, repeatedly subtract n until the result becomes negative, and say this occurs after $q + 1$ subtractions. At the preceding step, the remainder (i.e., the number of “leftovers”) $r = N - nq$ must have been between 0 and $n - 1$. If $N \leq 0$, argue similarly, but *add* n repeatedly until, after q additions, the result is positive.

Proof. (Existence). Let N in \mathbf{Z} and $n > 0$ be given. Consider the set S of integers of the form $N - nk$ with k in \mathbf{Z} , namely, integers that can be obtained from N by adding or subtracting n repeatedly. Let $S^+ = S \cap \mathbf{N}$ be the set of non-negative integers in S .

The set S^+ is non-empty: If $0 \leq N$, then $N = N - n \cdot 0 \in S^+$. If instead $N < 0$, then since $1 - n \leq 0$, we have $0 \leq (1 - n)N = N - nN$, i.e., $N - nN \in S^+$.

By the well-ordering principle, the non-empty set S^+ has a smallest element, say r . Since $r \in S$, by definition there exists an integer q such that $r = N - nq$, or $N = nq + r$.

Because r is the *smallest* element of S^+ , we have $0 \leq r < n$: If instead $n \leq r$ were true, it would follow that $0 \leq r - n < r$, which would mean $r - n$ in S^+ is smaller than r . This completes the proof of the “existence” part of the theorem.

(Uniqueness). We wish to show that N can be written in only one way as $N = nq + r$ with $0 \leq r < n$.

Suppose $N = nq_1 + r_1 = nq_2 + r_2$ with $0 \leq r_1 < n$ and $0 \leq r_2 < n$. We may assume $r_1 \leq r_2$ without loss of generality; if not, swap the names of these numbers. Rearranging,

$$nq_1 - nq_2 = n(q_1 - q_2) = r_2 - r_1.$$

The left-hand side is an integer multiple of n . The right-hand side is non-negative, but no larger than $r_2 < n$. Their common value is therefore a non-negative integer multiple of n that is strictly smaller than n , namely zero. In other words, $r_1 = r_2$ and $q_1 = q_2$. This completes the proof of uniqueness. \square

Remark 3.16. The conclusion of the division algorithm may look peculiar when $N < 0$. For example, if $N = -30$ and $n = 11$, the division algorithm gives $-30 = -3 \cdot 11 + 3$, while most peoples' intuition gives $-30 = -2 \cdot 11 - 8$ (cf. Figure 3.1). Both equations are correct, of course, but the condition $0 \leq r < n$ forces us to use $-3 \cdot 11 + 3$ as the unique representation of -30 as a multiple of 11 plus a remainder.

Definition 3.17. Let N be an integer. We say N is *even* if there exists an integer q such that $N = 2q$. We say N is *odd* if there exists an integer q such that $N = 2q + 1$.

Remark 3.18. By Theorem 3.15, every integer is either even or odd, and no integer is both.

Exercises

Exercise 3.1. In each part, a denotes an integer satisfying an equation. Solve for a , giving a proof based on the axioms and theorems in this chapter.

$$(a) \ a^2 = 1. \qquad (b) \ a^3 = a. \qquad (c) \ a^4 = 1.$$

Exercise 3.2. Let a and b be integers. Use the axioms to prove:

$$\begin{aligned} (a) \ (a - b)(a + b) &= a^2 - b^2. \\ (b) \ (a + b)^2 - (a - b)^2 &= 4ab. \\ (c) \ (a + b)^2 + (a - b)^2 &= 2(a^2 + b^2). \end{aligned}$$

- (d) Suppose a and b are not integers, but entities satisfying all the additive and multiplicative axioms except the commutative law of multiplication. (That is, suppose $ba \neq ab$ in general.) Expand the left-hand side of the identity in each of parts (a)–(c).

Exercise 3.3. An integer N is a *square* if there exists an integer n such that $N = n^2$. Prove that:

- (a) Squares of consecutive integers differ by an odd integer.
- (b) Every square leaves a remainder of 0 or 1 on division by 4.
- (c) If m and n are integers, not both even, then $n^2 \neq 2m^2$.

Exercise 3.4. Suppose n_1 and n_2 are integers that leave respective remainders of r_1 and r_2 on division by 2.

- (a) What remainder does $n_1 + n_2$ leave on division by 2?
- (b) What remainder does $n_1 n_2$ leave on division by 2?
- (c) In each of the preceding parts, make a 2×2 table to list all possible outcomes.

Exercise 3.5. Suppose n_1 and n_2 are integers that leave respective remainders of r_1 and r_2 on division by 5. Describe the remainders left by $n_1 + n_2$ and $n_1 n_2$ on division by 5.

Exercise 3.6. In each part, let a , b , c , and d be integers.

- (a) If $a < b$ and $c < d$, then $a + c < b + d$.
- (b) If $0 < a < b$ and $0 < c < d$, then $0 < ac < bd$.

Exercise 3.7. Consider the set \mathbf{C} of complex numbers equipped with the operations of complex addition and multiplication. Show that there exists no subset \mathbf{C}^+ that is closed under addition and multiplication and satisfies the trichotomy axiom.

Hint: Start by showing that if \mathbf{C}^+ is closed under multiplication, then every perfect square is an element of \mathbf{C}^+ .

Exercise 3.8. Let a and b be integers. Prove there exist integers u and v such that $u + v = a$ and $u - v = b$ if and only if a and b are both even or both odd, and find formulas for u and v in terms of a and b .

Exercise 3.9. If $a \in \mathbf{Z}$, define the *absolute value* of a by

$$|a| = \begin{cases} a & \text{if } 0 \leq a \\ -a & \text{if } a < 0 \end{cases}$$

For arbitrary integers a and b , prove:

- (a) $|-a| = |a|$. (b) $|ab| = |a| \cdot |b|$.
 (c) $-|a| \leq a \leq |a|$. (d) $|b| \leq a$ if and only if $-a \leq b \leq a$.

- (e) $|a + b| \leq |a| + |b|$. (The *triangle inequality*.)

Hint: Apply part (c) to the integers a and b separately, add the inequalities, and use part (d).

- (f) $||a| - |b|| \leq |a - b|$. (The *reverse triangle inequality*.)

Hint: First apply the triangle inequality to $a = (a - b) + b$ and to $b = a + (b - a)$.

Exercise 3.10. Let a and b be integers. Define the *minimum* and *maximum* of a and b by

$$\min(a, b) = \begin{cases} a & \text{if } a \leq b, \\ b & \text{if } b < a, \end{cases} \quad \max(a, b) = \begin{cases} b & \text{if } a \leq b, \\ a & \text{if } b < a. \end{cases}$$

- (a) Prove $a + b = \min(a, b) + \max(a, b)$.

Give two proofs: One involving verbal explanation, and one based on manipulation of formulas.

- (b) Prove $|a - b| = \max(a, b) - \min(a, b)$.

Suggestion: Again, give two proofs. For the algebraic proof, consider two cases, $a \leq b$ and $b < a$.

- (c) Use parts (a) and (b) and Exercise 3.8 to find algebraic formulas for $\min(a, b)$ and $\max(a, b)$.

Chapter 4

Mappings and Relations

As in calculus, a “function” or “mapping” is a rule associating a unique “output” to each “input”. While this intuitive description is adequate for informal work, rigorous mathematics requires more precision: The sets of allowable inputs and potential outputs must be made an intrinsic part of a function.

Example 4.1. Consider the familiar squaring function $f(x) = x^2$, where x ranges over the set of real numbers. If we set $y = f(x)$, we might wish to “solve” for x in terms of y . At first glance this is trivial: set $x = \sqrt{y}$. Unfortunately, closer inspection reveals two fatal flaws. First, if $y < 0$, there is no real x satisfying $x^2 = y$. In this context, the square root is *undefined*. Second, if $y > 0$, there exist *two* values of x with $x^2 = y$; the input x is not a function of the output y , so the square root is *not well-defined*. In either event, we have not associated a unique output to each input.

In high school, you learned to avoid complications with square roots by only considering non-negative numbers y , and agreeing that \sqrt{y} always refers to the non-negative square root. Technically you are no longer inverting the function $f(x) = x^2$ with x real, but a *different function defined by the same formula*, for which the allowable inputs and potential outputs have been explicitly restricted.

Remark 4.2. The squaring function is arguably artificial in this respect, but for other familiar functions, such as the circular trig functions, the inability to invert causes genuine annoyances. Consider longitude (measured in degrees) as a function of position on the earth. Upon circumnavigating the earth to the east, longitude increases by 360° . But this cannot be the whole story; if it were, each geographic location

would have multiple longitudes, any two differing by a whole multiple of 360° .

Instead, when you circumnavigate the globe in an eastward direction, you must cross a line where longitude “jumps down” by 360° . This discontinuity is a mathematical artifact of the impossibility of inverting sine and cosine to recover longitude continuously as a real-valued function of position on the earth.

The earth is approximately spherical and rotates with respect to the distant stars. A *sidereal day*, or 24 hours, is the time required for the earth to rotate 360° with respect to the stars. This *duration* is the same for all points on the earth, but the *starting time* (midnight) depends on one’s longitude. By international treaty, the earth’s surface is divided into twenty-four *time zones*, each a sector of longitude 15° wide (with substantial allowances for geographical and political boundaries). The times in neighboring zones differ by one hour.

The global discontinuity of longitude has a notable practical consequence: the existence of the International Date Line, an imaginary “cut” along the surface of the earth joining the south and north poles, along which local time “jumps” by 24 hours, affecting global travelers and international stock traders alike.

4.1 Mappings, Images, and Preimages

Before giving the formal definition of a mapping, we need to construct an appropriate set universe.

Definition 4.3. Let A and B be sets. Their *Cartesian product* $A \times B$ is the set of all “ordered pairs” from A and B ,

$$A \times B = \{(a, b) : a \in A \text{ and } b \in B\}.$$

Example 4.4. The Cartesian plane $\mathbf{R} \times \mathbf{R} = \mathbf{R}^2$ is the set of ordered pairs of real numbers. Similarly, $\mathbf{R} \times \mathbf{R} \times \mathbf{R}$, or \mathbf{R}^3 , is Cartesian space, the set of ordered triples of real numbers.

Example 4.5. If $A = \{a, b, c\}$ and $B = \{0, 1\}$, then $A \times B$ is the six-element set $\{(a, 0), (b, 0), (c, 0), (a, 1), (b, 1), (c, 1)\}$ in the left-hand diagram in Figure 4.1.

For the same set B , $B \times B = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$.

Example 4.6. If $A = \emptyset$ or $B = \emptyset$, then $A \times B = \emptyset$.

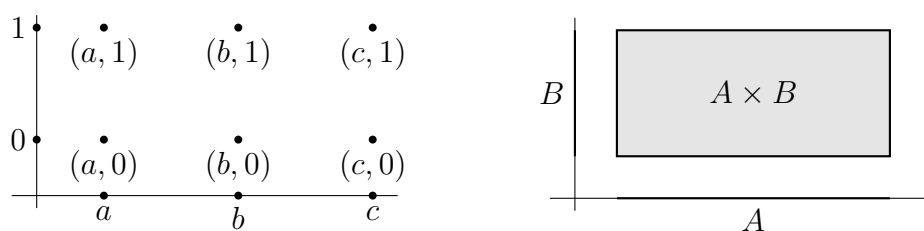


Figure 4.1: Cartesian products.

An abstract Cartesian product can be visualized conveniently by depicting the set A on a horizontal axis and the set B on a vertical axis, and taking the set of points lying above or below A and to the left or right of B . The right-hand diagram in Figure 4.1 shows the case where A and B are intervals.

Definition 4.7. A mapping f from A to B is a subset f of $A \times B$ satisfying the following condition:

For every a in A , there exists a unique b in B such that $(a, b) \in f$.

The set A is the *domain* of f , and B is the *target*. We write $b = f(a)$, and call b the *value* of f at a . We also say a is *mapped to* b by f , or that f *maps* a to b .

The notation $f : A \rightarrow B$, read “ f from A to B ”, signifies that f is a mapping from A to B . A mapping $f : A \rightarrow B$ associates a unique value b in the target to each element a of the domain. If the Cartesian product $A \times B$ is viewed as a rectangle, a mapping is a “graph” in the sense of calculus, namely a subset intersecting each vertical line in the rectangle exactly once. The vertical line at horizontal position a intersects f (a.k.a. the graph of f) at location $b = f(a)$.

Remark 4.8. If B is arbitrary (empty or not), there is a unique mapping $f : \emptyset \rightarrow B$, namely the empty set.

If A is non-empty, there exists no mapping $f : A \rightarrow \emptyset$.

Definition 4.9. Let $f : A \rightarrow B$ be a mapping. If $S \subseteq A$, we define the *image* of S under f to be the set

$$f(S) = \{b \text{ in } B : b = f(s) \text{ for some } s \text{ in } S\} \subseteq B,$$

see Figure 4.2. In particular, the *image of* f is the set $f(A) \subseteq B$ of all values of f .

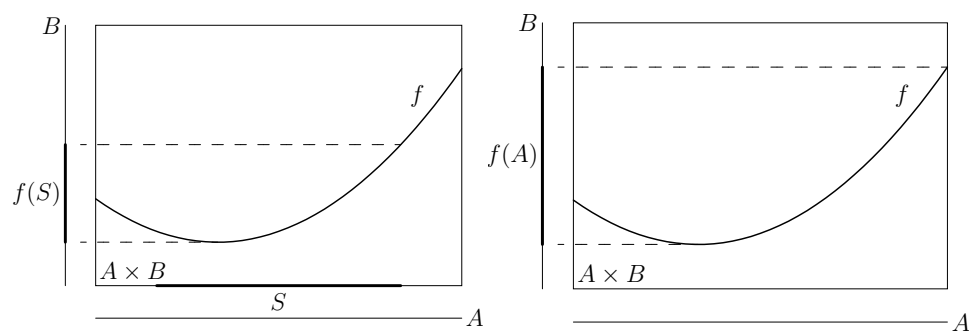


Figure 4.2: The image of a set under a mapping.

Definition 4.10. Let $f : A \rightarrow B$ be a mapping. If $T \subseteq B$, we define the *preimage* of T under f to be the set

$$f^{-1}(T) = \{a \text{ in } A : f(a) \in T\} \subseteq A$$

of elements of the domain mapped into T by f , see Figure 4.3.

Remark 4.11. A mapping $f : A \rightarrow B$ may be viewed as a “poll” taken of a population A , with responses in the set B . The image under f of a set $S \subseteq A$ is the set of responses from individuals in S . The preimage of a set $T \subseteq B$ is the set of individuals whose responses are in T .

Example 4.12. If A is a non-empty set, we define the *identity mapping* $I_A : A \rightarrow A$ by $I_A(a) = a$ for all a in A . Under the identity map, every set is its own image, and its own preimage.

Example 4.13. Let A and B be non-empty sets. For each b in B , there is a *constant mapping* $c_b : A \rightarrow B$ defined by $c_b(a) = b$ for all a in A .

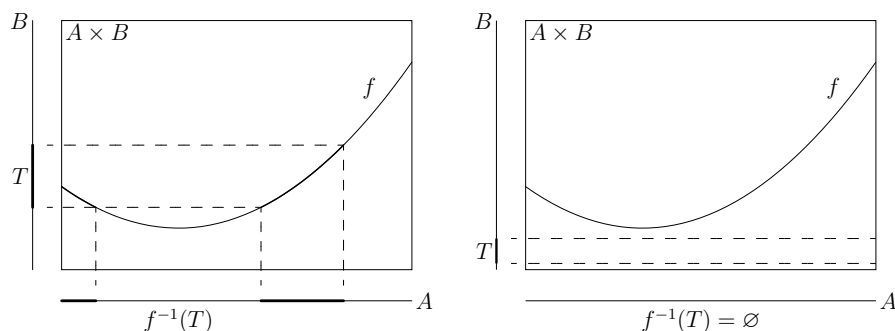


Figure 4.3: The preimage of a set.

The image of an arbitrary non-empty subset of A is the singleton $\{b\}$. The preimage of a set T is either the empty set (if $b \notin T$) or the entire domain A (if $b \in T$).

Proposition 4.14. *Let $f : A \rightarrow B$ be a mapping, S_1 and S_2 subsets of A , and T_1 and T_2 subsets of B . Then*

- (i) $f(S_1 \cup S_2) = f(S_1) \cup f(S_2)$.
- (ii) $f(S_1 \cap S_2) \subseteq f(S_1) \cap f(S_2)$.
- (iii) $f^{-1}(T_1 \cup T_2) = f^{-1}(T_1) \cup f^{-1}(T_2)$.
- (iv) $f^{-1}(T_1 \cap T_2) = f^{-1}(T_1) \cap f^{-1}(T_2)$.

Proof. To prove two sets are equal, we must establish inclusions in both directions. Assume S_1 and S_2 are subsets of A .

(The inclusion $f(S_1 \cup S_2) \subseteq f(S_1) \cup f(S_2)$). If $b \in f(S_1 \cup S_2)$, then by definition there exists an element a in $S_1 \cup S_2$ such that $f(a) = b$. Since either $a \in S_1$ or $a \in S_2$ by definition of the union of sets, either $b \in f(S_1)$ or $b \in f(S_2)$, which means $b \in f(S_1) \cup f(S_2)$. This proves $f(S_1 \cup S_2) \subseteq f(S_1) \cup f(S_2)$.

(The inclusion $f(S_1) \cup f(S_2) \subseteq f(S_1 \cup S_2)$). If $b \in f(S_1) \cup f(S_2)$, there exists an a in $S_1 \subseteq S_1 \cup S_2$ such that $f(a) = b$ or there exists an a in $S_2 \subseteq S_1 \cup S_2$ such that $f(a) = b$. In either case, there exists an a in $S_1 \cup S_2$ such that $f(a) = b$, which means $b \in f(S_1 \cup S_2)$. This proves $f(S_1) \cup f(S_2) \subseteq f(S_1 \cup S_2)$.

The other parts are entirely similar, and are left to you. □

4.2 Surjectivity and Injectivity

Our inability to invert the map $f : \mathbf{R} \rightarrow \mathbf{R}$, $f(x) = x^2$, had two aspects: When we wrote $y = f(x)$, some y were associated with no values of x , and some were associated with multiple values of x .

Definition 4.15. A mapping $f : A \rightarrow B$ is *surjective* if for every b in B , there exists an a in A such that $f(a) = b$.

We may use the term *onto* with identical meaning. Algebraically, f is surjective if, regardless of b , the equation $f(a) = b$ can be solved (perhaps non-uniquely) for a . In other words, f is surjective if every element of the target is a value of f .

In terms of sets, f is surjective if the preimage of $T = \{b\}$ is non-empty for each b in B , or the image of f is the entire target, $f(A) = B$, see Figure 4.4.

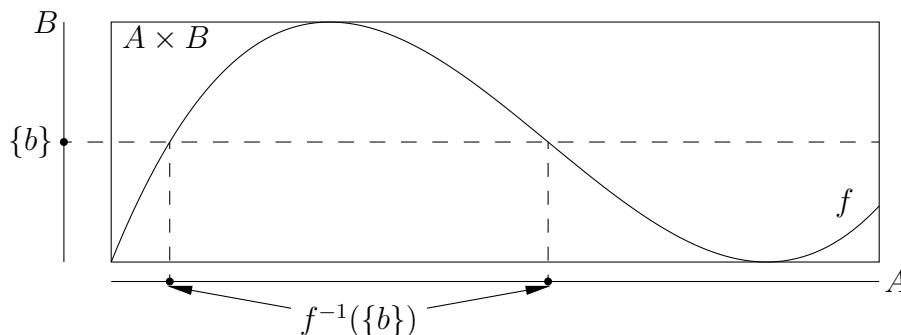


Figure 4.4: The preimage of a point under a surjective mapping.

Geometrically, a mapping $f : \mathbf{R} \rightarrow \mathbf{R}$ is surjective if every horizontal line hits the graph of f at least once.

Definition 4.16. Let $f : A \rightarrow B$ be a mapping. Points a_1 and a_2 in A are *identified by f* if $f(a_1) = f(a_2)$, namely if a_1 and a_2 are mapped to the same value by f .

Definition 4.17. A mapping f is *injective* if $f(a_1) = f(a_2)$ implies $a_1 = a_2$. Contrapositively, if $a_1 \neq a_2$, then $f(a_1) \neq f(a_2)$.

We sometimes use the phrase *one-to-one* with identical meaning. A mapping f is injective if and only if f does not identify any pairs of distinct elements. Equivalently, the preimage of an arbitrary singleton $T = \{b\}$ contains at most one element.

Geometrically, $f : \mathbf{R} \rightarrow \mathbf{R}$ is injective if every horizontal line hits the graph of f at most once.

Remark 4.18. Continuing Remark 4.11, a mapping $f : A \rightarrow B$ is surjective if every allowable answer to the poll is given by at least one individual. Similarly, f is injective if no two people give the same response; knowledge of the response uniquely determines the individual who gave that response.

Definition 4.19. A mapping $f : A \rightarrow B$ is *bijective* if f is both surjective and injective.

Remark 4.20. If $f : A \rightarrow B$ is bijective, then each element a in A corresponds to exactly one element b in B . For this reason, many authors use the phrase “one to one correspondence” to connote a bijection. We avoid this name, since it can be easily confused with a “one to one mapping”, which is not quite the same thing.

Remark 4.21. Algebraically, if $f : A \rightarrow B$ is bijective, the equation $f(a) = b$ can be solved uniquely for each b in B . Procedurally, f “relabels” elements of the set A using elements of B as names.

Example 4.22. Define $f_1 : \mathbf{R} \rightarrow [0, \infty)$ by $f_1(x) = x^2$. This mapping is surjective (every non-negative real y can be written as $x^2 = f_1(x)$ for at least one real x), but not injective (since $f_1(-1) = 1 = f_1(1)$, but $-1 \neq 1$).

Example 4.23. Define $f_2 : (0, \infty) \rightarrow \mathbf{R}$ by $f_2(x) = x^2$. This mapping is not surjective (there is no real x such that $x^2 = f_2(x) = -1$), but *is* injective. To establish injectivity, suppose $a_1^2 = f_2(a_1) = f_2(a_2) = a_2^2$. Subtracting and factoring, we find $0 = a_2^2 - a_1^2 = (a_2 - a_1)(a_2 + a_1)$, which implies $a_1 = a_2$ or $a_1 = -a_2$. The latter is impossible since a_1 and a_2 are positive by hypothesis.

We have shown that if $f_2(a_1) = f_2(a_2)$, then $a_1 = a_2$. Since a_1 and a_2 were arbitrary, f_2 is injective.

Note carefully that the mappings f_1 and f_2 in these examples are defined by the same formula, but have distinct domains and/or targets.

Example 4.24. Let $\zeta = e^{2\pi i/3} = \frac{1}{2}(-1 + i\sqrt{3})$, and consider the set $A = \{1, \zeta, \zeta^2\} \subseteq \mathbf{C}^\times$, the set of non-zero complex numbers. Since ζ is a cube root of unity, $(\zeta^2)^2 = \zeta^4 = \zeta$ and $(\zeta^2)^3 = 1$.

The mapping $f : A \rightarrow A$ defined by $f(z) = z^2$ is bijective: $1 = f(1)$, $\zeta = f(\zeta^2)$, and $\zeta^2 = f(\zeta)$.

The mapping $g : A \rightarrow A$ defined by $f(z) = z^3$ is neither injective nor surjective. Indeed, $f(z) = 1$ for every z in A .

Example 4.25. Define $f : \mathbf{Z} \rightarrow \mathbf{Z}$ by $f(a) = 1 - a$. Prove f is bijective.

(Injectivity). Let a_1 and a_2 be arbitrary integers, and assume that $f(a_1) = f(a_2)$. By the definition of f , $1 - a_1 = 1 - a_2$, so $a_1 = a_2$ by elementary algebra. Since a_1 and a_2 were arbitrary, f is injective.

(Surjectivity). Informally, we wish to solve $b = f(a) = 1 - a$ for a in terms of b . Rearrangement gives $a = 1 - b$.

Formally, let b be an arbitrary integer, and consider the integer $a = 1 - b$. Since $f(a) = f(1 - b) = 1 - (1 - b) = b$, we have shown that for every integer b , there exists an integer a such that $f(a) = b$. This means f is surjective.

Example 4.26. Let $f : \mathbf{Z} \rightarrow \mathbf{Z}$ be defined by $f(a) = 1 - 2a$. Prove f is injective (one-to-one) but not surjective (onto).

(Injectivity). Let a_1 and a_2 be integers, and assume $f(a_1) = f(a_2)$, i.e., $1 - 2a_1 = 1 - 2a_2$. Subtracting the left side from the right gives $0 = 2a_1 - 2a_2 = 2(a_1 - a_2)$. By Theorem 3.12 (ii), $a_1 - a_2 = 0$ as well. Since $f(a_1) = f(a_2)$ implies $a_1 = a_2$, f is injective.

(Non-surjectivity). To show f is not surjective, it suffices to exhibit an integer not in the image of f . Let $b = 0$. The equation $f(a) = b$ becomes $1 - 2a = 0$, or $1 = 2a$. There exists no integer a satisfying this condition, which means 0 is not in the image of f .

Example 4.27. Define $f : \mathbf{Z}^+ \rightarrow \mathbf{Z}$ by

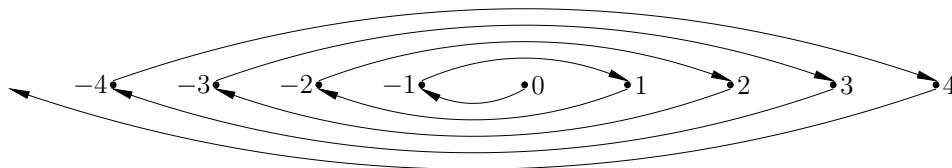
$$f(a) = \begin{cases} \frac{a-1}{2} & \text{if } a \text{ is odd,} \\ -\frac{a}{2} & \text{if } a \text{ is even.} \end{cases}$$

Prove f is bijective. (Informally, there are just as many positive integers as there are integers!)

(Initial exploration). To understand f intuitively, list its first several values. The inputs (elements of the domain) are 1, 2, 3, 4, \dots . To find an output, determine whether the input is even or odd, and evaluate the corresponding formula. Thus $f(1) = 0$ (1 is odd), $f(2) = -1$ (2 is even), $f(3) = 1$, $f(4) = -2$, $f(5) = 2$, and so forth:

a	1	2	3	4	5	6	7	8	9
$f(a)$	0	-1	1	-2	2	-3	3	-4	4

In words, f alternately “counts off” one negative, one positive. Using arrows to indicate successive values:



Since the same value is never achieved twice, f is injective. Since every integer value is achieved, f is surjective. We must convert this intuition into a formal proof.

(Injectivity). Let a_1 and a_2 be integers, and assume $f(a_1) = f(a_2)$. Because f is defined “piecewise”, it’s most convenient to consider three separate cases.

Case 1: a_1 and a_2 both odd. By hypothesis and the definition of f , $(a_1 - 1)/2 = (a_2 - 1)/2$, and elementary algebra implies $a_1 = a_2$.

Case 2: a_1 and a_2 both even. Here, $-a_1/2 = -a_2/2$, and again we find $a_1 = a_2$.

Case 3: a_1 and a_2 have opposite parity (one is odd, one is even). Without loss of generality, we may assume a_1 is odd and a_2 is even. (Otherwise, swap their names.) Since $f(a_2) < 0 \leq f(a_1)$, the hypothesis $f(a_1) = f(a_2)$ is false. Said contrapositively, if $f(a_1) = f(a_2)$, we are not in Case 3.

Since the conclusion $a_1 = a_2$ followed in each case, we have shown f is injective.

(Surjectivity). Let b be an arbitrary integer, and consider two cases:

Case 1: $b < 0$. Let $a = -2b$. Since a is an even integer, we have $f(a) = -a/2 = b$; there exists an a such that $f(a) = b$ provided $b < 0$.

Case 2: $0 \leq b$. Let $a = 1 + 2b$. Since a is odd, $f(a) = (a - 1)/2 = 2b/2 = b$; there exists an a such that $f(a) = b$ provided $0 \leq b$.

Since every integer b is either negative or non-negative, we have handled all possibilities. In each case, there exists an integer a such that $f(a) = b$, so f is onto.

Example 4.28. Let A be an arbitrary set, and let $\mathcal{P}(A)$ be its power set. The following argument of G. Cantor shows there is no surjection $f : A \rightarrow \mathcal{P}(A)$.

Let $f : A \rightarrow \mathcal{P}(A)$ be an arbitrary mapping. For each a in A , the value $f(a)$ is a *subset* of A , so the statement $a \in f(a)$ is meaningful for each a . Let

$$T = \{a \text{ in } A : a \notin f(a)\}.$$

To prove f is not surjective, it suffices to show $f(t) \neq T$ for all t in A . We will prove that if $f(t) = T$ for some t , then set theory is logically inconsistent. Contrapositively, if set theory is logically consistent then $f(t) \neq T$ for all t in A .

If $f(t) = T$, we may ask which alternative is true: $t \notin T$ or $t \in T$. By the definition of T , if $t \in f(t) = T$, then t fails to satisfy the

criterion for membership in T , so $t \notin T$. However, if $t \notin f(t) = T$, then t satisfies the criterion of membership, so $t \in T$. In summary, the statement $t \in T$ is logically equivalent to its negation $t \notin T$. This completes the proof.

4.3 Composition and Inversion

Definition 4.29. Let $f : A \rightarrow B$ and $g : B \rightarrow C$ be mappings. Their *composition* is the mapping $g \circ f : A \rightarrow C$ defined by

$$(g \circ f)(a) = g(f(a)) \quad \text{for each } a \text{ in } A.$$

In this situation we say g is *composable with* f .

Remark 4.30. In words, plug the output of f into g ; the resulting output is $(g \circ f)(a)$.

When context clearly signifies composition of functions, the operator symbol \circ may be omitted, and the composition $g \circ f$ denoted gf .

Proposition 4.31. *Mapping composition is associative: If $f : A \rightarrow B$, $g : B \rightarrow C$, and $h : C \rightarrow D$ are composable mappings, then $h(gf) = (hg)f$ as mappings from A to D .*

Proof. If a is an arbitrary element of A , then

$$\begin{aligned} [h \circ (g \circ f)](a) &= h[(g \circ f)(a)] = h[g(f(a))] \\ &= (h \circ g)(f(a)) = [(h \circ g) \circ f](a). \end{aligned} \quad \square$$

Surjectivity and injectivity of mappings f and g are related to whether or not the composition gf is surjective and/or injective. Think of two functions “cooperating”, with g acting on the output of f . If f achieves every value in B and g achieves every value in C , then in tandem they achieve every value in C . Similarly, if neither g nor f identifies any pair of distinct points, then gf does not either. Before reading further, you should express these observations formally as logical implications and try to prove them.

Proposition 4.32. *Let $f : A \rightarrow B$ and $g : B \rightarrow C$ be mappings.*

- (i) *If f and g are surjective, then gf is surjective.*
- (ii) *If f and g are injective, then gf is injective.*

Proof. (i). Suppose $f : A \rightarrow B$ and $g : B \rightarrow C$ are surjective. Let c in C be arbitrary. Because g is surjective, there exists a b in B such that $g(b) = c$. Since f is surjective, there exists an a in A such that $f(a) = b$. But $(gf)(a) = g(f(a)) = g(b) = c$. We have shown that for every c in C , there exists an a in A such that $(gf)(a) = c$, which by definition means gf is surjective.

(ii). Exercise 4.4 (a). □

Conversely, suppose we know gf is surjective, or that gf is injective. What can we deduce about f and g ?

In our cooperation metaphor, if gf achieves every value in C , then g itself must as well, since any value not achieved by g is certainly not achieved by gf . Thus, if gf is surjective, then g is surjective.

Similarly, if f identifies some pair of points, then gf identifies that pair as well, since g cannot split asunder what f has joined. Formally, if gf is injective, then f is injective, Exercise 4.4 (b).

The following examples show nothing more can be deduced.

Example 4.33. Let $f : [-1, 1] \rightarrow \mathbf{R}$ and $g : \mathbf{R} \rightarrow [-1, 1]$ be defined by $f(x) = \arcsin x$, $g(x) = \sin x$. The mapping f is injective but not surjective (why?), g is surjective but not injective (why?), while $gf : [-1, 1] \rightarrow [-1, 1]$ is the identity map (which is bijective), and $fg : \mathbf{R} \rightarrow \mathbf{R}$ is neither injective nor surjective.

Example 4.34. An arbitrary mapping $f : A \rightarrow B$ can be “factored” into the composition of an injection followed by a surjection. Define $\gamma_f : A \rightarrow A \times B$ and $\pi_2 : A \times B \rightarrow B$ by

$$\gamma_f(a) = (a, f(a)), \quad \pi_2(a, b) = b.$$

Geometrically, “ γ_f lifts a to the graph of f ” and “ π_2 projects $A \times B$ onto the second factor.” Clearly, $f = \pi_2 \circ \gamma_f : A \rightarrow B$, γ_f is injective, and π_2 is surjective.

Inversion of Mappings

Definition 4.35. Let A and B be sets. A mapping $f : A \rightarrow B$ is *invertible* if there exists a mapping $g : B \rightarrow A$ that *inverts* f , i.e., such that $g \circ f$ is the identity map of A and $f \circ g$ is the identity map of B .

Remark 4.36. If $f : A \rightarrow B$ is invertible and $g : B \rightarrow A$ inverts f , then $(g \circ f)(a) = a$ for all a in A and $(f \circ g)(b) = b$ for all b in B . That is,

$$(4.1) \quad \text{For all } a \text{ in } A \text{ and all } b \text{ in } B, g(b) = a \text{ if and only if } f(a) = b.$$

Proposition 4.37. *Let A and B be sets, $f : A \rightarrow B$ a mapping.*

- (i) *f is invertible if and only if f is bijective.*
- (ii) *If f is invertible, there exists a unique map inverting f .*

Remark 4.38. Both conclusions hold (with essentially vacuous proof) if either A or B is the empty set. It suffices to assume A, B are non-empty.

Proof. (i). Assume f is invertible, and let g be a mapping that inverts f , i.e., that satisfies $gf = I_A$ and $fg = I_B$. If $f(a_1) = f(a_2)$ for some a_1 and a_2 in A , applying g to both sides gives $a_1 = a_2$, so f is injective. If b is an arbitrary element of B , and if $a = g(b)$, then $f(a) = (fg)(b) = b$, so f is surjective.

Conversely, suppose f is bijective: For each b in B , there exists a unique a in A such that $b = f(a)$. Define $g(b) = a$. This prescription defines a mapping $g : B \rightarrow A$ that satisfies (4.1), so f is invertible.

- (ii). If $g_1, g_2 : B \rightarrow A$ invert f , then

$$g_1 = g_1 \circ I_B = g_1 \circ (f \circ g_2) = (g_1 \circ f) \circ g_2 = I_B \circ g_2 = g_2. \quad \square$$

A mapping f is invertible if and only if f is injective and surjective. We now consider what happens if each condition holds individually.

Left Inverses

Assume f is one-to-one; not every element of B need be a value of f , but every value (every element of $f(A)$, the image of A under f) is achieved exactly once. We may define $h : f(A) \rightarrow A$ by the analog of (4.1): For all b in $f(A)$, $h(b) = a$ if and only if $f(a) = b$.

If we apply f to a in A , then apply h to $b = f(a)$, we find

$$(hf)(a) = h(f(a)) = h(b) = a \quad \text{for all } a \text{ in } A.$$

That is, $hf = I_A$, the identity map on A ; h is a *left inverse* of f .*

*In general, $fh \neq I_B$, the identity map on B , since (i) h is defined only on the image of f , and (ii) the image of fh , which is a subset of the image of f , may be a *proper* subset of B .

In order to obtain a mapping $g : B \rightarrow A$ satisfying $gf = I_A$, we must “enlarge” the domain of h ; any convenient “rule” will do. For example, pick an element a_0 in A arbitrarily, and define, for b in B ,

$$g(b) = \begin{cases} a & \text{if } b = f(a) \text{ for some } a \text{ in } A \\ a_0 & \text{otherwise} \end{cases}$$

The easy verification that $gf = I_A$ is left to you.

Example 4.39. Define $f : \mathbf{R} \rightarrow \mathbf{R}$ by $f(x) = e^x$, see Figure 4.5, left. For each $y > 0$ (namely, for each y in the image of f), we have $y = e^x$ if and only if $x = \ln y$. Define

$$g(y) = \begin{cases} \ln y & \text{if } y > 0, \\ 0 & \text{if } y \leq 0, \end{cases}$$

see Figure 4.5, right. Then $(gf)(x) = g(f(x)) = x$ for all x ; what about $f(g(y))$?

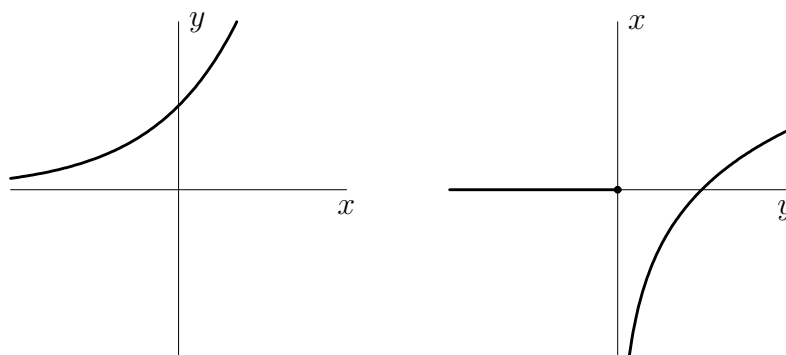


Figure 4.5: A left inverse of $f(x) = e^x$.

Right Inverses

Assume f is onto; every element of B is a value of f , but some values b may be achieved at distinct points: $f(a_1) = f(a_2)$ but $a_1 \neq a_2$. Define $g : B \rightarrow A$ by the following prescription: For each b in B , use the Axiom of Choice to pick an a in A such that $f(a) = b$, and define $g(b) = a$.*

*The Axiom of Choice asserts that if $\{S_i\}_{i \in I}$ is a collection of non-empty sets indexed by a set I , it is possible to choose, for each i in I , an element x_i of S_i .

It is straightforward to check that $fg = I_B$, the identity map on B .^{*} Any particular g defined this way is called a *branch* of f^{-1} .

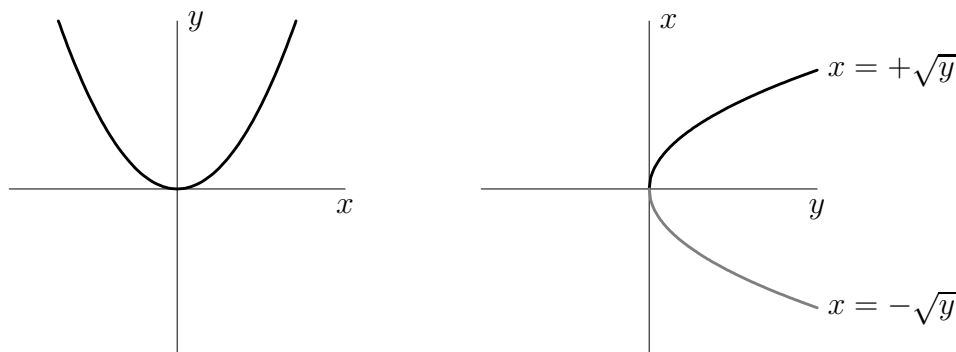


Figure 4.6: Right inverses of $f(x) = x^2$.

Example 4.40. Define $f : \mathbf{R} \rightarrow [0, \infty)$ by $f(x) = x^2$, see Figure 4.6, left. For each $y > 0$, there exist two real x such that $f(x) = y$, namely $x = \pm\sqrt{y}$. In particular, there are two “obvious” branches of f^{-1} , defined by $g_{\pm}(y) = \pm\sqrt{y}$ for $y \geq 0$, see Figure 4.6, right. (There are infinitely many other choices, though all are discontinuous.) For any such choice, $(fg)(y) = f(g(y)) = y$ for all $y \geq 0$. What about $g(f(x))$?

4.4 Equivalence Relations

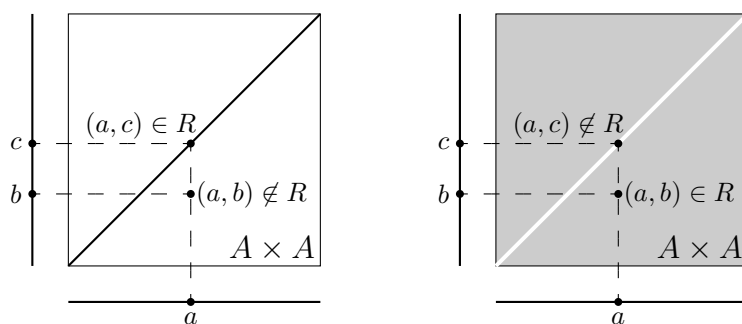
Definition 4.41. Let A be a non-empty set. A *relation* on A is a subset $R \subseteq A \times A$. Elements a and b of A are *R-related*, written aRb , if $(a, b) \in R$.

Example 4.42. The *equality* relation $=$ on A is defined by the *diagonal* $R = \Delta = \{(a, a) : a \in A\}$.

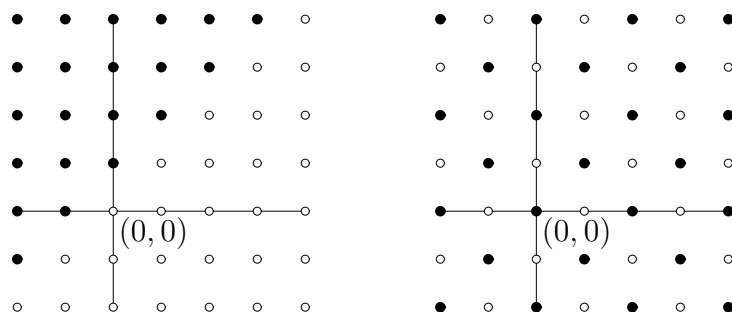
Example 4.43. The *inequality* relation \neq is the complement of the equality relation, $A \times A \setminus \Delta = \{(a_1, a_2) : a_1 \neq a_2\}$.

Remark 4.44. Generally, if $R_2 = A \times A \setminus R_1$, the relations R_1 and R_2 are logical opposites: One relation holds for a pair of elements if and only if the other fails for the same pair.

^{*}In general, $gf \neq I_A$, the identity map on A , since if $f(a_1) = b = f(a_2)$ for $a_1 \neq a_2$, we cannot have both $g(b) = a_1$ and $g(b) = a_2$.

Figure 4.7: Equality and inequality: $a \neq b$ and $a = c$.

Example 4.45. Let $A = \mathbf{Z}$ be the set of integers. The *less-than* relation is the set $R = \{(n_1, n_2) : n_1 < n_2\}$.

Figure 4.8: Less-than and parity on \mathbf{Z} .

Example 4.46. Again, let $A = \mathbf{Z}$. The *parity* relation on \mathbf{Z} is the set $R = \{(n_1, n_2) : n_2 - n_1 \text{ is even}\}$. Two integers are related if and only if they are both even or both odd.

Example 4.47. Let $f : A \rightarrow A$ be a mapping. Viewing f as a subset of $A \times A$ defines the *maps-to-under- f* relation on A : aRb if and only if $f(a) = b$, if and only if a maps to b under f .

Definition 4.48. Let R be a relation on a set A . We say R is

- *reflexive* if aRa for all a in A ;
- *symmetric* if, for all a and b in A , aRb implies bRa ;
- *transitive* if, for all a , b , and c in A , aRb and bRc imply aRc .

Example 4.49. Though not a formal example, the “friendship” relation may help you assimilate the conditions in the preceding definition. Let A be some set of people, and let aRb mean “ b is a friend of a ”.

R is reflexive if and only if every person is their own friend; R is symmetric if and only if all friendships are mutual; R is transitive if and only if every friend-of-a-friend is a friend.

Definition 4.50. A reflexive, symmetric, and transitive relation is an *equivalence relation*.

If R is an equivalence relation on A and $a \in A$, the *equivalence class* of a is the set

$$[a] = \{x \text{ in } A : aRx\} \subseteq A$$

comprising all elements related to a .

Example 4.51. Equality is an equivalence relation on an arbitrary set: For all a , b , and c , we have $a = a$ (reflexivity), $a = b$ implies $b = a$ (symmetry), and if $a = b$ and $b = c$, then $a = c$ (transitivity).

Inequality is symmetric, but neither reflexive nor transitive.

Less-than is transitive (if $a < b$ and $b < c$, then $a < c$), but neither reflexive nor symmetric.

The parity relation is an equivalence relation: For all integers a , b , and c , $a - a$ is even (reflexivity), if $b - a$ is even (aRb) then $a - b$ is even (bRa), and if $b - a$ and $c - b$ are even (aRb and bRc), then $c - a = (c - b) + (b - a)$ is even (aRc).

Equivalence Relations and Partitions

Let A be a non-empty set. Recall that a *partition* of A is a collection of non-empty, disjoint subsets whose union is A . Partitions and equivalence relations are two ways of viewing a single mathematical structure: Every equivalence relation gives rise to a partition, every partition gives rise to an equivalence relation, and these associations are inverse to each other.

Proposition 4.52. *Let R be an equivalence relation on A . The equivalence classes of R partition A .*

Proof. Since $a \in [a]$ for each a , every element of A lies in at least one equivalence class. It remains to prove that two arbitrary equivalence classes $[a]$ and $[b]$ are either disjoint or identical. To prove this it suffices

to show that if $[a] \cap [b] \neq \emptyset$ (i.e., the classes are not disjoint), then $[a] = [b]$.

Let's first run through the argument using the friendship metaphor. If a and b have a friend in common, then a and b are themselves friends (transitivity). Consequently, every friend of a is a friend of b (transitivity again) and *vice versa*, so a and b have exactly the same set of friends.

Formally, if $[a] \cap [b] \neq \emptyset$, there exists a c in A such that $c \in [a] \cap [b]$. Consequently, aRc and bRc . By symmetry of R , aRc and cRb , and by transitivity aRb . This means $a \in [b]$ and $b \in [a]$.

It is now easy to prove $[a] \subseteq [b]$ and $[b] \subseteq [a]$: If $x \in [a]$, then xRa , and since aRb , transitivity guarantees xRb , meaning $x \in [b]$. Reversing the roles of a and b completes the argument.

We have shown that non-disjoint equivalence classes are identical, so the set of equivalence classes of R is indeed a partition of A . \square

Remark 4.53. Conversely, if A is partitioned into subsets $\{A_i\}_{i \in I}$, there is an induced equivalence relation defined by aRb if and only if there exists an index i such that $a \in A_i$ and $b \in A_i$. Informally, aRb if and only if both elements lie in the same subset of the partition. Be sure to convince yourself that R is an equivalence relation, and that the partition induced by R is the original partition.

Example 4.54. The equivalence classes of the equality relation are the singletons, sets having one element: $[a] = \{a\}$ for each a in A .

Example 4.55. The parity relation on \mathbf{Z} has two equivalence classes: $[0] = 2\mathbf{Z}$ and $[1] = 2\mathbf{Z} + 1$.

Partitions and Prejudice

Our minds organize the external world by categorizing, unconsciously identifying people, objects, or phenomena that share some attribute.

Example 4.56. A physicist, a statistician, and a mathematician saw a flock of 100 sheep, of which one was black. The physicist said, "We can deduce that one in 100 sheep is black." The statistician said, "No, only that *in this sample of 100 sheep*, one is black." The mathematician corrected, "No, we can only deduce that one sheep in this sample is black *on one side*."

Often we cope fluently with such hierarchies: a particular mandarin orange, mandarin oranges, oranges, citrus fruit, fruit. . . . At other times, prejudice deceives us into identifying individuals according to superficial characteristics (such as gender, ethnicity, religion, or scientific field) and incorrectly presuming “all such people are alike”.

In mathematics, we can sometimes turn prejudice to good use. Perhaps we don’t care which integer we’re dealing with, but only if it’s even or odd, or if it leaves a remainder of 5 on division by 12. Maybe we’re dealing with pairs of points in the plane, but don’t care where they’re located, only that the second is located one unit to the right of the first. In such cases, an equivalence relation allows us to formalize our prejudice and discard irrelevant information.

Let A be a set, R an equivalence relation on A , and $\{A_i\}_{i \in I}$ the partition of A into equivalence classes. Each “index” i is associated with the non-empty set $A_i \subseteq A$, and the index set I is in bijective correspondence with the set of equivalence classes. We call the set of equivalence classes the *quotient* of A by R , denoted $I = A/R$ and read “ A modulo R ” (or “ $A \bmod R$ ” for short). *Elements* of A/R are *collections* of objects in A . The equivalence relation R is “unable to distinguish” elements of A_i , so when R “looks at” A it “sees” $I = A/R$.

Example 4.57. Two real numbers θ_1 and θ_2 determine the same longitude on the earth if and only if their difference is a multiple of one full turn, say 360° . To formalize this in the language of quotients, let $A = \mathbf{R}$ be the set of real numbers (a.k.a. the number line), and define the relation R by $\theta_1 R \theta_2$ if and only if $\theta_2 - \theta_1$ is an integer multiple of 360. By an argument entirely similar to that given for the parity relation in Example 4.51, R is an equivalence relation.

The set of equivalence classes is indexed by the half-open interval $[0, 360)$, since every angle is equivalent mod R to a unique number between 0 and 360 (excluding 360, which is equivalent to 0). We call this set the “space of angles”.

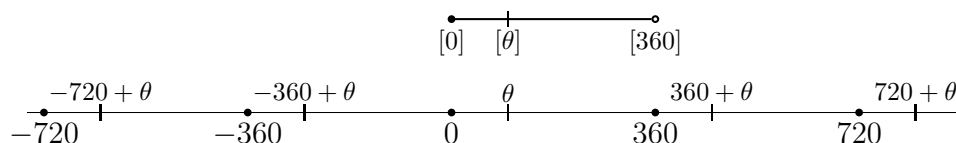


Figure 4.9: The number line, and the space of angles.

Mappings and Equivalence Classes

Let A be non-empty, R an equivalence relation on A , and $f : A \rightarrow B$ a mapping. We will often be interested in trying to define an “induced” map \bar{f} from the quotient set $\bar{A} = A/R$ to B .

Think of the elements of an equivalence class $[a]$ as a clique of friends who are polled by f , the question being “Which element of B do you map to?” If the clique responds unanimously (“We all map to b ”), then by fiat \bar{f} maps $[a]$ in \bar{A} to b in B . If *every* clique reaches a unanimous decision, there is a mapping $\bar{f} : A/R \rightarrow B$ defined by $\bar{f}([a]) = f(a)$.

If the responses are mixed for some clique $[a]$, then \bar{f} is undefined; a mapping must be single-valued for every input, but the members of $[a]$ do not decide unanimously where to be mapped by f .

Definition 4.58. Let $f : A \rightarrow B$ be a mapping, and R an equivalence relation on A . We say f is *well-defined* modulo R , or f is *constant on equivalence classes* of R , if aRa' implies $f(a) = f(a')$. If f is well-defined modulo R , we define the *induced mapping* $\bar{f} : A/R \rightarrow B$ by $\bar{f}([a]) = f(a)$ for each a in A .

Remark 4.59. If R is an equivalence relation on A , there is a “quotient map” $\Pi : A \rightarrow A/R$ defined by $\Pi(a) = [a]$. If $f : A \rightarrow B$ is well-defined modulo R and $\bar{f} : A/R \rightarrow B$ denotes the induced mapping, then $f = \bar{f} \circ \Pi$. We say “ f factors through A/R ”.

Example 4.60. Let $A = \mathbf{Z}$ be the set of integers, R the parity relation, and $f : \mathbf{Z} \rightarrow \{1, -1\}$ the mapping defined by $f(a) = (-1)^a$. Under f , every even integer maps to 1 and every odd integer maps to -1 , so f is well-defined modulo parity. Intuitively, to compute $(-1)^a$ for some integer a , we only need to know whether a is even or odd.

The quotient space $A/R = \{[0], [1]\} = \{2\mathbf{Z}, 2\mathbf{Z} + 1\}$ is a set having two elements, and the induced map $\bar{f} : A/R = \{2\mathbf{Z}, 2\mathbf{Z} + 1\} \rightarrow \{1, -1\}$, defined by

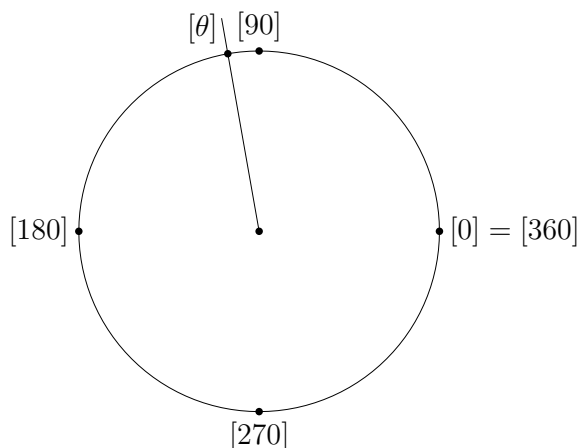
$$\bar{f}([0]) = (-1)^0 = 1, \quad \bar{f}([1]) = (-1)^1 = -1,$$

is bijective.

Example 4.61. Let $A = \mathbf{Z}$, R the parity relation, and $f : \mathbf{Z} \rightarrow \mathbf{Z}$ defined by $f(a) = a^2$. The integers 0 and 2 are elements of $[0]$, but $f(0) = 0 \neq 4 = f(2)$. Thus f is not well-defined modulo parity.

This should be no surprise: To compute the square of an integer a , it is not enough to know whether a is even or odd.

Example 4.62. Let $A = \mathbf{R}$ be the set of real numbers, R the “longitude” relation, and define $f : \mathbf{R} \rightarrow \mathbf{R}^2$ by $f(t) = (\cos t, \sin t)$, the standard trigonometric parametrization of the circle, with trig functions in “degrees mode”.



If $\theta_2 - \theta_1$ is an integer multiple of 360, then $\cos \theta_1 = \cos \theta_2$ and $\sin \theta_1 = \sin \theta_2$, so $f(\theta_1) = f(\theta_2)$. Consequently, there is an induced mapping from the space of angles to the unit circle in the plane. In words, f factors through locations on the earth.

Since $\cos \theta_1 = \cos \theta_2$ and $\sin \theta_1 = \sin \theta_2$ if and only if $\theta_2 - \theta_1$ is an integer multiple of 360, the mapping \bar{f} is bijective, so *the space of angles may be regarded as the unit circle*. Geometrically, each equivalence class $[\theta]$ corresponds to a unique point of the unit circle.

Exercises

Exercise 4.1. Let $f : A \rightarrow B$ be a mapping, and let U and V be subsets of A . Prove the following:

- (a) If $U \subseteq V$, then $f(U) \subseteq f(V)$.
- (b) $f(A \setminus U) = f(A) \setminus f(U)$.
- (c) If f is injective and $f(U) \subseteq f(V)$, then $U \subseteq V$.

Exercise 4.2. Let $f : A \rightarrow B$ and $g : B \rightarrow C$ be mappings, and assume $S \subseteq A$, $T \subseteq C$.

- (a) Prove $g(f(S)) = (gf)(S)$.

(b) Prove $f^{-1}(g^{-1}(T)) = (gf)^{-1}(T)$.

Exercise 4.3. Let $f : A \rightarrow B$ be a mapping.

(a) Assume $T \subseteq B$ is arbitrary. Prove $f(f^{-1}(T)) \subseteq T$, and that equality holds if f is surjective. Give an example of a mapping f and a set T for which the inclusion is proper.

(b) Assume $S \subseteq A$ is arbitrary. Prove $S \subseteq f^{-1}(f(S))$, and that equality holds if f is injective. Give an example of a mapping f and a set S for which the inclusion is proper.

Exercise 4.4. (a) Let $f : A \rightarrow B$ and $g : B \rightarrow C$ be injective. Prove gf is injective. (This is the second assertion of Proposition 4.32.)

(b) Suppose gf is injective. Prove f is injective.
Suggestion: Prove the contrapositive.

Exercise 4.5. Let m and b be integers, and define a mapping $f : \mathbf{Z} \rightarrow \mathbf{Z}$ by $f(x) = mx + b$.

(a) Prove f is injective if and only if $m \neq 0$.

(b) Find necessary and sufficient conditions on m and b for f to be surjective. If f is bijective, find a formula for the inverse mapping.

Exercise 4.6. Let $f : A \rightarrow B$ be a mapping. If $S \subseteq A$, define the *restriction* of f to S to be the mapping $f|_S : S \rightarrow B$ defined by $f|_S(a) = f(a)$ for all a in S .

(a) Prove that f is injective if and only if $f|_S$ is injective for *every* subset S of A .

(b) Assume f is a bijection. Prove that if S is a non-empty subset of A , then the restriction $f|_S$ is a bijection from S to $f(S)$ and the restriction $f|_{A \setminus S}$ is a bijection from $A \setminus S$ to $B \setminus f(S)$.

Exercise 4.7. Let m , n , and q be positive integers.

(a) Let A be a set containing m elements, B a set containing n elements, and assume $m > nq$. Prove that if $f : A \rightarrow B$ is a mapping, then there exists a b in B such that $f^{-1}(\{b\})$ contains at least $q + 1$ elements. (This result is known as the *Pigeonhole Principle*. If you distribute $m > nq$ pigeons among n holes, then some hole contains more than q pigeons.)

Suggestion: Write B as a union of singleton sets, and use Proposition 4.14. The contrapositive may be more natural to prove.

- (b) With the same notation, let $f : A \rightarrow B$ be a mapping. Prove that if f is injective, then $m \leq n$, and that if f is surjective, then $m \geq n$. Show by example that both converse statements are false.
- (c) With the same notation, assume $m = n$, and let $f : A \rightarrow B$ be a mapping. Prove f is injective if and only if f is surjective. (Suggestion: Use part (b) to prove that f is injective if and only if $f(A)$ contains m elements, if and only if f is surjective.)

Exercise 4.8. Define $f : \mathbf{C} \rightarrow \mathbf{C}$ by $f(z) = z^2$.

- (a) By writing $z = x + iy$ with x and y real, calculate the real and imaginary parts of $f(z)$.
- (b) By writing $z = re^{i\theta}$ with $r \geq 0$ and θ real, re-calculate $f(z)$, and use your result to describe the geometric action of the mapping f .
- (c) Find the preimages of the singletons $\{1\}$, $\{-1\}$, $\{i\}$, and $\{\rho e^{i\phi}\}$.

Exercise 4.9. Repeat the preceding question for $f : \mathbf{C} \rightarrow \mathbf{C}$ defined by $f(z) = z^3$. In part (c), do you notice any “geometric pattern”?

Exercise 4.10. Let $n > 1$ be an integer, and define $f : \mathbf{C} \rightarrow \mathbf{C}$ by $f(z) = z^n$. By writing $z = re^{i\theta}$, describe the geometric action of f , and find the preimage of $\{\rho e^{i\phi}\}$. If $\rho > 0$, how many points are in the preimage, and how are these points situated geometrically in \mathbf{C} ?

Exercise 4.11. Let $g : \mathbf{R} \rightarrow \mathbf{R}$ be a real-valued function of one real variable. We say g is *even* if $g(-x) = g(x)$ for all x in \mathbf{R} , and that g is *odd* if $g(-x) = -g(x)$ for all x in \mathbf{R} . (Analogous formulas define the notions of “even” and “odd” functions whose domain and/or target is \mathbf{Z} or any other set in which negatives are defined.)

- (a) Find all functions that are *both* even and odd.
- (b) Let $f : \mathbf{R} \rightarrow \mathbf{R}$ be an arbitrary function. Show the functions

$$f_{\text{even}}(x) = \frac{1}{2}[f(x) + f(-x)], \quad f_{\text{odd}}(x) = \frac{1}{2}[f(x) - f(-x)]$$

are even and odd, respectively.

- (c) Suppose there exist an even function E and an odd function O such that $f(x) = E(x) + O(x)$ for all real x . Find formulas for E and O . Hint: Compute $f(-x)$.

- (d) Prove every function $f : \mathbf{R} \rightarrow \mathbf{R}$ can be written *uniquely* as the sum of an even function and an odd function. These functions are called the *even part* and *odd part* of f .
- (e) Find the even and odd parts of $f(x) = x^3 - 2x^2 + x + 1$, $g(x) = e^x$, and $h(x) = \cos x$.

Exercise 4.12. The *hyperbolic functions* \cosh and \sinh are defined by

$$\cosh x = \frac{1}{2}(e^x + e^{-x}), \quad \sinh x = \frac{1}{2}(e^x - e^{-x}), \quad x \text{ real.}$$

- (a) Show that $\cosh^2 - \sinh^2 = 1$. Carefully sketch the graphs of \cosh and \sinh on a single set of axes. Suggestion: First calculate $\cosh \pm \sinh$.
- (b) Show that for all real x ,
- $$\cosh(2x) = \cosh^2 x + \sinh^2 x, \quad \sinh(2x) = 2 \cosh x \sinh x.$$
- (c) Show that $\cosh' = \sinh$ and $\sinh' = \cosh$.
- (d) The *hyperbolic tangent* and *hyperbolic secant* functions are

$$\tanh = \frac{\sinh}{\cosh}, \quad \operatorname{sech} = \frac{1}{\cosh^2}.$$

Carefully sketch their graphs on a single set of axes, show that $\tanh^2 = 1 - \operatorname{sech}^2$, and find formulas for \tanh' and sech' .

- (e) Find an algebraic formula for the inverse function \tanh^{-1} . Hint: Solve $y = \tanh x$ for x by cross-multiplying and rearranging.
- (f) Find algebraic formulas for \sinh^{-1} , and for two branches of \cosh^{-1} . Use algebra to show the branches of \cosh^{-1} differ by a sign. Hint: Solve (e.g.) $y = \sinh x$ for x by multiplying through by e^x and rearranging to get a quadratic in e^x ; then use the quadratic formula.

Exercise 4.13. Let x and y be arbitrary real numbers. Show that

$$\begin{aligned} \cosh(x+y) &= \cosh x \cosh y + \sinh x \sinh y, \\ \sinh(x+y) &= \sinh x \cosh y + \cosh x \sinh y, \\ \tanh(x+y) &= \frac{\tanh x + \tanh y}{1 + \tanh x \tanh y}. \end{aligned}$$

Exercise 4.14. Let ϕ be a real number, and recall Euler's formula

$$e^{i\phi} = \cos \phi + i \sin \phi.$$

- (a) Express $e^{-i\phi}$ in terms of $\cos \phi$ and $\sin \phi$.
- (b) Show that

$$\cos \phi = \frac{e^{i\phi} + e^{-i\phi}}{2}, \quad \sin \phi = \frac{e^{i\phi} - e^{-i\phi}}{2i}.$$

- (c) Show that for all real ϕ ,

$$\cosh(i\phi) = \cos \phi, \quad \sinh(i\phi) = i \sin \phi.$$

(The hyperbolic functions are defined in Exercise 4.12.)

Exercise 4.15. Let A be a non-empty set, and let $R = \emptyset \subseteq A \times A$. Prove R is symmetric and transitive, but not reflexive.

Exercise 4.16. Define a relation R on \mathbf{Z} by aRb if and only if $|a| = |b|$.

- (a) Prove R is an equivalence relation.
- (b) Let $f : \mathbf{Z} \rightarrow \mathbf{Z}$ be defined by $f(a) = a^2$. Is f well-defined mod R ?
- (c) Let $g : \mathbf{Z} \rightarrow \mathbf{Z}$ be defined by $g(a) = 3a$. Is g well-defined mod R ?
- (d) Prove $f : \mathbf{Z} \rightarrow \mathbf{Z}$ is well-defined mod R if and only if f is an even function, see Exercise 4.11.

Exercise 4.17. Let $A = \mathbf{Z}$, and define a relation R by aRb if and only if $b - a$ is an integer multiple of 4.

- (a) Prove R is an equivalence relation, and find the equivalence classes of R , and describe the quotient \mathbf{Z}/R .
- (b) Let $f : \mathbf{Z} \rightarrow \mathbf{C}$ be defined by $f(a) = (-1)^a$. Prove f is well-defined mod R . Is \bar{f} injective?
- (c) Let $g : \mathbf{Z} \rightarrow \mathbf{C}$ be defined by $g(a) = i^a$. Is g well-defined mod R ? If so, is \bar{g} injective?

Exercise 4.18. Let $f : A \rightarrow A$ be a mapping, and suppose the “maps-to” relation, aRb if and only if $b = f(a)$, is an equivalence relation. What can you say about f ?

Exercise 4.19. Let R be an equivalence relation on A . If $f : A \rightarrow B$ is a mapping such that $a_1 R a_2$ if and only if $f(a_1) = f(a_2)$, i.e., whose level sets are precisely the equivalence classes of R , prove that the induced mapping $\bar{f} : A/R \rightarrow B$ is injective.

Exercise 4.20. Let $f : A \rightarrow B$ be a mapping, and define a relation on A by $a_1 R a_2$ if and only if $f(a_1) = f(a_2)$.

- (a) Prove R is an equivalence relation, and the equivalence classes of R are preimages of singletons, namely *level sets* of f : $f^{-1}(\{b\})$ for some b in B .
- (b) Let $f : \mathbf{R} \rightarrow \mathbf{R}$ be defined by $f(x) = x^2$. Describe the equivalence classes of f .
- (c) Let $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ be defined by $f(x, y) = x^2 + y^2$. Describe the equivalence classes of f .

Exercise 4.21. Let $f : A \rightarrow B$ be a mapping. A mapping $g : A \rightarrow B$ is said to be *constant on the level sets of f* if $f(a_1) = f(a_2)$ implies $g(a_1) = g(a_2)$. (Compare the two preceding questions.)

- (a) Define $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ by $f(x, y) = x^2 + y^2$. Which of the following are constant on the level sets of f ?

$$g_1(x, y) = (1 - \sqrt{x^2 + y^2})^2, \quad g_2(x, y) = x^2 - y^2, \quad g_3(x, y) = 1.$$

- (b) For a general mapping $f : A \rightarrow B$, prove the following are equivalent:
 - (i) g is constant on the level sets of f .
 - (ii) There exists a mapping $\phi : B \rightarrow B$ such that $g = \phi \circ f$. (In this situation we say “ g is a function of f ”.)

Exercise 4.22. Imagine a world where the natural numbers are known, but the integers are not. For example, $2 = 1 + x$ “has a solution”, but $1 = 2 + x$ “has no solution”. For a time, mathematicians cope by inventing symbols, such as “ -1 ”, to denote “negative” numbers. These fictitious entities turn out to be so useful that logical care demands they be placed on a firm logical foundation.

This exercise outlines an implementation, taking its cue from the equation $n_1 = n_2 + x$. The goal is to construct—using only natural

numbers and operations of set theory—a larger collection of “numbers” that contains a copy of \mathbf{N} and in which the equation $m = n + x$ has a solution when m and n are numbers *of the more general type*.

Intuitively, an integer will be an *ordered pair of natural numbers*. The pair $x = (n_1, n_2)$ corresponds to the solution of $n_1 = n_2 + x$. For example, the pair $(1, 2)$ corresponds to the solution of $1 = 2 + x$, namely to the integer $x = -1$.

Many different pairs represent the same number; $(4, 1)$, $(9, 6)$, and $(1968, 1965)$ all correspond to 3. Two pairs (n_1, n_2) and (m_1, m_2) represent the same number exactly when $n_1 - n_2 = m_1 - m_2$, that is, when $n_1 + m_2 = n_2 + m_1$. We are therefore led to define the relation

$$(4.2) \quad (m_1, m_2) \sim (n_1, n_2) \quad \text{if and only if } n_1 + m_2 = n_2 + m_1.$$

(a) Prove (4.2) defines an equivalence relation on the set $X = \mathbf{N} \times \mathbf{N}$.

(An *integer* is an equivalence class of $\mathbf{N} \times \mathbf{N}$ with respect to (4.2).)

(b) Define the *sum* of two integers by adding representatives:

$$(m_1, m_2) \oplus (n_1, n_2) = (m_1 + n_1, m_2 + n_2).$$

Show that “addition” \oplus is well-defined modulo (4.2). Explicitly, if $(m_1, m_2) \sim (m'_1, m'_2)$ and $(n_1, n_2) \sim (n'_1, n'_2)$, then

$$(m_1, m_2) \oplus (n_1, n_2) \sim (m'_1, m'_2) \oplus (n'_1, n'_2).$$

(c) Motivated by the idea that (m_1, m_2) and (n_1, n_2) represent $m_1 - m_2$ and $n_1 - n_2$, define the *product* of two integers by

$$(m_1, m_2) \odot (n_1, n_2) = (m_1 n_1 + m_2 n_2, m_1 n_2 + n_1 m_2).$$

Show that “multiplication” \odot is well-defined modulo (4.2).

(d) Using commutativity and associativity of addition and multiplication *of natural numbers*, verify that \oplus and \odot are commutative and associative.

(e) Show that if m and n are natural numbers, then

$$(m, 0) \oplus (n, 0) = (m + n, 0),$$

$$(m, 0) \odot (n, 0) = (mn, 0).$$

That is, the equivalence class $[(n, 0)]$ corresponds to the natural number n , so we have succeeded in building a copy of \mathbf{N} inside \mathbf{Z} .

(f) Show that $m = n + x$, i.e., $(m_1, m_2) = (n_1, n_2) \oplus (x_1, x_2)$, has a solution for all integers m and n .

Chapter 5

Induction and Recursion

The set \mathbf{N} of natural numbers is a mathematician's prototype of an infinite list. There is a “first element” 0 and a notion of “successorship”, the unique successor of n in \mathbf{N} being denoted $n + 1$. Every natural number arises by starting with 0 and taking successors, and every natural number except 0 is the successor of a unique natural number. We denote natural numbers with their familiar Hindu-Arabic symbols, $\mathbf{N} = \{0, 1, 2, 3, 4, 5, \dots\}$.

This chapter introduces the technique of *mathematical induction* for proving infinitely many statements provided the statements are suitably structured in a list indexed by the natural numbers. Mathematical induction is particularly useful for investigating structures defined by *recursion*, such as numerical sequences whose first term is given and where subsequent terms are obtained by a definite rule from prior terms.

5.1 Mathematical Induction

Suppose someone tells you the sum of the first n positive odd integers is equal to n^2 . What basis do you have for believing this claim?

As a start, you might verify a few instances by hand. For example, $1 + 3 + 5 = 9 = 3^2$, so the claim is true when $n = 3$. Perhaps skeptical, you add the first ten odd integers, or the first twenty, each time verifying the claim. Perhaps you are starting to believe.

Logically, however, testing special cases leaves you no closer to complete certainty. Have you tried adding the first hundred thousand odd integers? The first billion? Finding a single counterexample would prove the claim false, but no matter how many cases you verify, there

remain infinitely many unverified cases.

The technique of *mathematical induction* allows us to resolve such questions with a finite proof. The idea is to break the statement “For every natural number n , the sum of the first n odd positive integers is equal to n^2 ” into an infinite list of statements. Here, we take

$$P(n) \qquad 1 + 3 + 5 + \cdots + (2n - 1) = n^2.$$

To say $P(100)$ is true, for example, means the sum of the first hundred odd integers is equal to $10,000 = 100^2$. (An “empty” sum is 0, so $P(0)$ reads $0 = 0$.)

The original statement may be rephrased “For every natural number n , $P(n)$ is true.” This single statement P encapsulates the infinite list of statements: $P(0)$ is true, $P(1)$ is true, $P(2)$ is true, etc.

In order to establish the truth of P , it suffices to prove $P(0)$ is true (the *base case*), and to prove $P(k)$ implies $P(k + 1)$ for every k in \mathbf{N} (the *inductive step*). The legal contract is as follows.

Theorem 5.1 (Principle of Mathematical Induction). *Let $P(n)$, with $n = 0, 1, 2, \dots$, be a family of statements. If $P(0)$ is true, and if $P(k)$ implies $P(k + 1)$ for each $k \geq 0$, then $P(n)$ is true for all $n \geq 0$.*

Remark 5.2. It is sometimes convenient to take an index $n_0 > 0$ for the base case. In this event, one must prove $P(n_0)$ is true, and establish that $P(k)$ implies $P(k + 1)$ for $k \geq n_0$.

To see intuitively why the base case and inductive step are enough, consider the consequences of “ $P(0)$ is true, and $P(k)$ implies $P(k + 1)$ for all $k \geq 0$ ”. Taking $k = 0$, the inductive step says $P(0)$ implies $P(1)$. But $P(0)$ is *true* by the base case, so $P(1)$ is also *true* by the inductive step. Now repeat the argument, taking $k = 1$. By the inductive step, $P(1)$ implies $P(2)$, but $P(1)$ is *true*, so $P(2)$ is also true. Continuing in this fashion, $P(3)$ is true, and $P(4)$, and so forth, *ad infinitum*. The chain of deduction may be represented as a sequence of arrows:

$$\begin{array}{c} \underbrace{P(0)}_{\text{Base case}} \implies P(1) \implies P(2) \implies P(3) \implies \dots \\ \implies P(k) \implies P(k + 1) \implies \dots \end{array}$$

Proof of the Principle of Mathematical Induction. The proof proceeds by contraposition. Suppose $P(n)$ is false for some natural number n . It

suffices to show that either the base case is false, or else the inductive step fails for some index.

Let $A = \{m \text{ in } \mathbf{N} : P(m) \text{ is false}\}$. Because $P(n)$ is false for some n , the set $A \subseteq \mathbf{N}$ is non-empty. By the well-ordering property, Axiom O4. for the integers, there is a smallest element a_0 in A . By definition of A , $P(a_0)$ is false, but $P(m)$ is true if $m < a_0$.

If $a_0 = 0$, then the base case is false.

If $a_0 > 0$, then $a_0 - 1 \in \mathbf{N}$, and $P(a_0 - 1)$ is true but $P(a_0)$ is false. That is, the inductive step is invalid for $k = a_0 - 1$.

Contrapositively, if the base case is true and the inductive step is valid for all k in \mathbf{N} , then $P(n)$ is true for all n in \mathbf{N} . \square

Example 5.3. Let's see how induction works in detail, with $P(n)$ as above. As noted earlier, the base case $P(0)$ reads $0 = 0$ because an empty sum is 0. (If this seems suspicious, note that the first odd positive integer is equal to 1^2 , so $P(1)$ is also true).

Next, assume inductively that $P(k)$ is true for some fixed (but arbitrary) natural number k . The sum of the first $(k + 1)$ odd positive integers is equal to the sum of the first k plus the $(k + 1)$ th. By hypothesis, the sum of the first k is equal to k^2 . We therefore deduce

$$\underbrace{1 + 3 + 5 + \cdots + (2k - 1)}_{=k^2 \text{ by } P(k)} + (2k + 1) = k^2 + (2k + 1) = (k + 1)^2$$

by algebra. This equation says the sum of the first $(k + 1)$ odd positive integers is equal to $(k + 1)^2$. By assuming $P(k)$, we proved $P(k + 1)$.

To summarize, the base case $P(0)$ is true, and the inductive step, $P(k)$ implies $P(k + 1)$, is valid for each natural number k . By the Principle of Mathematical Induction, $P(n)$ is true for all $n \geq 0$.

Remark 5.4. Our use of n or k to denote an arbitrary natural number in an inductive proof signifies a subtle but important distinction. In this book, $P(n)$ refers to the general statement of an inductive list, whose truth value is to be established. By contrast, $P(k)$ refers to a general statement that is “inductively true”: We assume “for the sake of argument” that $P(k)$ is true for some (particular but arbitrary) k , and try to deduce $P(k + 1)$.

To emphasize, in an inductive proof we never assume $P(k)$ is true without qualification. To do so (for arbitrary k) would be to assume the conclusion we wish to establish, namely that $P(k)$ is true for every $k \geq 0$.

Example 5.5. Consider the statement, “For all $n \geq 1$, $n^2 + n + 41$ is prime.” Checking cases may convince you this statement is true. Taking $n = 2, 5, 20$, and 100 respectively asserts that $2^2 + 2 + 41 = 47$, $5^2 + 5 + 41 = 71$, $20^2 + 20 + 41 = 461$, and $100^2 + 100 + 41 = 10141$ are primes, all true statements.

This example demonstrates the danger of relying merely on checking cases. Note that $n^2 + n + 41 = n(n + 1) + 41$. Can you use this fact to find two (or more) values of n for which $n^2 + n + 41$ is not prime?

Complete mastery of mathematical induction is essential. It is our fundamental technique for proving infinite families of statements when they can be listed in such a way that each statement implies the next.

Example 5.6. The *Tower of Hanoi* puzzle consists of seven disks of decreasing size, stacked on one of three spindles. The object is to move the entire stack to one of the other spindles, moving only one disk at a time, and never placing a larger disk atop a smaller one. The initial configuration is shown in Figure 5.1.

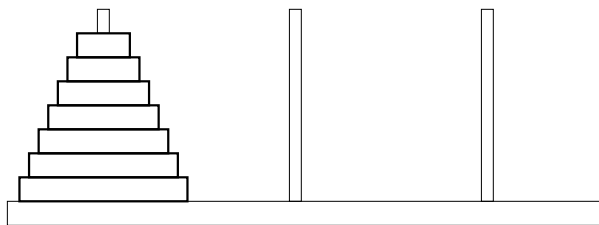


Figure 5.1: The Tower of Hanoi, initial configuration.

How many individual transfers are required to “solve” the Tower of Hanoi? To bring the power of mathematical induction to bear, we generalize the puzzle, allowing n disks rather than seven. Let $T(n)$ denote the number of individual transfers required to move a stack of n disks subject to the rules above.

Clearly $T(1) = 1$; a single transfer moves a “stack” of one disk. For two disks, a bit of thought shows the task can be done in three transfers and no fewer: $T(2) = 3$. It’s worthwhile to experiment with a stack of three or four coins of different sizes before reading further.

The puzzle with $(n + 1)$ disks can be solved as follows: Move the top n disks from spindle 1 to spindle 2 (taking $T(n)$ transfers), then move the bottom disk to spindle 3 (one transfer), and finally move the stack of n disks from spindle 2 to spindle 3 (another $T(n)$ transfers). This

strategy is clearly optimal, since the bottom disk cannot be transferred until the rest of the stack has been moved away. Tallying the number of transfers, we find $T(n+1) = 2T(n) + 1$.

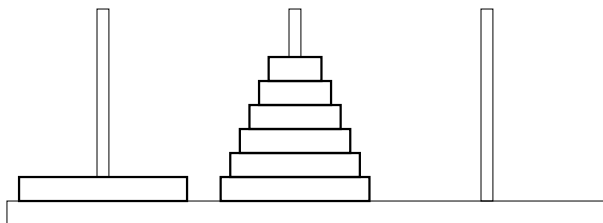


Figure 5.2: The Tower of Hanoi, intermediate configuration.

The original question could now be answered by successively calculating $T(3) = 2T(2) + 1 = 7$, $T(4) = 2T(3) + 1 = 15$, and so forth. Having formulated a general problem, however, we are led to ask “How many transfers are required to move a stack of n disks?” This is no longer a finite issue, since there are infinitely many puzzles, one for each positive integer n .

To proceed further, we must *guess a formula for $T(n)$* . The sequence 1, 3, 7, 15 (for towers of one, two, three, and four disks respectively) might lead us to suspect the number of transfers is one less than a power of two: $T(n) = 2^n - 1$. Call this equation $P(n)$. To see if this guess is correct, we will attempt to prove $P(n)$ is true for all n .

The statement $P(1)$ says $T(1) = 1 = 2^1 - 1$, which is true. This establishes the base case. Next, assume inductively that $P(k)$ is true for some k , namely $T(k) = 2^k - 1$. By the hierarchical strategy described earlier,

$$T(k+1) = 2T(k) + 1 = 2(2^k - 1) + 1 = 2 \cdot 2^k - 2 + 1 = 2^{k+1} - 1.$$

Thus, $P(k)$ implies $P(k+1)$ for each natural number k .

Since the base case is true and the inductive step is valid for each $k \geq 1$, our guess at a formula for $T(k)$ was correct by the Principle of Mathematical Induction. As a special case, we find that $2^7 - 1 = 127$ transfers are required to move a stack of seven disks.

Remark 5.7. Both luck and skill are involved in solving this type of problem. Looking at the “data” 1, 3, 7, we might have found other plausible formulas, such as $(n-1)^2 + (n-1) + 1 = n^2 - n + 1$.

The inductive step weeds out incorrect guesses such as this one. If $T(k) = k^2 - k + 1$ for some k , then

$$\begin{aligned} T(k+1) &= (k+1)^2 - (k+1) + 1 = k^2 + k + 1, \\ 2T(k) + 1 &= 2k^2 - 2k + 3, \end{aligned}$$

and these are generally different.

Recursive Definition

Definition 5.8. Let A be a non-empty set. A mapping $s : \mathbf{N} \rightarrow A$ is called an *infinite sequence* in A . The value $s(n) = s_n$ is called the *n th term*.

Remark 5.9. Intuitively, an infinite sequence is an *ordered list* $(s_n)_{n=0}^{\infty}$ of elements of A . Take care to distinguish a sequence and its image $\{s_n\}_{n=0}^{\infty}$, the unordered set of terms. The integer sequence defined by $s(n) = (-1)^n$ takes only two values, 1 and -1 , but has infinitely many terms, one for each natural number.

In practice, it may be convenient to take an integer n_0 as lowest index, in which case we write $(s_n)_{n=n_0}^{\infty}$ rather than $(s_{m+n_0})_{m=0}^{\infty}$.

A sequence may be defined by a formula, but many interesting sequences are specified by an initial value s_0 and a *recursion formula*, a rule for finding s_{n+1} if the terms $s_0, s_1, s_2, \dots, s_n$ are known.

Example 5.10. Let $(a_k)_{k=0}^{\infty}$ be a sequence of complex numbers. The associated sequence $(s_n)_{n=0}^{\infty}$ of *partial sums* is defined recursively by

$$s_0 = a_0, \quad s_{n+1} = s_n + a_{n+1} \quad \text{for } n \geq 0.$$

For example, successively expanding this definition gives

$$\begin{aligned} s_3 &= s_2 + a_3 \\ &= (s_1 + a_2) + a_3 \\ &= (s_0 + a_1) + a_2 + a_3 \\ &= a_0 + a_1 + a_2 + a_3, \end{aligned}$$

the sum of the terms a_k for $0 \leq k \leq 3$. This concept arises frequently enough to get special notation:

$$s_n = \sum_{k=0}^n a_k = a_0 + a_1 + a_2 + \cdots + a_n.$$

The recursion relation is written

$$(5.1) \quad \sum_{k=0}^{n+1} a_k = \left(\sum_{k=0}^n a_k \right) + a_{n+1},$$

which says the sum over $0 \leq k \leq n+1$ is obtained by adding the $(n+1)$ th term to the sum over $0 \leq k \leq n$.

Example 5.11. The sequence of *Fibonacci numbers* is defined recursively by

$$f_1 = f_2 = 1, \quad f_{k+2} = f_{k+1} + f_k \quad \text{for } k \geq 1.$$

The sum of two consecutive Fibonacci numbers gives the next. Thus

$$\begin{aligned} f_3 &= f_2 + f_1 = 1 + 1 = 2, \\ f_4 &= f_3 + f_2 = 2 + 1 = 3, \\ f_5 &= f_4 + f_3 = 3 + 2 = 5, \\ f_6 &= f_5 + f_4 = 5 + 3 = 8, \end{aligned}$$

and so on. The next several terms are 13, 21, 34, 55, 89, 144, ...

The Fibonacci sequence is generated by a *two-term recursion*, involving two consecutive terms of the sequence.

Our next example merits a formal definition.

Definition 5.12. The *factorial* $n!$ of a non-negative integer n is defined recursively by

$$0! = 1, \quad (n+1)! = (n+1) \cdot n! \quad \text{for } n \geq 0.$$

Remark 5.13. The convention $0! = 1$ is justified by Proposition 5.30 below, which counts the orderings of a set of n elements.

Example 5.14. Expanding the recursive definition for $n = 5$ gives $5! = 5 \cdot 4! = 5 \cdot 4 \cdot 3! = 5 \cdot 4 \cdot 3 \cdot 2! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 0!$, and this may be evaluated using the initial condition $0! = 1$. Thus $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

Generally, if n is a *positive* integer, then $n!$ is the product of the integers from 1 to n . The first ten or so factorials are worth memorizing, and their approximate sizes worth remembering:

n	$n!$	n	$n!$	n	$n!$	n	$n!$
1	1	5	120	9	362 880	13	6 227 020 800
2	2	6	720	10	3 628 800	14	87 178 291 200
3	6	7	5 040	11	39 916 800	15	1 307 674 368 000
4	24	8	40 320	12	479 001 600	16	20 922 789 888 000

There is a close kinship between mathematical induction and recursive definition. Both involve a base case, and information at one “level” expressed in terms of the same information “at lower levels”. Mathematical induction is usually the most natural way to study properties of recursively-defined sequences. For example, the recursion relation

$$T_1 = 1, \quad T_{k+1} = 2T_k + 1, \quad \text{for } k \geq 1$$

uniquely specifies the *closed formula* $T_k = 2^k - 1$, see Example 5.6 above. This sequence gives the number of moves required to solve the Tower of Hanoi with k disks.

Example 5.15. Songs can be recursive. The well-known drinking song “ n bottles of beer on the wall” has nested verses leading into each other, with the number of bottles decreasing by one at each verse. For each $n > 0$, the song ends after a finite amount of time: The recursion is *properly terminated*.

Samuel Beckett’s play *Waiting for Godot* contains an improperly-terminated recursion:

“A dog went into the kitchen and stole a crust of bread,
 So cook up with a ladle and beat him ’til he was dead.
 Then all the dogs came running and dug the dog a tomb,
 And carved upon the tombstone for eyes of dogs to come:
 “A dog went into the kitchen...
 (Repeat *ad infinitum*.)

Each successive verse gets nested within quotes one layer more deeply.

Example 5.16. The recursion rule $f(0) = 0$, $f(n) = nf(n - 1)$ for $n \leq 0$, is improperly terminated, *cf.* Definition 5.12. Attempting to evaluate $f(-1)$ gives an infinite regression:

$$f(-1) = (-1)f(-2) = (-1)(-2)f(-3) = \dots$$

Example 5.17. Computer hackers are fond of mischievous recursion.* The operating system “GNU”, often paired with the Linux kernel to form GNU/Linux, is a Unix-like multi-tasking environment whose name self-referentially stands for “GNU’s Not Unix”.

*Since before the days of personal computers, the term *hacker* has referred to technological problem-solvers and playful, creative programming enthusiasts. The term picked up associations with computer criminals only decades later. “Old-time” hackers use the term *cracker* for criminals who gain unauthorized access to computer systems.

5.2 Applications

This section introduces two “substantial” applications of recursive definitions and mathematical induction: integer exponentiation and the laws of exponents, and formulas for counting mappings between two finite sets.

Exponentiation

In school, you learned that x^n stands for “ n factors of x multiplied together”. Eventually this idea was extended to include exponents that are negative integers, zero, fractions, and general real numbers. This section develops the theory of exponentiation for integer exponents from first principles.

Remark 5.18. Defining exponentiation for rational exponents is essentially a matter of algebra, though establishing existence of numbers such as $2^{1/2}$ requires “analysis”, the branch of mathematics dealing with limits. Even *defining* exponentiation with irrational exponents requires analysis. Non-integer exponents lie beyond the scope of this book.

Definition 5.19. Let α be a complex number, and define

$$\alpha^0 = 1, \quad \alpha^{k+1} = \alpha \cdot \alpha^k \quad \text{for } k \geq 0.$$

Remark 5.20. The expression “ α^k ” is read “ α to the k ” or “the k th power of α ”. Intuitively, α^k is the result of multiplying k factors of α . (Convince yourself by expanding α^4 using the definition, in a similar manner to Example 5.14.)

The “empty product” convention $\alpha^0 = 1$ is justified by the “law of exponents”, $\alpha^{m+n} = \alpha^m \cdot \alpha^n$ for all integers m and n , see Theorem 5.23.

Remark 5.21. We have *defined* $0^0 = 1$. This convention makes formulas “work” in boundary cases. (Many mathematicians leave 0^0 undefined.)

Definition 5.22. If $\alpha = a + bi$ is a *non-zero* complex number and k is a positive integer, we define $\alpha^{-k} = (\alpha^{-1})^k$, the k th power of the *reciprocal*

$$\alpha^{-1} = \frac{1}{\alpha} = \frac{a - bi}{a^2 + b^2}.$$

Theorem 5.23 (The Law of Exponents). *If α, β are non-zero complex numbers,*

$$\left. \begin{array}{ll} \text{(i)} & (\alpha\beta)^n = (\alpha^n)(\beta^n), \\ \text{(ii)} & \alpha^{m+n} = \alpha^m \cdot \alpha^n, \\ \text{(iii)} & \alpha^{nm} = (\alpha^n)^m, \end{array} \right\} \quad \text{for all integers } m \text{ and } n.$$

In particular, $\alpha^{-n} = (\alpha^n)^{-1}$ for all $\alpha \neq 0$ and all integers n .

Remark 5.24. Conceptually, these results amount to counting the number of factors in a product. For instance, α^{m+n} represents a product of $(m+n)$ factors all equal to α ; such a product can be separated into a product of m factors and a product of n factors, i.e., into $\alpha^m \cdot \alpha^n$.

To give formal proofs, we use mathematical induction. Associativity and commutativity of complex multiplication are used freely. For best results, try to prove each part yourself from the inductive statement provided before reading the book's proof.

Proof. (i). For each natural number n , consider the statement

$$P(n) \qquad (\alpha\beta)^n = (\alpha^n)(\beta^n).$$

The base case $P(0)$ reduces to $1 = 1$, which is true. Assume inductively that $P(k)$ is true for some $k \geq 0$. We have

$$\begin{aligned} (\alpha\beta)^{k+1} &= (\alpha\beta)^k(\alpha\beta) && \text{Definition of exponentiation,} \\ &= (\alpha^k\beta^k)(\alpha\beta) && \text{Inductive hypothesis,} \\ &= \alpha^k(\beta^k \cdot \alpha)\beta && \text{Associativity,} \\ &= \alpha^k(\alpha \cdot \beta^k)\beta && \text{Commutativity,} \\ &= (\alpha^k \cdot \alpha)(\beta^k \cdot \beta) && \text{Associativity,} \\ &= (\alpha^{k+1})(\beta^{k+1}) && \text{Definition of exponentiation.} \end{aligned}$$

Since $P(0)$ is true and $P(k)$ implies $P(k+1)$ for all $k \geq 0$, the statement $P(n)$ is true for all $n \geq 0$ by the principle of mathematical induction.

If $n < 0$, replace α and β by their multiplicative inverses, and recall that $\alpha^{-n} = (\alpha^{-1})^n$ by definition. Further, taking $\beta = \alpha^{-1}$, we have

$$(\alpha^n)(\alpha^{-1})^n = 1^n = 1 = (\alpha^n)(\alpha^n)^{-1};$$

by cancellation, $\alpha^{-n} = (\alpha^{-1})^n = (\alpha^n)^{-1}$ for all n .

(ii). We first assume m and n are non-negative integers. For each n in \mathbf{N} , consider the statement

$$P(n) \quad \alpha^{m+n} = \alpha^m \cdot \alpha^n \quad \text{for all } m \text{ in } \mathbf{N}.$$

This *single statement* may be viewed as an infinite family of statements, one for each natural number m , with n a fixed natural number.

The base case $P(0)$ asserts $\alpha^{m+0} = \alpha^m \cdot \alpha^0$ for all m , which is true since $m + 0 = m$ for all m and $\alpha^0 = 1$. Next, assume inductively that $P(k)$ is true for some k , namely that

$$\alpha^{m+k} = \alpha^m \cdot \alpha^k \quad \text{for all } m \text{ in } \mathbf{N}.$$

For all m , we have

$$\begin{aligned} \alpha^{m+k+1} &= \alpha^{m+k} \cdot \alpha && \text{Definition of exponentiation} \\ &= (\alpha^m \cdot \alpha^k) \cdot \alpha && \text{Inductive hypothesis} \\ &= \alpha^m \cdot (\alpha^k \cdot \alpha) && \text{Associativity} \\ &= \alpha^m \cdot \alpha^{k+1} && \text{Definition of exponentiation} \end{aligned}$$

which establishes the inductive step. By the principle of mathematical induction, $\alpha^{m+n} = \alpha^m \cdot \alpha^n$ for all non-negative m and n .

If m and n are both non-positive, conclusion (ii) of the theorem now follows immediately by replacing α with α^{-1} .

It remains to check the case where one exponent is positive, the other negative. Without loss of generality, say $-m$ and n are positive.

Suppose first that $0 \leq m + n$. Since $-m$ is positive, the preceding argument shows

$$\alpha^n = \alpha^{(m+n)+(-m)} = \alpha^{m+n} \cdot \alpha^{-m}.$$

Multiplying both sides by α^m gives $\alpha^{m+n} = \alpha^m \cdot \alpha^n$, as claimed.

If instead $m + n < 0$, then $0 < -(m + n)$, and

$$\alpha^{-m} = \alpha^{-(m+n)+n} = \alpha^{-(m+n)} \cdot \alpha^n;$$

rearranging establishes the asserted claim.

(iii). For m and n non-negative, this follows by induction on the statement

$$P(n) \quad \alpha^{nm} = (\alpha^n)^m \quad \text{for all } m \text{ in } \mathbf{N}.$$

To establish the inductive step, note that

$$\alpha^{n(m+1)} = \alpha^{nm+n} = \alpha^{nm} \cdot \alpha^n = (\alpha^n)^m \cdot \alpha^n = (\alpha^n)^{m+1}$$

by (ii) and the definition of exponentiation. As a fringe benefit, we find that

$$(\alpha^m)^n = \alpha^{mn} = \alpha^{nm} = (\alpha^n)^m \quad \text{for all } \alpha \neq 0, \text{ all } m \text{ and } n \text{ in } \mathbf{N}.$$

If $m < 0$ or $n < 0$, replace α by α^{-1} and use $(\alpha^{-1})^n = \alpha^{-n}$. \square

Counting Mappings and Subsets

Throughout this section, m and n denote natural numbers (i.e., non-negative integers), and $\underline{\mathbf{m}}$ and $\underline{\mathbf{n}}$ are sets containing m and n elements. When we need to list elements, we write $\underline{\mathbf{m}} = \{1, 2, \dots, m\}$, with the understanding that $\underline{\mathbf{m}} = \emptyset$ if $m = 0$.

We give formulas for the number of mappings from $\underline{\mathbf{m}}$ to $\underline{\mathbf{n}}$, the number of *injective* mappings, and the number of distinct images of injective mappings, i.e., the number of m -element subsets of $\underline{\mathbf{n}}$.

Traditionally, these formulas are justified by informal “counting” arguments. In the interests of mathematical rigor, we use the counting arguments to “guess” the formulas, but establish the formulas using mathematical induction.

Proposition 5.25. *There are precisely n^m mappings from $\underline{\mathbf{m}}$ to $\underline{\mathbf{n}}$.*

Remark 5.26. This result is our first substantial justification for defining $0^0 = 1$. (The laws of exponents are compatible with $0^0 = 0$.)

Proof. Informally, each element of $\underline{\mathbf{m}}$ can be sent to any of n distinct values in $\underline{\mathbf{n}}$. Since these choices are independent, the total number of mappings is the product of m factors of n , i.e., n^m .

Formally, let $M_{m,n}$ denote the (unknown) number of distinct mappings from $\underline{\mathbf{m}}$ to $\underline{\mathbf{n}}$. The proof proceeds by induction on m :

$$P(m) \qquad M_{m,n} = n^m \qquad \text{for every } n \geq 0.$$

If $m = 0$, there exists a unique mapping from $\underline{\mathbf{m}}$ to $\underline{\mathbf{n}}$: Indeed, $\emptyset \times \underline{\mathbf{n}} = \emptyset$, and the empty set (the unique subset of $\emptyset \times \underline{\mathbf{n}}$) vacuously satisfies the definition of a mapping with domain \emptyset . That is, the base case $P(0)$ is true: $M_{0,n} = 1 = n^0$ for all $n \geq 0$.

Assume inductively that $P(k)$ is true for some $k \geq 0$. A mapping $f : \underline{\mathbf{k} + 1} \rightarrow \underline{\mathbf{n}}$ is uniquely determined by

- The $M_{k,n} = n^k$ choices of the restriction $f|_{\underline{\mathbf{k}}} : \underline{\mathbf{k}} \rightarrow \underline{\mathbf{n}}$;
- The n choices of $f(k + 1)$.

By the inductive hypothesis and the definition of exponentiation,

$$M_{k+1,n} = M_{k,n} \cdot n = n^k \cdot n = n^{k+1} \quad \text{for every } n \geq 0. \quad \square$$

Remark 5.27. This argument correctly predicts that there exist *no* mappings from $\underline{\mathbf{m}}$ to $\underline{\mathbf{0}} = \emptyset$ if $m \geq 1$. If necessary, re-examine the definition of a mapping to see why the empty set does not define a mapping with non-empty domain and empty target.

Definition 5.28. Let \mathcal{U} be a set of n elements. An *ordering* of \mathcal{U} is a bijection $s : \underline{\mathbf{n}} \rightarrow \mathcal{U}$, namely a listing $(s_k)_{k=1}^n$ of the elements of \mathcal{U} .

Example 5.29. The set $\mathcal{U} = \{a, b, c\}$ be ordered in six ways. In “alphabetical” order:

$$(a, b, c), \quad (a, c, b), \quad (b, a, c), \quad (b, c, a), \quad (c, a, b), \quad (c, b, a).$$

Proposition 5.30. *Let \mathcal{U} a set of n elements. There exist $n!$ distinct orderings of \mathcal{U} .*

Remark 5.31. The proposition justifies the definition $0! = 1$: The unique mapping $f : \emptyset \rightarrow \emptyset$ is vacuously bijective. It therefore suffices to prove the proposition for $n \geq 1$.

Proof. Informally, if $n \geq 1$, there are n ways to choose s_1 , and then $(n - 1)$ ways to choose a distinct s_2 , then $(n - 2)$ ways to choose s_3 , and so on. The total number of choices, i.e., the number of ways of ordering \mathcal{U} , is therefore $n(n - 1)(n - 2) \cdots 3 \cdot 2 \cdot 1 = n!$.

Formally, let C_n denote the (unknown) number of orderings of a set of n elements. The proof proceeds by induction on n :

$$P(n) \qquad C_n = n!.$$

The base case $C_1 = 1!$ asserts there is $1! = 1$ ordering of a set of one element. This is clear: There is only one mapping between sets containing one element, and this mapping is obviously a bijection.

Assume inductively that $C_k = k!$ for some $k \geq 1$. Let \mathcal{U} be a set containing $(k+1)$ elements. We wish to count the number of bijections $f : \underline{k+1} \rightarrow \mathcal{U}$. By Exercise 4.6 (b), every bijection f is uniquely determined by

- The $(k+1)$ choices of element $f(k+1)$ of \mathcal{U} ;
- The C_k choices of bijection $f|_{\underline{k}} : \underline{k} \rightarrow \mathcal{U} \setminus \{f(k+1)\}$.

By the inductive hypothesis and the definition of factorials,

$$C_{k+1} = (k+1) C_k = (k+1) k! = (k+1)!. \quad \square$$

Example 5.32. A 52-card deck of playing cards can be shuffled into

$$52! = 80,658,175,170,943,878,571,660,636,856,403,766, \\ 975,289,505,440,883,277,824,000,000,000,000,$$

or about $8.065817517 \times 10^{67}$ orderings.

To put the vastness of this number into perspective, the age of the visible universe is roughly 4.4×10^{17} seconds, and the visible universe is estimated to contain roughly 10^{80} atoms, give or take a couple of orders of magnitude. Since a human body contains roughly 10^{27} atoms, the visible universe contains enough matter to form $10^{80} \div 10^{27} = 10^{53}$ card dealers.* If these dealers shuffled decks once every second (with no rest breaks for 13.7 billion years), they would have performed roughly $4.4 \times 10^{17} \times 10^{53} = 4.4 \times 10^{70}$ shuffles since the big bang, enough to expect to have seen every possible ordering, but only a few hundred times. In our actual universe, where the total number of earthly card shuffles surely does not exceed 10^{15} (a billion dealers each shuffling one million times), we can safely say the number of shufflings ever seen is a vanishingly small fraction of all possible shufflings.

Definition 5.33. Let \mathcal{U} be a set of n elements. An *ordered m -set* from \mathcal{U} is an injective mapping $f : \underline{m} \rightarrow \mathcal{U}$. The image of an ordered m -set is the (*associated*) *unordered m -set*.

Remark 5.34. These terms are not standard, and are introduced primarily for convenience in the remainder of this section.

*Most of the universe consists of hydrogen, while a card dealer is largely made up of elements more than ten times heavier than hydrogen. Further, our estimate puts aside necessary support infrastructure: planets with habitable surface environments, resort cities, and casinos with all-you-can-eat buffets.

Example 5.35. The 4-element set $\mathcal{U} = \{a, b, c, d\}$ has twelve ordered 2-sets. In alphabetical order:

$$\begin{array}{cccccc} (a, b) & (a, c) & (a, d) & (b, a) & (b, c) & (b, d) \\ (c, a) & (c, b) & (c, d) & (d, a) & (d, b) & (d, c). \end{array}$$

Because of the ordering, $(b, a) \neq (a, b)$ and so forth. Each unordered 2-set appears exactly twice. (Why?)

Proposition 5.36. *Let \mathcal{U} be a set of n elements. If $0 \leq m \leq n$, there exist precisely*

$$n(n-1)(n-2) \cdots (n-m+1) = \frac{n!}{(n-m)!}$$

distinct ordered m -sets from \mathcal{U} .

Proof. Informally, following the idea of Proposition 5.36, there are n ways to choose s_1 , and then $(n-1)$ ways to choose a distinct s_2 , then $(n-2)$ ways to choose s_3 , etc., and $(n-m+1)$ ways to choose s_m .

Formally, let $O_{m,n}$ denote the (unknown) number of ordered m -sets from \mathcal{U} . The proof proceeds by induction on m :

$$P(m) \quad O_{m,n} = \frac{n!}{(n-m)!} \quad \text{for every } n \geq 0.$$

The unique mapping $s : \emptyset \rightarrow \mathcal{U}$ is vacuously injective. That is, the base case $P(0)$ is true: $O_{0,n} = 1 = n!/n!$.

Assume inductively that $P(k)$ is true for some integer k satisfying $0 \leq k \leq n-1$. Every injective mapping $s : \underline{k+1} \rightarrow \mathcal{U}$ is uniquely determined by

- $O_{k,n}$ choices of restriction $f|_{\underline{k}} \rightarrow \mathcal{U}$;
- The $(n-k)$ “remaining” choices of element $f(k+1)$ in the complement $\mathcal{U} \setminus f(\underline{k})$.

That is, $O_{k+1,n} = O_{k,n} \cdot (n-k)$. By the inductive hypothesis,

$$O_{k+1,n} = \frac{n!}{(n-k)!} \cdot (n-k) = \frac{n!}{(n-k-1)!} = \frac{n!}{(n-(k+1))!}. \quad \square$$

Remark 5.37. If $n < m$, the number of ordered m -sets from \mathcal{U} is 0. Be sure you understand this statement conceptually, and trace through the inductive step for $m = n$ to see why this outcome is predicted.

We come to the major goal of this subsection: Counting (unordered) m -element subsets of a set of n elements.

Definition 5.38. Let \mathcal{U} be a set of n elements. The *binomial coefficient* $\binom{n}{m}$, read “ n choose m ”, is defined to be the number of distinct subsets of \mathcal{U} having precisely m elements.

Remark 5.39. Each binomial coefficient $\binom{n}{m}$ is a non-negative integer, and $\binom{n}{m} = 0$ unless $0 \leq m \leq n$.

Further, $\binom{n}{m} = \binom{n}{n-m}$: If \mathcal{U} contains n elements, then to each m -element subset A of \mathcal{U} is uniquely associated its complement $\mathcal{U} \setminus A$, having $(n - m)$ elements.

Proposition 5.40. *If m and n are arbitrary integers, then*

$$\binom{n}{m} = \begin{cases} \frac{n!}{m!(n-m)!} & \text{if } 0 \leq m \leq n, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. If $m < 0$ or $n < m$, there are no m -sets from \mathcal{U} .

Suppose $0 \leq m \leq n$, and let \mathcal{U} be a set of n elements.

By Proposition 5.36, there are precisely $n!/(n-m)!$ ordered m -sets from \mathcal{U} . By Proposition 5.30, each unordered m -set from \mathcal{U} is associated to precisely $m!$ ordered m -sets from \mathcal{U} . Combining these observations,

$$\frac{n!}{(n-m)!} = O_{m,n} = m! \cdot \binom{n}{m}, \quad \text{or} \quad \binom{n}{m} = \frac{n!}{m!(n-m)!}. \quad \square$$

5.3 The Binomial Theorem

The identity $(\alpha + \beta)^2 = \alpha^2 + 2\alpha\beta + \beta^2$ is doubtless familiar. The *binomial theorem* generalizes to arbitrary positive integer powers $(\alpha + \beta)^n$.

Theorem 5.41 (Binomial Theorem). *If α and β are complex numbers and n is a non-negative integer,*

$$\begin{aligned} (\alpha + \beta)^n &= \sum_{k=0}^n \binom{n}{k} \alpha^{n-k} \beta^k \\ &= \binom{n}{0} \alpha^n + \binom{n}{1} \alpha^{n-1} \beta + \binom{n}{2} \alpha^{n-2} \beta^2 + \cdots + \binom{n}{n} \beta^n. \end{aligned}$$

Proof. Conceptually, the n -fold product

$$(\alpha + \beta)^n = (\alpha + \beta)(\alpha + \beta) \cdots (\alpha + \beta)$$

is expanded by the following procedure:

- Pick an arbitrary integer k with $0 \leq k \leq n$;
- Distribute k check marks among the n copies of $(\alpha + \beta)$;
- If a copy of $(\alpha + \beta)$ is unchecked, choose α from that copy; otherwise choose β . Multiply the resulting n factors to get $\alpha^{n-k}\beta^k$.
- Sum over all k and all ways of distributing k check marks.

By the first and third points, the expanded product has the form

$$(\alpha + \beta)^n = \text{---} \alpha^n + \text{---} \alpha^{n-1}\beta + \text{---} \alpha^{n-2}\beta^2 + \cdots + \text{---} \alpha\beta^{n-1} + \text{---} \beta^n$$

for some coefficients. By the second and fourth points, the coefficient of $\alpha^{n-k}\beta^k$ is $\binom{n}{k}$, the number of distinct ways of distributing k check marks among n parenthesized binomials. This completes the proof. \square

Pascal's Triangle

The binomial coefficients for any particular exponent n can be found in the $(n + 1)$ th row of a recursive diagram known as “Pascal’s triangle”.

Imagine expanding successive powers of $(\alpha + \beta)$ recursively, in as lazy a manner as possible. Because

$$(\alpha + \beta)^{n+1} = (\alpha + \beta)^n(\alpha + \beta) = (\alpha + \beta)^n\alpha + (\alpha + \beta)^n\beta,$$

knowledge of $(\alpha + \beta)^n$ can be “recycled” in calculating $(\alpha + \beta)^{n+1}$.

Starting from $n = 2$ and $(\alpha + \beta)^2 = \alpha^2 + 2\alpha\beta + \beta^2$, we find

$$\begin{aligned} (\alpha + \beta)^3 &= (\alpha^2 + 2\alpha\beta + \beta^2)\alpha + (\alpha^2 + 2\alpha\beta + \beta^2)\beta \\ &= \alpha^3 + 2\alpha^2\beta + \alpha\beta^2 \\ &\quad + \alpha^2\beta + 2\alpha\beta^2 + \beta^3 \\ &= \alpha^3 + 3\alpha^2\beta + 3\alpha\beta^2 + \beta^3. \end{aligned}$$

	0	0	0	1	0	0	0	0	...
0	0	0	1	1	0	0	0	0	...
	0	0	1	2	1	0	0	0	...
0	0	1	3	3	1	0	0	0	...
	0	1	4	6	4	1	0	0	...
0	1	5	10	10	5	1	0	0	...
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
	$k = -1$	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$...	
$n = 0$	0	1	0	0	0	0	0	0	...
$n = 1$	0	1	1	0	0	0	0	0	...
$n = 2$	0	1	2	1	0	0	0	0	...
$n = 3$	0	1	3	3	1	0	0	0	...
$n = 4$	0	1	4	6	4	1	0	0	...
$n = 5$	0	1	5	10	10	5	1	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Table 5.1: Pascal's triangle, "classic" format (top) and tabular.

Recycling this formula gives

$$\begin{aligned}
 (\alpha + \beta)^4 &= (\alpha^3 + 3\alpha^2\beta + 3\alpha\beta^2 + \beta^3)\alpha + (\alpha^3 + 3\alpha^2\beta + 3\alpha\beta^2 + \beta^3)\beta \\
 &= \alpha^4 + 3\alpha^3\beta + 3\alpha^2\beta^2 + \alpha\beta^3 \\
 &\quad + \alpha^3\beta + 3\alpha^2\beta^2 + 3\alpha\beta^3 + \beta^4 \\
 &= \alpha^4 + 4\alpha^3\beta + 6\alpha^2\beta^2 + 4\alpha\beta^3 + \beta^4.
 \end{aligned}$$

To be sure you understand, check in detail that

$$(\alpha + \beta)^5 = \alpha^5 + 5\alpha^4\beta + 10\alpha^3\beta^2 + 10\alpha^2\beta^3 + 5\alpha\beta^4 + \beta^5.$$

A useful pattern is emerging: If the coefficients for $(\alpha + \beta)^n$ are laid out in a row, "duplicate" the row underneath itself, shift the second copy one entry to the right, and add in columns to get the coefficients for $(\alpha + \beta)^{n+1}$. Alternatively, "pad" the coefficients on either side with an infinite row of 0's. Then, to get the entries in the "next" row, add each entry to its neighbor on the left. The result is *Pascal's triangle*.

Table 5.1 shows two versions of Pascal's triangle. On top is the "classic" format, in which each entry below the first row is the sum of its "parents", the two nearest entries in the preceding row.

In the tabular formatting of Pascal's triangle, the entries are arranged so that n indexes the rows and k indexes the columns.

Example 5.42. The binomial theorem can be used with specific numbers. For example,

$$\begin{aligned} 11^3 &= (10 + 1)^3 = 10^3 + 3 \cdot 10^2 \cdot 1 + 3 \cdot 10 \cdot 1^2 + 1^3 \\ &= 1000 + 300 + 30 + 1 = 1331. \end{aligned}$$

The digits constitute the fourth row of Pascal's triangle.

Exercises

Exercise 5.1. Use the indicated strategies to find the sum of the first n positive integers.

- Compute a few special cases, formulate a conjecture, and use mathematical induction to prove your formula is correct.
- Starting with the formula for the sum of the first n odd positive integers, increment each summand by 1 to get the sum of the first n even positive integers. Add these sums to get the sum of the first $N = 2n$ integers, and express the result in terms of N .
- Observe $1 + 2 + \cdots + (n - 1) + n = n + (n - 1) + \cdots + 2 + 1$; add these expressions to each other and group the respective first terms, second terms, and so forth. Now solve for the unknown sum.

Exercise 5.2. Use the binomial theorem to calculate $9^3 = (10 - 1)^3$, $11^4 = (10 + 1)^4$, and $12^3 = (10 + 2)^3$. Do not use a calculator.

Exercise 5.3. Let $x \geq 0$ be a real number and n a non-negative integer.

- Use mathematical induction to prove $(1 + x)^n \geq 1 + nx$.
- Use the binomial theorem to prove $(1 + x)^n \geq 1 + nx$.

Exercise 5.4. Let r be a complex number, and let

$$S_n(r) = \sum_{k=0}^n r^k = 1 + r + r^2 + \cdots + r^n.$$

- Use induction to prove $1 + rS_n(r) = S_{n+1}(r)$ for all $n \geq 0$.

- (b) Use part (a) and the identity $S_{n+1}(r) = S_n(r) + r^{n+1}$ to prove $(1-r)S_n(r) = 1 - r^{n+1}$ for all $n \geq 0$.
- (c) Find a closed expression (the *finite geometric series formula*) for $S_n(r)$. (Handle the cases $r = 1$ and $r \neq 1$ separately.)
- (d) Calculate (and simplify): $\sum_{k=0}^n 9 \cdot \left(\frac{1}{10}\right)^k$, $\sum_{k=0}^n (-1)^k$, $\sum_{k=0}^{100} \frac{1}{2} \cdot \left(\frac{1}{4}\right)^k$.

Exercise 5.5. Recursively define the *double factorial* of a non-negative integer n by

$$0!! = 1, \quad 1!! = 1, \quad n!! = n(n-2)!!$$

- (a) Without using a calculator, evaluate the double factorials up to $10!!$.
- (b) Informally, $n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$. Find similar expressions for $(2n)!!$ and $(2n+1)!!$.
- (c) Express $(2n)!!(2n-1)!!$ (for $1 \leq n$) and $(2n)!!(2n+1)!!$ (for $0 \leq n$) in terms of ordinary factorials. Use mathematical induction to prove your formulas are correct.
- (d) If $n = 2k$ is even, find a formula for $n!!$ in terms of $k!$. Use mathematical induction to prove your formula is correct.
- (e) Use mathematical induction to prove $(2k)!!$ is the number of ways of partitioning a set of $2k$ elements into k pairs.

Exercise 5.6. With the help of the binomial theorem, expand:

- (a) $(a+b)^3$, $(a-b)^3$, and $\frac{1}{2}[(a+b)^3 \pm (a-b)^3]$.
- (b) $(a+b)^4$, $(a-b)^4$, and $\frac{1}{2}[(a+b)^4 \pm (a-b)^4]$.
- (c) $(a+b)^6$. (Hint: Extend Pascal's triangle.)

Exercise 5.7. Suppose $i^2 = -1$. Use the binomial theorem to expand:

- (a) $(x+iy)^2$. (b) $(x+iy)^3$. (c) $(x+iy)^4$.

In each part, separate the real and imaginary parts.

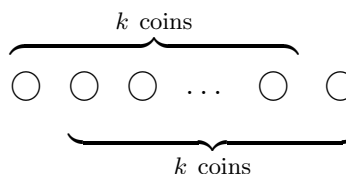
Exercise 5.8. Use the binomial theorem to establish the identities:

- (a) $\sum_{k=0}^n \binom{n}{k} = 2^n$ for $n \geq 0$. (b) $\sum_{k=0}^n (-1)^k \binom{n}{k} = 0$ for $n \geq 1$.

Exercise 5.9. Let α , β , and γ be complex numbers, and $n \geq 0$ an integer. State and prove a “trinomial theorem” for $(\alpha + \beta + \gamma)^n$.

Exercise 5.10. Consider the following “proof” that all coins have the same denomination:

Let $P(n)$ be the statement “In a set of n coins, all the coins have the same denomination.” Now, $P(1)$ is clearly true (a single coin has a single denomination), so the base case is true. Assume inductively that $P(k)$ is true for some $k > 1$, and divide an arbitrary set of $(k+1)$ coins into two groups as shown at right.



By the inductive hypothesis, the first k coins all have the same denomination, and the last k coins have the same denomination. Since these two sets “overlap” as shown, all the coins have the same denomination. Since $P(k)$ implies $P(k+1)$ for all $k > 1$, $P(n)$ is true for all n , namely, all the coins have the same denomination. Where, exactly, are the logical flaws in this argument?

Exercise 5.11. Let $n \geq 2$ be an integer, and consider the complex number $\zeta = e^{2\pi i/n}$, an n th root of unity, see Example 2.39. Prove that

$$1 + \zeta + \zeta^2 + \cdots + \zeta^{n-1} = \sum_{k=0}^{n-1} \zeta^k = 0$$

in two ways:

- (a) Using the geometric series formula from Exercise 5.4.
- (b) Calling the unknown sum S , and multiplying by ζ .

Exercise 5.12. Let $n \geq 2$ be an integer.

- (a) Show that the polynomial $z^n - 1$ factors as

$$z^n - 1 = (z - 1)(z - \zeta)(z - \zeta^2) \cdots (z - \zeta^{n-1}) = \prod_{k=0}^{n-1} (z - \zeta^k).$$

Hint: Each side has the same roots and the same leading coefficient.

(b) Use part (a) and the geometric series formula to show that

$$\sum_{j=0}^{n-1} z^j = 1 + z + z^2 + \cdots + z^{n-1} = \prod_{k=1}^{n-1} (z - \zeta^k).$$

(c) By setting $z = 1$ in part (b), prove that

$$n = \prod_{k=1}^{n-1} (1 - \zeta^k) = \prod_{k=1}^{n-1} |1 - \zeta^k|.$$

[This identity has a beautiful geometric interpretation: Inscribe a regular n -gon in the unit circle. Fix a vertex, and consider the $(n-1)$ chords joining that vertex to each of the other vertices. The product of the lengths of these chords is n , the number of sides of the polygon.]

Exercise 5.13. Under modest technical assumptions, a complex-valued function f on the interval $[-\pi, \pi]$ can be approximated by *Fourier polynomials*, namely “linear combinations” of the form

$$\frac{a_0}{2} + \sum_{k=1}^N [a_k \cos(k\phi) + b_k \sin(k\phi)], \quad a_k, b_k \text{ complex scalars.}$$

The identities below are useful in showing that a suitable sequence of Fourier polynomials “converges to f ”.

Let ϕ be a real number. Recall that by Exercise 4.14,

$$\cos \phi = \frac{e^{i\phi} + e^{-i\phi}}{2}, \quad \sin \phi = \frac{e^{i\phi} - e^{-i\phi}}{2i}.$$

(a) Prove that if $n \geq 0$, then

$$\sum_{k=-n}^n e^{ik\theta} = 1 + 2 \sum_{k=1}^n \cos(k\theta).$$

Hint: On the left, pair up the terms $e^{ik\theta}$ and $e^{-ik\theta}$.

(b) Sum the left-hand series in the preceding part, and prove that

$$1 + 2 \sum_{k=1}^n \cos(k\theta) = \frac{\sin(n + \frac{1}{2})\theta}{\sin \frac{1}{2}\theta}.$$

Hints: Use the geometric series formula from Exercise 5.4; multiply and divide the resulting fraction by $e^{-i\theta/2}$; and use the formula above for $\sin \phi$.

Chapter 6

Binary Operations

Ordinary addition of integers may be viewed as mapping each ordered pair (a, b) of integers to the integer $a + b$. Multiplication of integers has precisely the same *abstract* description, sending an ordered pair of integers to an integer. Moreover, the axioms for these particular operations (page 34) share features, including associativity, commutativity, and existence of identity elements.

This chapter discusses “binary operations”, mathematical functions that accept an ordered pair of objects of some type and return an object of the same type. Algebraic notions, such as associativity and identity elements, make sense in this general setting.

Once we have established a property of general binary operations, such as uniqueness of identity elements, we are assured the property holds automatically each time we encounter a new example, whether it be addition and multiplication of complex numbers, or composition of maps from a set X to itself.

Definition 6.1. Let A be a non-empty set. A *binary operation* on A is a mapping $\mu : A \times A \rightarrow A$.

Remark 6.2. Conceptually, a binary operation is a rule for combining two elements of A to obtain an element of A . If a and b are elements of A , we usually write ab instead of $\mu(a, b)$. The expressions $a \cdot b$ or $a * b$ are used to emphasize the operation, especially when more than one binary operation is under consideration.

Example 6.3. The familiar operations of addition, multiplication, and subtraction are binary operations on \mathbf{Z} , the set of integers.

Division is *not* a binary operation on \mathbf{Z} since, for example, $1 \div 0$ and $1 \div 2$ do not represent integers.

Example 6.4. Addition, multiplication, and exponentiation are binary operations on the set \mathbf{Z}^+ of positive integers.

Example 6.5. Let X be an arbitrary set, and let $\mathcal{M}(X)$ be the set of all mappings $f : X \rightarrow X$. Function composition, $\mu(g, f) = g \circ f$, is a binary operation on $A = \mathcal{M}(X)$.

Example 6.6. Let X be an arbitrary set. The intersection operator defines a binary operation on $A = \mathcal{P}(X)$, the power set of X . Similarly, the union operator defines a binary operation on $\mathcal{P}(X)$.

When A is a finite set of n elements, a binary operation may be represented by a *Cayley table*, an $n \times n$ tabular listing of all products, $\mu(a, b) = ab$ being placed in the “ a th row” and “ b th column”.

Example 6.7. Let $A = \{E, O\}$ be a set with two elements, which we view as representing a general *even* integer (E) and a general *odd* integer (O). The Cayley table

$+$	E	O
E	E	O
O	O	E

expresses the fact that a sum of two even integers or two odd integers is even, while the sum of an even and an odd integer is odd.

Example 6.8. Let $A = \{a, b, c\}$ be a set with three elements. The following Cayley tables define binary operations on A :

μ_1	a	b	c	μ_2	a	b	c	μ_3	a	b	c
a	a	c	b	a	a	b	c	a	a	a	a
b	b	a	c	b	b	c	a	b	a	b	c
c	c	b	a	c	c	a	b	c	a	c	b

We have, e.g., $\mu_1(b, a) = b$, (second row, first column of the first table), while $\mu_1(a, b) = c$, $\mu_2(a, b) = b$, and $\mu_3(a, b) = a$ from the first row, second column of the respective tables.

Example 6.9. Let $B = \{0, 1, 2\}$. The following tables define binary operations on B “isomorphic to” those of Example 6.8:

μ_1	0	1	2	μ_2	0	1	2	μ_3	0	1	2
0	0	2	1	0	0	1	2	0	0	0	0
1	1	0	2	1	1	2	0	1	0	1	2
2	2	1	0	2	2	0	1	2	0	2	1

The concept of “abstractly identical” operations, formalized in Example 6.10, boils down to *relabeling*.

Example 6.10. Given a binary operation \cdot on a set A and a bijection $\phi : A \rightarrow B$, define a binary operation $*$ on B as follows:

$$(6.1) \quad \phi(a_1) * \phi(a_2) = \phi(a_1 \cdot a_2) \text{ for all } a_1 \text{ and } a_2 \text{ in } A.$$

In words, attach each element b in B to its “avatar”, the unique element a in A such that $\phi(a) = b$. Since the binary operation \cdot combines pairs of avatars, Equation (6.1) tells us how to combine elements of B .

This relationship can be visualized as a “commutative diagram”, Figure 6.1. Starting with a pair (a_1, a_2) in $A \times A$, there are two ways of getting to an element of B : (i) Multiply the elements in A (left edge) and map the product to B by ϕ (bottom edge), or (ii) map the elements individually to B (top edge) then multiply in B (right edge).

The diagram “commutes” because these two mapping compositions yield the same value for all pairs of inputs.

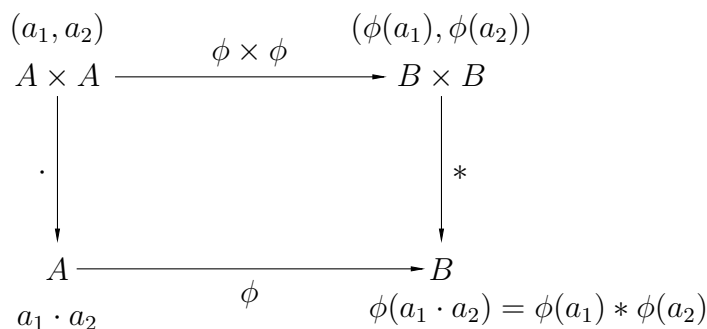


Figure 6.1: Isomorphism depicted as a commutative diagram.

Example 6.11. Consider Examples 6.8 (where $A = \{a, b, c\}$) and 6.9 ($B = \{0, 1, 2\}$). If we define $\phi : A \rightarrow B$ by $\phi(a) = 0$, $\phi(b) = 1$, and $\phi(c) = 2$, the respective binary operations are related as in Example 6.10. Be sure you understand this example in detail.

6.1 Properties of Binary Operations

Definition 6.12. Suppose μ is a binary operation on A , and $S \subseteq A$ is non-empty. We say S is *closed under μ* if $\mu(s_1, s_2) \in S$ for all s_1 and s_2 in S .

If S is closed under μ , then the “restricted” mapping $\mu : S \times S \rightarrow S$ is a binary operation on S .

Example 6.13. Let $\mu : \mathbf{Z} \times \mathbf{Z} \rightarrow \mathbf{Z}$ be addition: $\mu(a, b) = a + b$.

The set $S = 2\mathbf{Z}$ of even integers is closed under addition: A sum of even integers is even.

Similarly, the set $S = \mathbf{Z}^+$ of positive integers is closed.

The set $S = \{0, 1\}$ is *not* closed under addition: $s_1 = 1$ and $s_2 = 1$ are elements of S , but $s_1 + s_2 = 2 \notin S$.

The set $S = 2\mathbf{Z} + 1$ of odd integers is not closed in a particularly strong way: A sum of odd integers is *never* odd.

Remark 6.14. Note carefully that the existence of a *single pair* s_1, s_2 in S with $\mu(s_1, s_2)$ not in S is enough to prove S is not closed under μ . However, examples do not suffice to show S *is* closed under μ .

Associativity

By definition, a binary operation gives rise to “products” involving two elements. In practice, we often wish to combine three or more elements. This gives rise to a potential ambiguity: When we write a product “ abc ”, we might mean either

$$(ab)c = \mu(\mu(a, b), c) \quad \text{or} \quad a(bc) = \mu(a, \mu(b, c)).$$

In general these expressions represent different elements of A . For the binary operation μ_1 of Example 6.8, we have

$$(ba)c = bc = c, \quad b(ac) = bb = a.$$

Definition 6.15. A binary operation μ on a set A is *associative* if $a(bc) = (ab)c$ for all a, b , and c in A .

Example 6.16. Addition and multiplication are associative operations on the set of integers by Axioms A1. and M1, page 34. (Addition and multiplication are also associative on the larger sets of rational numbers, real numbers, and complex numbers.)

Example 6.17. Subtraction on \mathbf{C} is *not* associative: If $c \neq 0$, then

$$(a - b) - c \neq a - (b - c) = (a - b) + c.$$

Similarly, division is not associative on the set of non-zero complex numbers.

Example 6.18. Exponentiation defines a binary operation on \mathbf{Z}^+ , but this operation is not associative. Even in the special case $a = b = c$, we have $a^{(a^a)} = (a^a)^a$ if and only if $a = 1$ or $a = 2$.

Example 6.19. If A is a set, then as noted earlier, mapping composition is a binary operation on $\mathcal{M}(A)$, the set of all mappings $f : A \rightarrow A$. This operation is automatically associative by Proposition 4.31.

Example 6.20. Let A be a set. By Exercise 2.11, the operations of intersection and union are associative binary operations on $\mathcal{P}(A)$, the power set of A .

When a binary operation is associative, products of any finite number of factors may be grouped arbitrarily (preserving only the order of the factors) without changing the result. For example,

$$a((bc)d) = a(b(cd)) = (ab)(cd) = ((ab)c)d = \dots$$

Rather than proving directly that any two groupings of a product have the same value, we'll pick one specific grouping, and show that an arbitrary grouping has the same value.

Definition 6.21. Let A be a non-empty set equipped with a binary operation. A product of n elements is *grouped from the right* if pairs of factors are grouped from right to left:

$$a_1 \left(a_2 \left(a_3 \dots (a_{n-1} a_n) \dots \right) \right).$$

Proposition 6.22. *Let A be a non-empty set equipped with an associative binary operation. If a_1, \dots, a_n is an ordered n -tuple of elements of A , then every grouping of these n factors has the same value as the grouping from the right.*

In particular, an arbitrary grouping of the factors *taken in order from left to right* has the same value.

Proof. If n is an integer with $n \geq 3$, let $P(n)$ denote the statement:

Every m -fold product with $3 \leq m \leq n$ can be regrouped from the right without changing the value of the product.

The associative law says every threefold product can be regrouped from the right without changing the value of the product; $P(3)$ is true.

Assume inductively that $P(k)$ is true for some integer $k \geq 3$. Every grouping of a product of $(k+1)$ factors a_1, \dots, a_{k+1} may be viewed as a product $A_1 A_2$ of two factors, with each of A_1 and A_2 a product of k or fewer factors. By the inductive hypothesis, we may regroup A_1 from the right without changing the value of the product, say $A_1 = a_1 A'_1$. By associativity,

$$A_1 A_2 = (a_1 A'_1) A_2 = a_1 (A'_1 A_2).$$

The product $(A'_1 A_2)$ has k factors, and by the inductive hypothesis can be regrouped from the right without changing the value of the product. The previous equation therefore states that $A_1 A_2$ can be regrouped from the right without changing the value of the product. \square

Identity Elements

Definition 6.23. Let (A, μ) be a set equipped with a binary operation. An element e in A is an *identity element* for μ if

$$ea = ae = a \quad \text{for all } a \text{ in } A.$$

A binary operation may have no identity element at all. However, there can be at most one, for if e and e' are identity elements for μ , then $e = ee'$ (since e' is an identity element) and $ee' = e'$ (since e is an identity element), so $e = e'$.

Example 6.24. The integer 0 is the identity element for addition on \mathbf{Z} . The integer 1 is the identity element for multiplication on \mathbf{Z} .

Example 6.25. There is no identity element for subtraction. In other words, there exists no integer e such that $a - e = e - a = a$ for every integer a . (Why not?)

Example 6.26. An identity element can be located at a glance from a Cayley table: The corresponding row and column of the table will contain the same entries as the “index” entries across the top and down the left side. Consider the operations in Example 6.8:

μ_1	a	b	c	μ_2	a	b	c	μ_3	a	b	c
a	a	c	b	a	a	b	c	a	a	a	a
b	b	a	c	b	b	c	a	b	a	b	c
c	c	b	a	c	c	a	b	c	a	c	b

μ_1 has no identity element, while a is the identity element for μ_2 and b is the identity element for μ_3

Example 6.27. Let \mathcal{U} be a universe. Since $A \cup \emptyset = \emptyset \cup A = A$ for every subset $A \subseteq \mathcal{U}$, $e = \emptyset$ for the union operator.

Dually, since $A \cap \mathcal{U} = \mathcal{U} \cap A = A$ for all A , $e = \mathcal{U}$ is the identity element for the intersection operator.

Inverse Elements

Definition 6.28. Let (A, μ) be a set equipped with a binary operation, and assume there is an identity element for μ . If $a \in A$, then an element b in A is an *inverse* of a (with respect to μ) if

$$ab = ba = e.$$

Remark 6.29. It makes no sense to ask about inverses unless μ has an identity element. Moreover, even if μ has an identity element, a specific element a in A may or may not have an inverse.

Remark 6.30. If μ is *associative* and has an identity element e , then each element a has at most one inverse, see Exercise 6.3. Briefly, inverses are unique with respect to an associative operation. The inverse of a is normally denoted a^{-1} . A commutative binary operation is customarily denoted $+$, in which case the inverse of a is denoted $-a$.

Example 6.31. Let $A = \mathbf{Z}$. By Axioms A2. and A3., page 34, addition has identity element 0, and every integer a has an additive inverse, $-a$.

Example 6.32. Let $A = \mathbf{Z}$. By Axiom M2., page 34, multiplication has an identity element 1. The only invertible integers are 1 and -1 , each being its own inverse.

Example 6.33. Let $A = \mathbf{Q}^\times$ be the set of non-zero rational numbers, equipped with the operation of multiplication. The rational number 1 is the identity element, and every non-zero rational $a = p/q$ has an inverse, $a^{-1} = q/p$.

Example 6.34. The complex number 1 is the identity element for multiplication on the set \mathbf{C}^\times of non-zero complex numbers. If $\alpha = a + bi$ is non-zero, then $\alpha^{-1} = (a - bi)/(a^2 + b^2)$, see Example 2.29.

Remark 6.35. (Non-)existence of inverse elements can be read off a Cayley table. First locate the identity element e ; if none exists, nothing further need be done.

To seek the inverse of a specific element a , inspect the a th row of the table, looking for the identity element e . If e is found in the b th column (signifying $ab = e$), check to see whether $ba = e$ as well. If so, $a^{-1} = b$; otherwise a has no inverse.

Example 6.36. The operation μ_1 in Example 6.8 has no identity element, so the concept of inverses makes no sense.

For μ_2 , $a^{-1} = a$, $b^{-1} = c$, and $c^{-1} = b$; every element has an inverse.

For μ_3 , $b^{-1} = b$ and $c^{-1} = c$, but a has no inverse.

Example 6.37. Let A be a non-empty set, $\mathcal{M}(A)$ the set of mappings from A to A equipped with the operation of function composition.

The identity mapping $I_A : A \rightarrow A$ is the identity element for composition. A mapping $f : A \rightarrow A$ has an inverse in the sense of binary operations if and only if f is a bijection, if and only if f is invertible in the sense of mappings.

Example 6.38. In Example 6.27, we saw that the union operation on $\mathcal{P}(A)$ has identity element $e = \emptyset$. If $S \subseteq A$, an inverse of S is a set T such that $S \cup T = \emptyset$. No such set exists unless $S = \emptyset$, in which case $T = \emptyset$ satisfies the conditions for an inverse element. In other words, $\emptyset^{-1} = \emptyset$, and no other set has an inverse with respect to the union. See also Exercise 6.4.

Commutativity

Definition 6.39. A binary operation on A is *commutative* if $ab = ba$ for all a and b in A .

Example 6.40. By Axioms A4. and M4., page 34, addition and multiplication are commutative operations on \mathbf{Z} : $a + b = b + a$ and $ab = ba$ for all integers a and b .

Example 6.41. Subtraction is not commutative on \mathbf{Z} : $1 - 2 \neq 2 - 1$.

Example 6.42. Set union and intersection are commutative on $\mathcal{P}(A)$.

Example 6.43. Function composition is not commutative. For example, if f and $g : \mathbf{R} \rightarrow \mathbf{R}$ are defined by $f(a) = a + 1$ and $g(a) = a^2$, then $(f \circ g)(a) = a^2 + 1$, while $(g \circ f)(a) = (a + 1)^2 = a^2 + 2a + 1$.

Example 6.44. Many real-life activities do not commute: Putting on your socks and putting on your shoes; removing car keys from the ignition and closing the locked car door; turning off the electricity and repairing the wiring; looking both ways and crossing the street. In each case, the result of one activity has some bearing on the success or failure of the other.

Exercises

Exercise 6.1. Each part refers to the indicated binary operations on the set $A = \{0, 1, 2, 3\}$.

\ominus	0	1	2	3	\odot	0	1	2	3
0	0	3	2	1	0	0	0	0	0
1	1	0	3	2	1	0	1	2	3
2	2	1	0	3	2	0	2	0	2
3	3	2	1	0	3	0	3	2	1

- Is \ominus associative? Is \ominus commutative? Does \ominus have an identity element? If so, which elements of A have inverses?
- Show \odot is associative. Suggestion: First show that if any operand is 0 or 1, then $a \odot (b \odot c) = (a \odot b) \odot c$.
- Is \odot commutative? Does \odot have an identity element? If so, which elements of A have inverses?

Exercise 6.2. Suppose A is a set of n elements and $*$ is a binary operation on A .

- How many conditions must be checked to prove $*$ is commutative? (The answer is not n^2 .) Describe how a Cayley table for $(A, *)$ can be used to check commutativity.
- How many conditions must be checked to prove $*$ is associative?
- For each a in A , construct two binary operations

$$\lambda_a(x, y) = (x * a) * y, \quad \rho_a(x, y) = x * (a * y).$$

Note that the operation $*$ is associative if and only if these operations have the same Cayley table for all a in A . Use this criterion

(*Light's test for associativity*) to check whether the following operations are associative:

\oplus	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

\odot	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

Exercise 6.3. Let (A, μ) be a set equipped with an associative binary operation, assume μ has an identity element e , and assume $a \in A$. Prove that if b and b' are inverses of a , then $b = b'$.

Hint: If you're stuck, you can find a very similar proof in Chapter 3.

Exercise 6.4. Consider the intersection operation on $\mathcal{P}(A)$. Determine which subsets of A (if any) are invertible, and find the inverse of any invertible set, cf. Example 6.38.

Exercise 6.5. On the set \mathbf{Z} of integers, define a binary operation by $a * b = a + b - 1$.

- Prove $*$ is associative and commutative.
- Prove $*$ has an identity element.
- Prove every integer has an inverse with respect to $*$.
- Define $\phi : \mathbf{Z} \rightarrow \mathbf{Z}$ by $\phi(a) = a + 1$. Starting with the operation of ordinary addition, use the method of Example 6.10 to define a new binary operation μ on \mathbf{Z} . Try to re-do the first three parts of this question by “transferring” a property of addition to the corresponding property of μ .

Exercise 6.6. Let $A = \{0, 1, 2, 3\}$ and $B = \{a, b, c, d\}$. Each part concerns the following binary operation on A :

\cdot	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

- Is \cdot commutative? Does \cdot have an identity element? If so, which elements have inverses?

- (b) Define a bijection $\phi : A \rightarrow B$ by $\phi(0) = a$, $\phi(1) = b$, $\phi(2) = c$, $\phi(3) = d$. Write out the Cayley table for the induced operation $*$ on B .
- (c) Is $*$ commutative? Does $*$ have an identity element? If so, which elements have inverses? How are your answers related to your answers for part (a)?
- (d) Show the set $\{a, c\} \subseteq B$ is closed under $*$, and write out the Cayley table for $*$ restricted to $\{a, c\}$.
- (e) Find all proper subsets of B that are closed under $*$.

Exercise 6.7. For each integer n , define the mapping $T_n : \mathbf{Z} \rightarrow \mathbf{Z}$ by $T_n(x) = x + n$, and let $\mathcal{M}(T) = \{T_n : n \in \mathbf{Z}\}$.

- (a) Prove that each T_n is a bijection, and find a formula for the inverse mapping.
- (b) Show that $\mathcal{M}(T)$ is closed under composition of functions.
- (c) Show that $\mathcal{M}(T)$ contains an identity element, and that $\mathcal{M}(T)$ is closed under inversion.

Exercise 6.8. For each integer n , define the mapping $S_n : \mathbf{Z} \rightarrow \mathbf{Z}$ by $S_n(x) = nx$, and let $\mathcal{M}(S) = \{S_n : n \in \mathbf{Z}\}$.

- (a) Determine (with proof) which mappings S_n are injective, and which are surjective.
- (b) Show that $\mathcal{M}(S)$ is closed under composition of functions.
- (c) Show that $\mathcal{M}(S)$ contains an identity element. Which elements of $\mathcal{M}(S)$ are invertible?

Exercise 6.9. For each non-zero real number a , define the mapping $S_a : \mathbf{R} \rightarrow \mathbf{R}$ by $S_a(x) = ax$, and let $\mathcal{M}(S) = \{S_a : a \in \mathbf{R}^\times\}$.

- (a) Prove that each mapping S_a is a bijection, and find a formula for the inverse mapping.
- (b) Show that $\mathcal{M}(S)$ is closed under composition of functions.
- (c) Show that $\mathcal{M}(S)$ contains an identity element, and that $\mathcal{M}(S)$ is closed under inversion.

Exercise 6.10. Let A be a non-empty set. The set \mathbf{R}^A of real-valued functions on A consists of *all mappings* $f : A \rightarrow \mathbf{R}$. Define binary operations of addition and multiplication on \mathbf{R}^A by

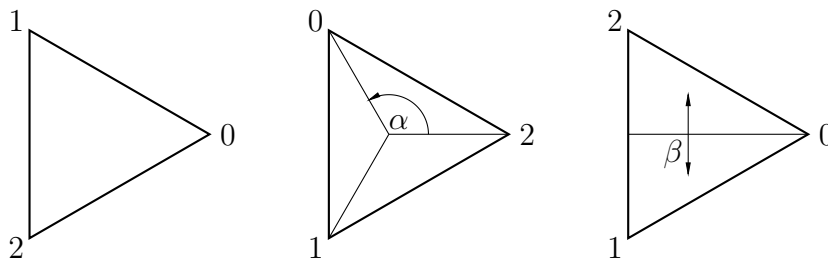
$$(f + g)(a) = f(a) + g(a), \quad (f \cdot g)(a) = f(a) \cdot g(a), \quad \text{for all } a \text{ in } A.$$

- (a) Prove that $+$ is an associative and commutative binary operation on \mathbf{R}^A , there is an identity element for $+$, and every element of \mathbf{R}^A has an additive inverse.
- (b) Prove that \cdot is an associative and commutative binary operation on \mathbf{R}^A . Does \cdot have an identity element? If so, which elements of \mathbf{R}^A have a multiplicative inverse?

Exercise 6.11. This question concerns a set of mappings from the plane \mathbf{R}^2 to itself. Let e be the identity map, α the counterclockwise quarter-turn about the origin, α^2 the half-turn about the origin (i.e., α performed twice), and α^3 the clockwise quarter turn about the origin (α performed three times). Let $A = \{e, \alpha, \alpha^2, \alpha^3\}$.

- (a) Find formulas for each element of A , and check that your formulas are geometrically sensible. (For example, $e(x, y) = (x, y)$.)
- (b) Show A is closed under mapping composition, and write out the Cayley table.
- (c) Is composition commutative on A ? Does composition have an identity element? If so, which elements of A have inverses?
- (d) Explain how your work in this exercise shows the binary operations of Exercise 6.6 are associative.

Exercise 6.12. Consider an equilateral triangle with vertices labeled 0, 1, 2 (below, left). Let α be a counterclockwise rotation by one-third of a turn about the center (middle), and β the reflection about the horizontal axis (right).



- (a) Sketch the six possible configurations of vertex labels on the triangle. (Three of them are shown.)
- (b) For each of the following compositions of maps, determine which of your sketches gives the corresponding vertex configuration: α^2 , α^3 , β^2 , $\alpha\beta$, $\alpha^2\beta$, $\beta\alpha$, $\beta\alpha^2$. (Note: Function composition is read right to left. For example, $\alpha\beta$ means “first apply β , then α ”.)
- (c) Let $A = \{e, \alpha, \alpha^2, \beta, \alpha\beta, \alpha^2\beta\}$. Write out the Cayley table for mapping composition on A .
Hints: Find a formula $\beta\alpha = \alpha^k\beta^\ell$ for suitable exponents k and ℓ . To simplify an arbitrary product of α s and β s, use this formula to move all the factors of α to the left.
- (d) Is composition commutative on A ? Does composition have an identity element? If so, which elements of A have inverses?
- (e) Find all proper, non-empty subsets of A that are closed under composition.
Hints: There are five of them. If $A' \subseteq A$ is closed under composition, and if $\alpha \in A'$ and $\beta \in A'$, then $A' = A$. (Why?)

The next two questions concern composition of certain functions. For notational brevity, write f^2 instead of ff , f^3 instead of fff , etc.

Exercise 6.13. In each part, $f(x) = \frac{1+x}{1-x}$ for $x \neq -1, 0, 1$.

- (a) Compute the compositions f^2 , f^3 , f^4 and f^5 . On the basis of your findings, what are f^{10} and f^{100} ?
- (b) Let S denote the set of distinct functions found in part (a). List the elements of S , show S is closed under composition, and make a Cayley table for S .
- (c) Show composition is a commutative operation on S , there is an identity element, and every element of S has an inverse in S .

Exercise 6.14. In each part, $f(x) = \frac{1}{1-x}$ and $g(x) = \frac{1}{x}$ for $x \neq 0, 1$.

- (a) Compute the compositions f^2 , f^3 , and f^4 . On the basis of your findings, what are f^{10} and f^{2010} ?

- (b) Compute the composition g^2 . Based on this finding, what is g^{1729} ?
- (c) Compute gf , gf^2 and fg , f^2g .
- (d) Let S denote the set of distinct functions found in parts (a)–(c). List the elements of S (there are six), and show each element of S has the form $f^k g^\ell$ for some integers k and ℓ . Show the set S is closed under composition, and make a Cayley table for S .
Hints: Show gf can be written in the stated form. Then argue that in any composition of f and g (in arbitrary order), the “factors” of f can be gathered on the left.
- (e) Let $h(x)$ be an arbitrary rational function obtainable by repeatedly composing f and/or g in arbitrary order. Use part (d) to show $h \in S$. (If you can see how, set up a formal argument using mathematical induction.)
- (f) Is composition a commutative operation on S ? Is there an identity element? Does every element of S have an inverse in S ?

Chapter 7

Groups

Many facts about algebra are not special to sets of numbers, but are instead consequences of properties of binary operations. The concept of a “group” gives us an abstract mathematical framework vastly generalizing the set of integers under addition.

Definition 7.1. Let G be a non-empty set and μ a binary operation on G . The pair (G, μ) is a *group* if

- (i) The binary operation μ is associative,
- (ii) There exists an identity element e for μ ,
- (iii) Every element of G has an inverse with respect to μ .

If μ is commutative, the group (G, μ) is *Abelian*. The *order* of (G, μ) is the number of elements in G .

Remark 7.2. When mathematicians work with a generic group, the operation is usually suppressed for convenience: One speaks of “the group G ”, and writes ab or $a \cdot b$ for the result of combining a and b under the operation of G . The element ab is generically termed the “product” of a and b , even though the group operation might be, for example, addition of integers.

In an effort to instill good habits, this book emphasizes the role of the operation. A group is *not* merely a set, but a set *together with a binary operation*: (G, \cdot) .

Example 7.3. The pair $(\mathbf{Z}, +)$ is an Abelian group of infinite order, the *additive group of integers*. This assertion restates the content of Axioms A1.–A4. for the integers, page 34.

Example 7.4. The pair (\mathbf{Z}, \cdot) , the set of integers under multiplication, is *not* a group; there is a unique identity element 1, but only integers having a multiplicative inverse are 1 and -1 , see Theorem 8.3.

Example 7.5. The set $G = \{-1, 1\} \subseteq \mathbf{Z}$ equipped with the operation of multiplication is an Abelian group of order 2. Any product formed from elements of G is an element of G :

\cdot	1	-1
1	1	-1
-1	-1	1

Since multiplication is both associative and commutative on \mathbf{Z} , *a fortiori* it enjoys these properties on G . Property (ii) is satisfied because the element 1 in G is the identity element for \cdot on all of \mathbf{Z} , and each element of G is its own inverse, so (iii) holds.

Example 7.6. The pair $(\mathbf{R}^\times, \cdot)$, comprising non-zero real numbers under ordinary multiplication, is an Abelian group of infinite order. The inverse of an element a is its reciprocal $a^{-1} = 1/a$. Indeed, the “exponential” notation for inverses in a general group comes from groups such as the multiplicative group of non-zero reals.

Theorem 7.7. Let (G, \cdot) be a group.

- (i) *The identity element is unique.*
- (ii) *Each element has a unique inverse.*
- (iii) *If $ab_1 = ab_2$, then $b_1 = b_2$.*
- (iv) *If $b_1a = b_2a$, then $b_1 = b_2$.*

Properties (iii) and (iv) are called the (left- and right-) *cancellation laws* in a group.

Proof. (i) If e and e' are identity elements for \cdot , then $e = ee' = e'$.

(ii) If $ab = ba = e$ and $ab' = b'a = e$, then

$$b = be = b(ab') = (ba)b' = eb' = b'.$$

(iii) If $ab_1 = ab_2$, then

$$b_1 = eb_1 = (a^{-1}a)b_1 = a^{-1}(ab_1) = a^{-1}(ab_2) = (a^{-1}a)b_2 = eb_2 = b_2.$$

The proof of (iv) is entirely similar, and left to you. \square

Remark 7.8. To prove $a^{-1} = b$ in a general group, it suffices by Theorem 7.7 (ii) to show $ab = ba = e$.

Example 7.9. Let (G, \cdot) be a group.

The identity element is its own inverse, since $ee = e$.

The inverse of a^{-1} is a , or $(a^{-1})^{-1} = a$, since $aa^{-1} = a^{-1}a = e$. In other words, a is the unique element b such that $a^{-1}b = e$.

Remark 7.10. Our proof of the cancellation law used the existence of inverses. However, the cancellation law can also hold in a non-group. For example, $(\mathbf{Z}^\times, \cdot)$, the set of non-zero integers under multiplication, satisfies the cancellation law despite the lack of multiplicative inverses. Theorem 3.12 (iii) asserts that if $a \neq 0$, and if b_1 and b_2 are integers such that $ab_1 = ab_2$, then $b_1 = b_2$.

Direct Products

Definition 7.11. Let (G_1, \cdot) and $(G_2, *)$ be groups. Their *direct product* is the group $(G_1, \cdot) \times (G_2, *)$ formed from the Cartesian product $G_1 \times G_2$ and the “componentwise” binary operation

$$(a_1, a_2)(b_1, b_2) = (a_1 \cdot b_1, a_2 * b_2).$$

The identity element is $e = (e_1, e_2)$, where e_i in G_i are the respective identity elements, and inverse of (a_1, a_2) is (a_1^{-1}, a_2^{-1}) , where the inverses are taken in the respective factors.

Example 7.12. Let $(\mathbf{R}, +)$ be the additive group of real numbers. The direct product $(\mathbf{R}, +) \times (\mathbf{R}, +) = (\mathbf{R} \times \mathbf{R}, +) = (\mathbf{R}^2, +)$ is the additive group of plane vectors, see Figure 7.1. The group operation is

$$(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2),$$

the identity element is $e = (0, 0)$, and the inverse of (x_1, x_2) is $(-x_1, -x_2)$.

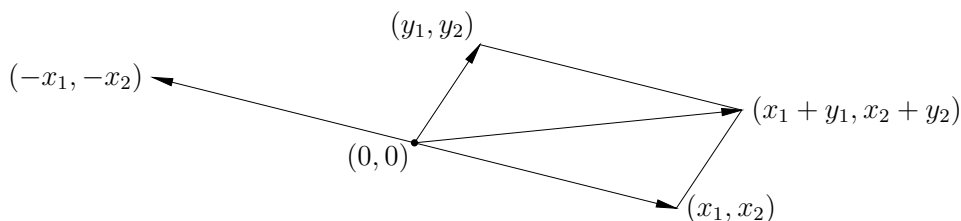


Figure 7.1: The additive group of plane vectors.

7.1 The Law of Exponents

Definition 7.13. Let (G, \cdot) be a group written multiplicatively. If $a \in G$, we define (*integer*) *powers* of a by

$$a^0 = e, \quad a^{k+1} = a^k \cdot a = \mu(a^k, a), \text{ for } k \geq 0,$$

and

$$a^{-k} = (a^{-1})^k \text{ for } k > 0.$$

Theorem 7.14 (The law of exponents). *Let (G, \cdot) be a group. If $a \in G$, then*

$$a^{m+n} = a^m \cdot a^n \text{ for all } m \text{ and } n \text{ in } \mathbf{Z}.$$

Particularly, $(a^n)^{-1} = a^{-n} = (a^{-1})^n$ for all n in \mathbf{Z} .

Proof. Assume first that $n \geq 0$, and consider the statement

$$P(n) \quad a^{m+n} = a^m \cdot a^n \text{ for all } m \geq 0.$$

This *single statement* may be viewed as an infinite family of statements, one for each non-negative value of m , with $n \geq 0$ fixed.

The base case $P(0)$ asserts $a^{m+0} = a^m \cdot a^0$ for all $m \geq 0$, which is true since $m + 0 = m$ for all m and $a^0 = e$. Next, assume inductively that $P(k)$ is true for some k , namely that

$$a^{m+k} = a^m \cdot a^k \text{ for all } m \geq 0.$$

For all $m \geq 0$, we have

$$\begin{aligned} a^{m+k+1} &= a^{m+k} \cdot a && \text{Definition of exponentiation} \\ &= (a^m \cdot a^k) \cdot a && \text{Inductive hypothesis} \\ &= a^m \cdot (a^k \cdot a) && \text{Associativity} \\ &= a^m \cdot a^{k+1} && \text{Definition of exponentiation} \end{aligned}$$

which establishes the inductive step. By the principle of mathematical induction, $a^{m+n} = a^m \cdot a^n$ for all non-negative m and n .

If m and n are both non-positive, the conclusion of the theorem now follows immediately by replacing a with a^{-1} .

It remains to check the case where one exponent is positive, the other negative. Without loss of generality, $m < 0 < n$. Suppose first that $m + n \geq 0$. Since $-m > 0$, the preceding argument shows

$$a^n = a^{(m+n)+(-m)} = a^{m+n} \cdot a^{-m},$$

and multiplying both sides by a^m gives $a^{m+n} = a^m \cdot a^n$, as claimed. The case $m + n < 0$ is entirely similar. \square

Example 7.15. In the additive group $(\mathbf{Z}, +)$, exponentiation (repeatedly combining a with itself k times) becomes ordinary integer multiplication, and we write ka instead of a^k . Theorem 7.14 says that for each a in \mathbf{Z} ,

$$(m + n)a = ma + na \quad \text{for all } m \text{ and } n \text{ in } \mathbf{Z}.$$

This is the ordinary distributive law, Axiom M4., page 34.

Example 7.16. Let (G, \cdot) be a group written multiplicatively. If $a \in G$, the identity

$$(a^m)^n = a^{mn} \quad \text{for all } m \text{ and } n \text{ in } \mathbf{Z}$$

holds. A proof of this property can be given in parallel to the proof of Theorem 7.14, see Exercise 7.22.

Example 7.17. Every child learns the identity $(ab)^2 = a^2b^2$ for real numbers a and b . It therefore comes as a rude shock to find this formula *fails* in general. Carefully expanding the definition of “squaring” reveals that

$$(ab)^2 = (ab)(ab) = abab, \quad \text{while} \quad a^2b^2 = aabb.$$

However, if $abab = aabb$, then

$$ba = a^{-1}(abab)b^{-1} = a^{-1}(aabb)b^{-1} = ab.$$

The converse holds similarly, so $(ab)^2 = a^2b^2$ if and only if $ba = ab$, if and only if a and b commute as elements of (G, \cdot) . Said another way, “ $(ab)^2 = a^2b^2$ ” is an identity only in an *Abelian* group.

Definition 7.18. A group (G, \cdot) is *cyclic* if there exists an element a in G such that $G = \{a^n : n \in \mathbf{Z}\}$. Such an element a is called a *generator* of G .

Example 7.19. The group $(\mathbf{Z}, +)$ is infinite cyclic, and has exactly two generators, 1 and -1 .

Example 7.20. The multiplicative group $(\{1, -1\}, \cdot)$ of Example 7.5 is finite cyclic; -1 is the only generator.

Remark 7.21. Every cyclic group (G, \cdot) is Abelian, since every element of G is a power of some fixed element a , and powers of a commute by the law of exponents: $a^n \cdot a^m = a^{n+m} = a^{m+n} = a^m \cdot a^n$.

Contrapositively, a non-Abelian group is not cyclic.

7.2 Subgroups

Definition 7.22. Let (G, \cdot) be a group and $H \subseteq G$ a non-empty subset. We say H is a *subgroup* of (G, \cdot) if the binary operation of G induces a binary operation on H , with respect to which the pair (H, \cdot) is itself a group.

Remark 7.23. To say “the operation of G induces a binary operation on H ” means that H is closed under the operation of G , see Definition 6.12. To say the pair (H, \cdot) is a group means in addition that H contains an identity element, and contains the inverse of each of its elements, see Theorem 7.32 below.

Example 7.24. Let $(G, \cdot) = (\mathbf{Z}, +)$ be the additive group of integers. The set $E = 2\mathbf{Z}$ of even integers is a subgroup: A sum of even integers is even, the identity element of G is even (hence is an element of E), and the inverse (in G) of an even integer is an even integer (E contains the inverse of each of its elements). (Compare Theorem 7.32 below.)

Example 7.25. The set $H = \mathbf{Q}$ of rational numbers is a subgroup of $(\mathbf{R}, +)$, the additive group of reals: A sum of rational numbers is rational, the identity element 0 of $(\mathbf{R}, +)$ is rational, and the additive inverse of a rational number is rational.

Example 7.26. The set $2\mathbf{Z} + 1$ of odd integers is not a subgroup of $(\mathbf{Z}, +)$, since a sum of odd integers is not odd.

Example 7.27. The set \mathbf{Z}^+ of positive integers is closed under addition, but is not a subgroup of $(\mathbf{Z}, +)$ since \mathbf{Z}^+ contains no identity element.

Example 7.28. The set \mathbf{N} of non-negative integers is closed under addition and contains an identity element, but is not a subgroup of $(\mathbf{Z}, +)$ since not every element of \mathbf{N} has an additive inverse in \mathbf{N} .

Example 7.29. Let $G = \{0, 1, 2, 3, 4, 5\}$. Define $a \cdot b$ by multiplying a and b as ordinary integers and taking the remainder on division by 6.

For example, $2 \cdot 4 = 2$, $2 \cdot 5 = 4$, $3 \cdot 2 = 0$, and $5 \cdot 5 = 1$.

The pair (G, \cdot) has an identity element $e = 1$, but is not a group since not every element has an inverse. (Which elements *do* have inverses?)

Let $H = \{2, 4\} \subseteq G$. Surprisingly, (H, \cdot) is a group, with identity element 4. Note carefully, however, that H is *not* a subgroup of (G, \cdot) , since (G, \cdot) is not itself a group.

To verify that (H, \cdot) is a group, it's convenient first to construct the Cayley table:

\cdot	2	4
2	4	2
4	2	4

(Closure). The set H is closed under \cdot because each product in the Cayley table is an element of H .

(Identity element). The element $e = 4$ is the identity element for \cdot , since $2 \cdot 4 = 4 \cdot 2 = 2$ and $4 \cdot 4 = 4$. (More abstractly, $4 \cdot a = a \cdot 4 = a$ for $a = 2$ and $a = 4$; see also Example 6.26.)

(Inverses). Each element of H is its own inverse, since $2 \cdot 2 = 4$ and $4 \cdot 4 = 4$.

Note carefully that in (G, \cdot) , neither 2 nor 4 is invertible.

Remark 7.30. Let (G, \cdot) be a group, $H \subseteq G$ non-empty. Example 7.29 raises the possibilities that the identity element in H might not be the identity element of G , and the notion of inversion in H might differ from inversion in G . The next result guarantees that these concerns are unfounded.

Proposition 7.31. *Let (G, \cdot) be a group with identity element e , and $H \subseteq G$ a subgroup.*

- (i) *The identity element of H is the identity element of G : $e \in H$.*
- (ii) *If $a \in H$, then the inverse of a in H coincides with the inverse of a in G .*

Proof. (i) By hypothesis, H has an identity element e' , and $e'e' = e'$. Now, this may be viewed as an equation in G . However, in G , the equation $ee' = e'$ also holds. By cancellation in G , $e = e'$.

(ii) Let b denote the inverse of a viewed as an element of H , and c the inverse of a viewed as an element of G . Since the identity element of H is the identity element of G by (i), $ab = e = ac$. By cancellation in G , $b = c$. \square

Recall that if $a \cdot b \in H$ for all a and b in H , then H is said to be *closed under \cdot* or *closed in (G, \cdot)* . Analogously, if $a^{-1} \in H$ for all a in H , then H is said to be *closed under inversion in (G, \cdot)* .

Theorem 7.32. *Let (G, \cdot) be a group and $H \subseteq G$ a non-empty subset. The following are equivalent.*

- (i) *For all x and y in H , $x^{-1}y \in H$.*
- (ii) *H is closed under \cdot , and closed under inversion in (G, \cdot) .*
- (iii) *H is a subgroup of (G, \cdot) .*

Proof. By hypothesis, H is non-empty; throughout, let a be an arbitrary element of H .

((i) implies (ii)). Taking $x = y = a$ in Condition (i), we have $e = a^{-1}a \in H$. Thus H contains the identity element of G .

Now, condition (i) with $x = a$ and $y = e$ implies $a^{-1} = a^{-1}e \in H$, which means H is closed under inversion.

Finally, let a and b be elements of H . Since H is closed under inversion, $a^{-1} \in H$. Condition (i) with $x = a^{-1}$ and $y = b$ implies $ab = x^{-1}y \in H$, so H is closed under the operation of G .

((ii) implies (iii)). Since H is closed under the operation of G , the operation of G induces a binary operation on H . Assume $a \in H$. Since H is closed under inversion, $a^{-1} \in H$; since H is closed under the operation of G , $e = a^{-1}a \in H$. Thus, H is a non-empty set equipped with a binary operation having an identity element and inverses. This means H is a subgroup of (G, \cdot) .

((iii) implies (i)). Let x and y be elements of H . Since H is a group under the operation of G , $x^{-1} \in H$. Since H is closed under the operation of G , $x^{-1}y \in H$. This establishes (i). \square

Corollary 7.33. *Let (G, \cdot) be a group, and $\{H_\alpha\}_{\alpha \in I}$ a family of subgroups of (G, \cdot) . The intersection $H = \cap_{\alpha} H_\alpha$ is a subgroup of (G, \cdot) .*

Proof. Since $e \in H_\alpha$ for every α , $e \in H$; in particular, $H \neq \emptyset$. To complete the proof, it suffices to establish condition (i) of the theorem. Let x and y be arbitrary elements of H . By the definition of intersection, x and y are elements of H_α for every α . Since H_α is a subgroup of (G, \cdot) , part (i) of the theorem implies $x^{-1}y \in H_\alpha$. Since α is arbitrary, $x^{-1}y \in \cap_{\alpha} H_\alpha = H$. This means H itself satisfies condition (i), so H is a subgroup of (G, \cdot) . \square

Generated Subgroups

Let (G, \cdot) be a group and $S \subseteq G$. In various guises, the question arises: What is the “smallest” subgroup of G containing S ? That is, if H is a subgroup of G and $S \subseteq H$, which elements of G must necessarily be in H ? The preceding corollary guarantees that this informal question has a well-defined mathematical answer.

Proposition 7.34. *Let (G, \cdot) be a group and $S \subseteq G$. The intersection $\langle S \rangle$ of all subgroups of G containing S is a subgroup of (G, \cdot) .*

If $H \subseteq G$ is a subgroup and $S \subseteq H$, then $\langle S \rangle \subseteq H$.

Proof. Let $\{H_\alpha\}$ be the family of all subgroups of G containing S . The intersection $\langle S \rangle = \cap_\alpha H_\alpha$ is a subgroup of (G, \cdot) by Corollary 7.33.

If $S \subseteq H$ for some subgroup H , then H is one of the H_α by definition, so obviously $\langle S \rangle = \cap_\alpha H_\alpha \subseteq H$. \square

Definition 7.35. Let (G, \cdot) be a group and $S \subseteq G$. The subgroup $\langle S \rangle$ is called the subgroup of G *generated by S* .

Remark 7.36. If $S \subseteq S'$, then $\langle S \rangle \subseteq \langle S' \rangle$. See also Exercise 7.23.

Example 7.37. In an arbitrary group (G, \cdot) , we have $\langle \emptyset \rangle = \{e\}$.

The second-simplest possibility, that $S = \{a\}$ is a singleton, yields a basic tool for investigating the structure of an arbitrary group.

Theorem 7.38. *Let (G, \cdot) be a group, and assume $a \in G$.*

- (i) *The set $H = \{a^n : n \in \mathbf{Z}\}$ of powers of a is a subgroup of (G, \cdot) .*
- (ii) *$H = \langle a \rangle$, the subgroup generated by $\{a\}$.*

Definition 7.39. The subgroup $\langle a \rangle = \{a^n : n \in \mathbf{Z}\} \subseteq (G, \cdot)$ is called the *cyclic subgroup generated by a* .

Proof. (i) The set $H \subseteq G$ is non-empty, and closed under the operation of G by the law of exponents: $a^m \cdot a^n = a^{m+n}$ for all m and n in \mathbf{Z} . Moreover, H is closed under inversion: If $a^m \in H$, then $(a^m)^{-1} = a^{-m}$ is an element of H . By Theorem 7.32 (ii), H is a subgroup of (G, \cdot) .

(ii) Let $\langle a \rangle \subseteq G$ be the smallest subgroup of (G, \cdot) containing a , namely the intersection of all subgroups of (G, \cdot) that contain a .

Since H is a subgroup of (G, \cdot) , $\langle a \rangle \subseteq H$ by Proposition 7.34.

To prove $H \subseteq \langle a \rangle$, note that $\langle a \rangle$ is closed under the operation of G . Since $a \in \langle a \rangle$, induction on m shows $a^m \in \langle a \rangle$ for every non-negative integer m . Further, $\langle a \rangle$ is closed under inversion, so $a^{-m} = (a^m)^{-1}$ is an element of $\langle a \rangle$ for every positive integer m . That is, $H \subseteq \langle a \rangle$. \square

Definition 7.40. Let (G, \cdot) be a group. If $a \in G$, the *order* of a in G is the smallest positive integer n such that $a^n = e$, or infinity if $a^n \neq e$ for all $n > 0$.

Example 7.41. In an arbitrary group (G, \cdot) , the identity element e has order 1, and is the only such element.

Example 7.42. In $(\mathbf{Z}, +)$, the additive group of integers, every non-zero integer has infinite order.

Example 7.43. In $(\mathbf{C}^\times, \cdot)$, the multiplicative group of non-zero complex numbers, the element -1 has order 2 since $(-1)^2 = 1$. The elements $\pm i$ have order 4 since $(\pm i)^4 = 1$ but $(\pm i)^k \neq 1$ if $k = 1, 2, 3$.

Remark 7.44. The order of an element a in a group (G, \cdot) turns out to be the order of (i.e., number of elements in) the cyclic subgroup $\langle a \rangle$ generated by a . This fact is not needed immediately, and will be more natural to prove later, see Corollary 12.11.

We briefly consider subgroups generated by more than one element.

Example 7.45. Let (G, \cdot) be a group, a and b elements of G such that neither is a power of the other, and $S = \{a, b\}$. The subgroup of G generated by S is potentially vastly complicated.

Certainly, $(\langle a \rangle \cup \langle b \rangle) \subseteq \langle S \rangle$, so if either a or b generates an infinite cyclic group, then the group $\langle S \rangle = \langle a, b \rangle$ is infinite. However, $\langle S \rangle$ contains much more than $\langle a \rangle \cup \langle b \rangle$. The general element is a *word* in a and b , namely, a finite product of powers of a , b , and their inverses, such as

$$aba, \quad a^2b, \quad aba^{-1}, \quad aba^{-1}b^{-1}, \quad a^6ba^{-101}b^2a^{-3}b^{12}.$$

If $ba \neq ab$, such expressions cannot generally be simplified.

In an *Abelian* group, however, any word in a and b can be rearranged uniquely so the powers of a are on the left and the powers of b are on the right. In this case,

$$\langle a, b \rangle = \{a^n b^m : n, m \in \mathbf{Z}\},$$

and $(a^{n_1} b^{m_1}) \cdot (a^{n_2} b^{m_2}) = a^{n_1+n_2} b^{m_1+m_2}$ by the law of exponents.

Definition 7.46. Let (G, \cdot) be a group. A subgroup $H \subseteq G$ is *finitely generated* if there exists a finite set $S \subseteq G$ such that $H = \langle S \rangle$.

Example 7.47. The following are finitely generated subgroups of $(\mathbf{R}, +)$:

$$\begin{aligned}\langle \tfrac{1}{2} \rangle &= \tfrac{1}{2}\mathbf{Z} = \{ \tfrac{n}{2} : n \in \mathbf{Z} \}, \\ \langle a \rangle &= a\mathbf{Z} = \{ an : n \in \mathbf{Z} \}, \quad (a \in \mathbf{R} \text{ fixed}) \\ \langle 1, \sqrt{2} \rangle &= \mathbf{Z} + \sqrt{2}\mathbf{Z} = \{ n_1 + n_2\sqrt{2} : n_1, n_2 \in \mathbf{Z} \}.\end{aligned}$$

The first two subgroups are cyclic; the third is not cyclic.

Example 7.48. The following are finitely generated subgroups of the multiplicative group of non-zero reals, $(\mathbf{R}^\times, \cdot)$:

$$\begin{aligned}\langle -1 \rangle &= \{1, -1\}, \\ \langle 2 \rangle &= \{2^n : n \in \mathbf{Z}\} = \{1, 2, \tfrac{1}{2}, 4, \tfrac{1}{4}, 8, \tfrac{1}{8}, \dots\}, \\ \langle a \rangle &= \{a^n : n \in \mathbf{Z}\} = \{1, a, \tfrac{1}{a}, a^2, \tfrac{1}{a^2}, a^3, \tfrac{1}{a^3}, \dots\}, \quad (a \neq 0 \text{ fixed}) \\ \langle -2, \tfrac{1}{3} \rangle &= \{ \tfrac{(-2)^n}{3^m} : n, m \in \mathbf{Z} \}.\end{aligned}$$

Example 7.49. The sets \mathbf{Q} and $\sqrt{2}\mathbf{Q} = \{r\sqrt{2} : r \in \mathbf{Q}\}$ are subgroups of $(\mathbf{R}, +)$, the additive group of reals.

The sets $\mathbf{Q}^\times = \{x \text{ in } \mathbf{Q} : x \neq 0\}$ and $\mathbf{R}^+ = \{x \text{ in } \mathbf{R}^\times : x > 0\}$ are subgroups of $(\mathbf{R}^\times, \cdot)$, the multiplicative group of non-zero reals.

None of these groups is finitely generated, see for example Exercise 7.24. Contrapositively, if S is a finite subset of any of these groups, then $\langle S \rangle$, the subgroup generated by S , is not the entire group.

Subgroups of $(\mathbf{Z}, +)$

In this section we use the division algorithm, Theorem 3.15, to characterize the subgroups of $(\mathbf{Z}, +)$, the additive group of integers. This result is of fundamental technical importance.

Theorem 7.50. *Let H be a subgroup of $(\mathbf{Z}, +)$, and assume $H \neq \{0\}$, i.e., H contains some non-zero integer. There exists a smallest positive integer d in H , and $H = d\mathbf{Z}$ is the set of integer multiples of d .*

Proof. (H contains a smallest positive element). Under the hypotheses of the theorem, there exists a non-zero integer m in H . Since H is a subgroup of $(\mathbf{Z}, +)$, $-m$ is also an element of H . One of m and $-m$ is positive, so $H^+ = H \cap \mathbf{Z}^+$, the set of positive integers in H , is non-empty. By the well-ordering principle, there exists a d in H^+ such that $d \leq k$ for all k in H^+ .

($d\mathbf{Z} \subseteq H$). The cyclic subgroup generated by d , namely $\langle d \rangle = d\mathbf{Z}$, is a subgroup of H by Theorem 7.38.

($H \subseteq d\mathbf{Z}$). Let n be an arbitrary element of H . By the division algorithm, there exist unique integers q and r such that $n = dq + r$ and $0 \leq r < d$. Since H is a subgroup of $(\mathbf{Z}, +)$ and since n and dq are elements of H , $r = n - dq \in H$. But by hypothesis d is the smallest positive element of H , so $r = 0$. This means every element of H is an integer multiple of d . \square

7.3 Groups of Complex Numbers

Recall that the *complex line* is the Cartesian plane in which the point (a, b) is viewed as the *complex number* $\alpha = a + bi$, with $i^2 = -1$. The Cartesian coordinates of α are the *real part* a and the *imaginary part* b .

Geometrically, the *imaginary unit* i may be viewed concretely either as the *location* $(0, 1)$ or as the *operation* of rotating the complex line about the origin $(0, 0)$ by a quarter turn counterclockwise. This rotation sends (a, b) to the point $(-b, a)$. Rotating twice sends (a, b) to $(-a, -b)$, giving geometric meaning to the mysterious equation $i^2 = -1$.

Additive Subgroups

The pair $(\mathbf{C}, +)$ is the *additive group* of complex numbers. The identity element is $0 = 0 + i0$, and the additive inverse of a complex number $\alpha = a + bi$ is its negative, $-\alpha = (-a) + (-b)i$.

Example 7.51. Let α be a non-zero complex number. The cyclic group generated by α is the infinite set of integer multiples

$$\langle \alpha \rangle = \{n\alpha : n \in \mathbf{Z}\} = \{na + (nb)i : n \in \mathbf{Z}\}.$$

Geometrically, $\langle \alpha \rangle$ is the set of complex numbers reached by starting from the origin and taking “steps” of $\pm\alpha$. Qualitatively, $\langle \alpha \rangle$ looks like a copy of the integers, possibly tilted and/or stretched.

Matters become more interesting if $\alpha_1 = a_1 + b_1i$ and $\alpha_2 = a_2 + b_2i$ are non-zero complex numbers and we consider the additive group

$$\langle \alpha_1, \alpha_2 \rangle = \{n_1\alpha_1 + n_2\alpha_2 : n_1, n_2 \in \mathbf{Z}\}.$$

Example 7.52. Let $\alpha_1 = 1$ and $\alpha_2 = i$. The group

$$\langle \alpha_1, \alpha_2 \rangle = \langle 1, i \rangle = \{n_1 1 + in_2 : n_1, n_2 \in \mathbf{Z}\} = \mathbf{Z} + i\mathbf{Z}$$

is the group of *Gaussian integers*. Geometrically, elements of $\mathbf{Z} + i\mathbf{Z}$ lie on the “unit square lattice”, Figure 7.2 (left).

If we view the small shaded square as a tile, the translates of this tile cover the entire complex line, with distinct tiles overlapping at most along a common edge. Algebraically, each complex number differs from some element in the shaded square by a Gaussian integer.

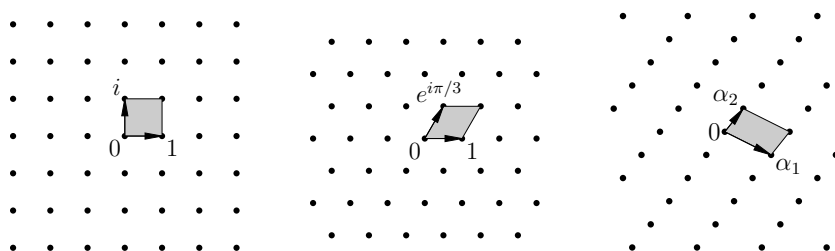


Figure 7.2: Subgroups of $(\mathbf{C}, +)$ generated by two elements.

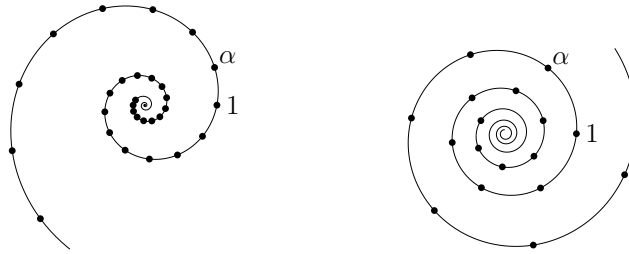
Example 7.53. Let $\alpha_1 = 1$ and $\alpha_2 = \frac{1}{2}(1 + i\sqrt{3}) = \cos \frac{\pi}{3} + i \sin \frac{\pi}{3}$. Elements of $G = \langle \alpha_1, \alpha_2 \rangle$ lie on the “honeycomb lattice”, Figure 7.2 (middle). As above, translates of the shaded parallelogram tile \mathbf{C} .

Example 7.54. Generally, if the ratio α_1/α_2 is non-real (i.e., the vectors α_1 and α_2 are non-parallel), then $\langle \alpha_1, \alpha_2 \rangle$ is by definition a *lattice*, Figure 7.2 (right).

Multiplicative Subgroups

The pair $(\mathbf{C}^\times, \cdot)$ is the *multiplicative group* of non-zero complex numbers. The identity element is $1 = 1 + i0$, and the inverse of a non-zero complex number $\alpha = a + bi$ is its reciprocal, $\alpha^{-1} = (a - bi)/(a^2 + b^2)$.

Recall that if $\alpha = re^{i\theta}$ is written in polar form, then $\alpha^n = r^n e^{in\theta}$ for every integer n . In particular, if $r \neq 1$, then distinct powers of α have different magnitude, and are therefore distinct. In this case, $\langle \alpha \rangle$ is an infinite cyclic group. The elements of $\langle \alpha \rangle$ lie on a curve known as a *logarithmic spiral*, Figure 7.3.

Figure 7.3: Infinite cyclic subgroups of $(\mathbf{C}^\times, \cdot)$.

Finite Subgroups of $(\mathbf{C}^\times, \cdot)$

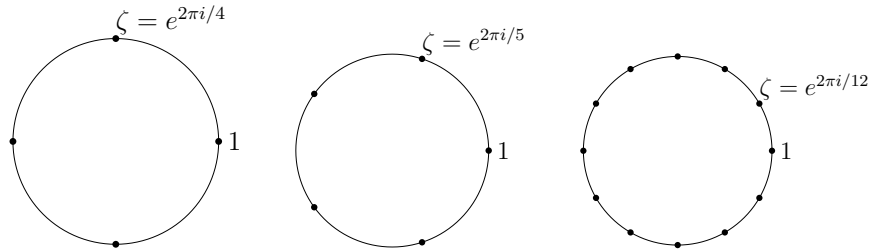
If $\alpha = e^{i\theta} = \cos \theta + i \sin \theta$ is a *unit complex number*, namely has magnitude 1, then every integer power of α is also a unit complex number, since $\alpha^n = e^{in\theta}$ for every integer n .

Now, $e^{i\phi} = \cos \phi + i \sin \phi = 1$ if and only if $\phi = 2\pi m$ for some integer m , so $\alpha^n = 1$ for some $n \neq 0$ if and only if $n\theta = 2\pi m$ for some integers m and n , if and only if $\theta = 2\pi(m/n)$ for some integers m and n , if and only if θ is a rational multiple of 2π .

Fix an integer $n > 1$, and consider the *root of unity* $\zeta_n = e^{2\pi i/n}$. The elements of $\langle \zeta_n \rangle$ are the complex numbers $1 = \zeta_n^0$, $\zeta_n = \cos \frac{2\pi}{n} + i \sin \frac{2\pi}{n}$, $\zeta_n^2 = \cos \frac{4\pi}{n} + i \sin \frac{4\pi}{n}$, \dots , namely

$$\{\zeta_n^k = \cos \frac{2\pi k}{n} + i \sin \frac{2\pi k}{n} : k = 0, \dots, n-1\}.$$

Geometrically, these points are the vertices of a regular n -gon inscribed in the unit circle, Figure 7.4.

Figure 7.4: Finite cyclic subgroups of $(\mathbf{C}^\times, \cdot)$.

Example 7.55. Let $n = 4$. By Euler's formula,

$$\zeta_4 = e^{2\pi i/4} = \cos \frac{\pi}{2} + i \sin \frac{\pi}{2} = 0 + i \cdot 1 = i,$$

$$\text{so } \langle \zeta_4 \rangle = \{\zeta_4^0, \zeta_4^1, \zeta_4^2, \zeta_4^3\} = \{1, i, -1, -i\}.$$

Exercises

Exercise 7.1. Let $G = \langle i \rangle = \{1, i, -1, -i\} \subseteq \mathbf{C}$, equipped with the operation of multiplication. (You may assume multiplication of complex numbers is associative.)

- (a) Write out the Cayley table for (G, \cdot) . Determine the identity element of (G, \cdot) , and find the inverse of each element.
- (b) Prove (G, \cdot) is cyclic, and find the generator(s).
Hint: Use the Cayley table to calculate the cyclic subgroup generated by each element of G .

Exercise 7.2. Suppose $a^4 = e$ in some group (G, \cdot) .

- (a) Prove $a^{4k} = e$ for all k in \mathbf{Z} . (Use induction to handle the case $0 \leq k$, and properties of exponents to handle the case $k < 0$.)
- (b) Prove $a^{-2} = a^2$.
- (c) Prove $a^{2011} = a^3 = a^{-1}$.

Exercise 7.3. Let (G, \cdot) be the multiplicative subgroup $O(1) = \{1, -1\}$.

- (a) Write out the four elements of the direct product $G \times G$ as ordered pairs. Labeling $G \times G = \{e, a, b, c\}$ (with e the identity element and the other three labels assigned to elements in any convenient manner), write out the Cayley table for $G \times G$. The group of order 4 defined by this table is called the *Klein 4-group*.
- (b) Prove $G \times G$ is Abelian, but not cyclic.
Hints: For the first assertion, you may find it easiest to prove that a direct product of Abelian groups is Abelian. For the second, use the Cayley table to calculate the elements of the cyclic subgroup generated by each element of $G \times G$.

Exercise 7.4. Let a and b be integers, and let

$$G = \langle a, b \rangle = \{ka + \ell b : k, \ell \in \mathbf{Z}\} \subseteq \mathbf{Z}.$$

- (a) Use Theorem 7.32 to prove G is a subgroup of $(\mathbf{Z}, +)$.
- (b) Prove G is infinite and contains a smallest positive element unless $a = b = 0$.

Exercise 7.5. Write out the three elements of the subgroup of $(\mathbf{C}^\times, \cdot)$ generated by the cube root of unity $\zeta_3 = e^{2\pi i/3}$. Use trigonometry to express these complex numbers in terms of radicals, and make a careful sketch. Make a Cayley table for this group. Show that ζ_3 and ζ_3^2 are inverses, giving as many arguments as you can find.

Exercise 7.6. As in the preceding exercise, write out the sixth roots of unity, namely the powers of $\zeta_6 = e^{2\pi i/6}$. Give your answer in both polar and Cartesian forms, and sketch these numbers carefully.

Exercise 7.7. Write out the eighth roots of unity, namely the powers of $\zeta_8 = e^{2\pi i/8}$. Give your answer in both polar and Cartesian forms, and sketch these numbers carefully.

Exercise 7.8. For each pair α_1 and α_2 of complex numbers, sketch several elements of the additive subgroup $\langle \alpha_1, \alpha_2 \rangle \subseteq (\mathbf{C}, +)$, and find a parallelogram as in Figure 7.2. Determine which (if any) of these five groups are subgroups of another group in the list.

- (a) $\alpha_1 = 2, \alpha_2 = i$.
- (b) $\alpha_1 = 1, \alpha_2 = 2i$.
- (c) $\alpha_1 = 1, \alpha_2 = 1 + i$.
- (d) $\alpha_1 = 1, \alpha_2 = \frac{1}{2}(-1 + \sqrt{3}i)$.
- (e) $\alpha_1 = 1 + i, \alpha_2 = 1 - i$.

Exercise 7.9. Let $\zeta = e^{4\pi i/5}$. List (in exponential form, not as radicals) the elements of the multiplicative cyclic subgroup $\langle \zeta \rangle$ generated by ζ , and make a Cayley table for this group.

Exercise 7.10. Let $\zeta = e^{2\pi i/6}$, and let $G = \langle \zeta \rangle \subseteq (\mathbf{C}^\times, \cdot)$ as in Exercise 7.6 above. For each element a of G , list the elements of the cyclic subgroup $\langle a \rangle$. Which elements are generators? When two elements of G generate the same cyclic subgroup, how are these elements related?

Exercise 7.11. In each part, a subgroup H of $(\mathbf{Z}, +)$ is given. Find the positive integer d such that $H = d\mathbf{Z}$.

- (a) $H = \langle 0, 3 \rangle$.
- (b) $H = \langle 2, 3 \rangle$.
- (c) $H = \langle 4, -6 \rangle$.

Exercise 7.12. In each part, a subgroup H of $(\mathbf{Z}, +)$ is given. Find the positive integer d such that $H = d\mathbf{Z}$.

- (a) $H = \langle 0, 6, 9 \rangle$. (b) $H = \langle -8, 12, 16 \rangle$.

Exercise 7.13. Consider the set S of functions introduced in Exercise 6.13.

- (a) Show S is an Abelian group under function composition.
 (b) By constructing an explicit bijection between sets and showing the Cayley tables correspond, show (S, \cdot) is abstractly equivalent to the group G in Exercise 7.1.

Exercise 7.14. Consider the set S of functions introduced in Exercise 6.14.

- (a) Show S is a non-Abelian group under function composition.
 (b) Find four distinct, proper, non-trivial subgroups of (S, \cdot) .
 Hint: For each non-identity element a in S , consider the cyclic subgroup generated by a .

Exercise 7.15. On the open interval $(-1, 1)$ of real numbers, define $a \oplus b = (a + b)/(1 + ab)$.

- (a) Prove \oplus is a binary operation on $(-1, 1)$. That is, if $-1 < a, b < 1$, then $-1 < a \oplus b < 1$.
 Hints: If $-1 < a, b < 1$, then $0 < 1 + ab$, $0 < (1 + a)(1 + b)$, and $0 < (1 - a)(1 - b)$. (Why?)
 (b) Prove \oplus is associative and commutative.
 (c) Prove \oplus has an identity element.
 (d) Prove every element of $(-1, 1)$ has an inverse with respect to \oplus .

Exercise 7.16. Let a and b be elements in some group (G, \cdot) .

- (a) Prove $(ab)^{-1} = b^{-1}a^{-1}$. (The *reverse order law* in a group.)
 (b) Prove $(ab)^{-1} = a^{-1}b^{-1}$ if and only if a and b commute.

Exercise 7.17. Let x and b be elements of a group (G, \cdot) . Prove $b^{-1}xb = e$ if and only if $x = e$.

Exercise 7.18. Suppose a and b are elements in some group (G, \cdot) .

- (a) Prove $(b^{-1}ab)^n = b^{-1}a^nb$ for all $n \geq 0$.
- (b) Prove $(b^{-1}ab)^{-1} = b^{-1}a^{-1}b$.
- (c) Conclude $(b^{-1}ab)^n = b^{-1}a^nb$ for all n in \mathbf{Z} .

Exercise 7.19. Let a and b be elements of a group. Prove that ba and ab have the same order (see Definition 7.40.), finite or infinite.

Hint: Use the results of the two preceding exercises.

Exercise 7.20. Let (G, \cdot) be a group, and assume every non-identity element of G has order 2. Prove (G, \cdot) is Abelian.

Hint: Use Example 7.17.

Exercise 7.21. Let (G, \cdot) be a group, and let a and b be *commuting* elements, i.e., such that $ba = ab$. Give a formal proof (using mathematical induction as necessary) that $(ab)^n = a^n b^n$ for all integers n . (Compare Example 7.17.)

Exercise 7.22. Let a be an element of a group. Prove

$$(a^m)^n = a^{mn} \quad \text{for all } m \text{ and } n \text{ in } \mathbf{Z},$$

see Example 7.16.

Exercise 7.23. Let (G, \cdot) be a group, $S \subseteq G$. Prove that $\langle\langle S \rangle\rangle = \langle S \rangle$; that is, the subgroup of (G, \cdot) generated by $\langle S \rangle$ is equal to the subgroup generated by S itself.

Suggestion: Prove that if H is a subgroup of (G, \cdot) , then $\langle H \rangle = H$.

Exercise 7.24. Prove the additive group $(\mathbf{Q}, +)$ is not finitely generated.

Suggestion: Show that if $S = \{r_1, \dots, r_n\}$ is a finite set of rational numbers, then

$$\langle S \rangle = \{m_1 r_1 + \dots + m_n r_n : m_i \in \mathbf{Z}\} \neq \mathbf{Q}$$

by finding an upper bound on the denominators of elements of $\langle S \rangle$. If your instructor requires a formal proof, you'll need to do induction on n , the number of elements of S .

Chapter 8

Divisibility and Congruences

This chapter and the next two are loosely organized around division in the integers. Along the way, we construct and study new “number systems” built from equivalence classes of integers.

Definition 8.1. Let a and b be integers. We say a *divides* b , and write $a \mid b$, if there exists an integer q such that $b = aq$. In this situation, we also say a is a *divisor* or a *factor* of b , or that b is a *multiple* of a .

Remark 8.2. Since $a(-q) = (-a)q = -(aq)$, the following are equivalent for all a and b : $a \mid b$, $-a \mid b$, and $a \mid -b$.

If $a > 0$, then by the division algorithm, $b = aq + r$ for unique integers q and r with $0 \leq r < a$. Consequently, if $a > 0$ then $a \mid b$ if and only if $r = 0$, if and only if there are no “leftovers” when b objects are divided into a piles.

Clearly, $a \mid 0$ and $-a \mid 0$ are true for every a , by taking $q = 0$. The reverse relation is much more stringent: If $0 \mid a$, then $a = 0$.

Similarly, the statements $1 \mid b$, $-1 \mid b$ are true for every integer b , by taking $q = b$ or $q = -b$ respectively. In words, 1 and -1 divide everything. The converse relation is interesting enough to state formally.

Theorem 8.3. If $a \in \mathbf{Z}$ and $a \mid 1$, then $a = \pm 1$.

Proof. We will prove the contrapositive: If $a \neq \pm 1$, then $a \nmid 1$. As noted above, $0 \nmid 1$, and $a \mid 1$ if and only if $-a \mid 1$, so it suffices to consider the case $1 < a$.

If $a \mid 1$, there exists an integer q such that $aq = 1$. However, if $1 < a$, then multiplying by q would give $0 < q < aq = 1$, which is false; there is no integer between 0 and 1. It follows that if $1 < a$, then $a \nmid 1$. \square

8.1 Residue Classes of Integers

Definition 8.4. Let $n > 0$ be a positive integer. Two integers a and b are *congruent mod n* if $n \mid (b - a)$. This relation is denoted $a \equiv b \pmod{n}$, or simply $a \equiv b$ if n is fixed throughout a discussion.

Theorem 8.5. Let $n > 0$ be an integer, and define a relation R on \mathbf{Z} by aRb if and only if $a \equiv b \pmod{n}$. The relation R is an equivalence relation.

Proof. See Exercise 8.6. □

Example 8.6. Equivalence classes mod n and their arithmetic are implicitly familiar. Since there are 12 hours in one turn of a clock, time-keeping works mod 12. At an early age, you learned that six hours after 7 o'clock is 1 o'clock, or that five hours before 3 o'clock is 10 o'clock. For military time, you'd work mod 24 instead, but the idea is the same.

Example 8.7. Days of the week are reckoned mod 7. The labels are not integers, of course, but names: {Sunday, Monday, ..., Saturday}.

Example 8.8. Angular measurements in degrees are made mod 360, because there are 360 degrees in one full turn of a circle. Angles of 270, 630, and -90 degrees represent the same geometric quantity.

Example 8.9. Moving from the sublime to the ridiculous, a cartoon character (usually the cat in a cat-and-mouse conflict) will sometimes acquire amnesia when given a sharp blow on the head. The cure, as everyone knows, is a second blow. The cat's state of mental health (amnesiac or cured) represents the number of cranial blows mod 2 the cat has received. The concept is so simple even young children understand it perfectly.

Theorem 8.10. Let $n > 0$ be an integer, and a and b arbitrary integers. Use the division algorithm to write $a = nq_1 + r_1$ and $b = nq_2 + r_2$ with $0 \leq r_1, r_2 < n$. The integers a and b are congruent mod n if and only if $r_1 = r_2$.

Proof. See Exercise 8.7. □

In words, two integers are congruent mod n if and only if they leave the same remainder on division by n .

Corollary 8.11. The relation “congruence mod n ” has precisely n equivalence classes: $[0], [1], \dots, [n - 1]$.

Proof. See Exercise 8.7. \square

Equivalence classes mod n may be visualized in at least two useful ways. The first is the “clock” picture of the set \mathbf{Z}_n of equivalence classes mod n , which may be drawn as a set of n equally-spaced points on a circle. The case $n = 12$, Figure 8.1 is essentially an ordinary analog clock, though by convention we place $[0]$ at the rightmost position and label counterclockwise, ending with the class $[n - 1]$ one space clockwise from $[0]$. This picture emphasizes the “cyclical” nature of residue classes. Adding 1 corresponds to traveling counterclockwise by one space. Adding n travels one full revolution, returning to the same residue class.

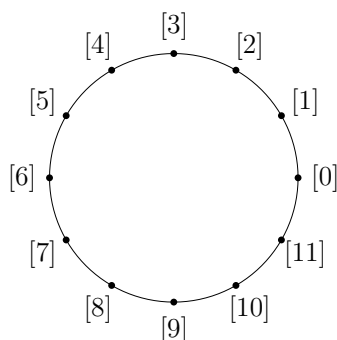


Figure 8.1: The set \mathbf{Z}_{12} of residue classes mod 12.

The second picture is the “unwrapping” of the clock onto a number line. For this, choose n distinct symbols, such as $[0], [1], \dots, [n - 1]$, and use these to label integer points on a numbers line.

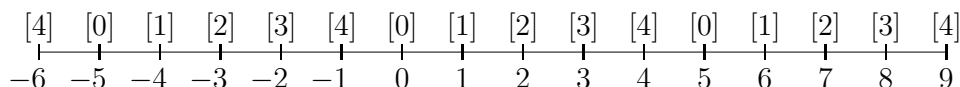


Figure 8.2: Residue classes mod 5 on the number line.

Modulo Arithmetic

As the examples above suggest, two residue classes mod n can be added in such a way that $[a] + [b] = [a + b]$. However, this equation is more subtle than it may first appear: a and b represent integers, so $a + b$ is also an integer, and we know every integer has a unique residue class

mod n . The potential snag is the *well-definedness* of the operation on the left-hand side. We might conceivably have $[a'] = [a]$ with $a' \neq a$ and $[b'] = [b]$ with $b' \neq b$. Before blithely assigning meaning to the expression “ $[a + b]$ ”, we must verify that if $[a'] = [a]$ and $[b'] = [b]$, then $[a' + b'] = [a + b]$.

Theorem 8.12. *Fix $n > 0$. Let a, a', b , and b' be integers such that $a \equiv a' \pmod{n}$ and $b \equiv b' \pmod{n}$, and let $[a] = [a']$ and $[b] = [b']$ denote their residue classes mod n . Then*

- (i) $a + b \equiv a' + b' \pmod{n}$, i.e., $[a + b] = [a' + b']$.
- (ii) $ab \equiv a'b' \pmod{n}$, i.e., $[ab] = [a'b']$.

The theorem is straightforward to prove directly, but breaking the proof into slightly smaller steps clarifies the main idea.

Lemma 8.13. *Let a, a' , and c be integers. If $a \equiv a' \pmod{n}$, then $a + c \equiv a' + c \pmod{n}$ and $ac \equiv a'c \pmod{n}$.*

Proof. By definition, $a \equiv a'$ if and only if $n \mid (a' - a)$.

Since $a' - a = (a' + c) - (a + c)$, we have $a \equiv a'$ if and only if $n \mid (a' - a) = (a' + c) - (a + c)$, if and only if $a + c \equiv a' + c$.

Further, if $n \mid (a' - a)$, then $n \mid (a' - a)c = a'c - ac$, i.e., $ac \equiv a'c$. \square

Proof of theorem. Let a, a', b, b' be integers such that $a \equiv a' \pmod{n}$ and $b \equiv b' \pmod{n}$. Then

$$\begin{aligned} a + b &\equiv a' + b \pmod{n} && \text{Lemma 8.13 with } c = b \\ &\equiv a' + b' \pmod{n} && \text{Lemma 8.13 with } c = a'. \end{aligned}$$

The proof for products is identical: $ab \equiv a'b \equiv a'b' \pmod{n}$. \square

Remark 8.14. If $a \not\equiv 0 \pmod{n}$ and $ab \equiv ac \pmod{n}$, it is *not* valid to deduce $b \equiv c$; the law of cancellation does not generally hold mod n . For example, $2 \cdot 3 \equiv 2 \cdot 0 \pmod{6}$, but even though $2 \not\equiv 0 \pmod{6}$, it is not true that $3 \equiv 0 \pmod{6}$.

Despite its seemingly-esoteric content, Theorem 8.12 has important practical consequences, two of which are explored in the next examples.

Example 8.15. To “solve” the congruence $67x \equiv 54 \pmod{5}$ is to find an integer x such that $67x$ leaves the same remainder as 54 on division by 5. We may replace 54 by any number congruent mod 5.

The obvious choice is 4, the remainder left by 54 on division by 5. By the second part of Theorem 8.12, we may replace 67 by any number congruent mod 5; the natural choice is 2. The original problem has therefore been recast in the much simpler form $2x \equiv 4 \pmod{5}$, which clearly has $x = 2$ as a solution.

Generally, when solving $ax \equiv b \pmod{n}$, the coefficients a and b may be reduced mod n and the resulting congruence is equivalent to the original.

Example 8.16. Consider the problem of simplifying $4^{2000} \pmod{63}$. The naive approach of calculating 4^{2000} (a number of over 1200 digits), then dividing by 63 and taking the remainder, is prohibitively complex. By the second part of Theorem 8.12, however, we may instead compute successive powers of 4, reducing mod 63 each time a running product exceeds 63.

Even this is prohibitive if carried out mechanically, but we can do better still. Indeed, $4^3 = 64 \equiv 1 \pmod{63}$, so $4^{3q} = (4^3)^q \equiv 1^q \equiv 1 \pmod{63}$ for all q . The preferred strategy, therefore, is to divide the original *exponent* 2000 by 3. Writing $2000 = 3 \cdot 666 + 2$, we have

$$4^{2000} = 4^{3 \cdot 666 + 2} = (4^3)^{666} \cdot 4^2 \equiv 4^2 \equiv 16 \pmod{63}.$$

The remainder in question is 16.

The same idea can be used to compute lengthy products mod n . Consider, for example,

$$10! = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \pmod{11}.$$

Judiciously gathering factors in pairs and reducing mod 11, $10 \equiv -1$, $9 \cdot 6 = 54 \equiv -1$, $8 \cdot 7 = 56 \equiv 1$, $5 \cdot 2 = 10 \equiv -1$, and $4 \cdot 3 = 12 \equiv 1$. Consequently,

$$10! \equiv (-1)(-1)(1)(-1)(1) = -1 \pmod{11}$$

by the second part of the theorem.

8.2 Greatest Common Divisors

Definition 8.17. Let a and b be integers. An integer c is a *common divisor* of a and b if $c \mid a$ and $c \mid b$.

Example 8.18. If $a = 12$ and $b = 18$, then the common divisors of a and b are $\pm 1, \pm 2, \pm 3$, and ± 6 .

If $a = b = 0$, then every integer is a common divisor of a and b . If at least one of a and b is non-zero, however, there is a common divisor d larger than every other common divisor.

Definition 8.19. Let a and b be integers, not both zero. An integer d is a *greatest common divisor* of a and b , denoted $\gcd(a, b)$, if

- (i) $d \mid a$ and $d \mid b$ (d is a common divisor of a and b).
- (ii) $0 < d$.
- (iii) If $c \mid a$ and $c \mid b$, then $c \mid d$ (every common divisor divides d).

Proposition 8.20. If a and b are not both zero, they have at most one \gcd .

Proof. Suppose d and d' both satisfy conditions (i)–(iii). Since d' divides both a and b , condition (iii) guarantees $d' \mid d$. Reversing the roles of d' and d shows $d \mid d'$.

Now, two positive integers, each dividing the other, must be equal: By hypothesis, there exist positive integers q_1 and q_2 such that $d = d'q_1$ and $d' = dq_2 = d'(q_1q_2)$, which implies $q_1q_2 = 1$. By Theorem 8.3, $q_1 = 1$, so $d = d'$. \square

The next result both guarantees the existence of the \gcd and provides a foundational characterization.

Theorem 8.21. Let a and b be integers, not both zero. The set

$$H = \langle a, b \rangle = \{ka + \ell b : k, \ell \in \mathbf{Z}\}$$

of integer linear combinations of a and b has a smallest positive element d , and $d = \gcd(a, b)$.

Proof. The set H is closed under addition and under inversion in $(\mathbf{Z}, +)$, so by Theorem 7.32 H is a subgroup of $(\mathbf{Z}, +)$. The integers a and b are not both zero, so $H \neq \{0\}$. By Theorem 7.50, H contains a smallest positive element d , and $H = d\mathbf{Z}$ is the set of multiples of d . It remains to show that the integer d satisfies conditions (i)–(iii) of Definition 8.19.

(d is a common divisor of a and b). The integers a and b are elements of H , so each is divisible by d .

Property (ii), $0 < d$, is immediate.

(Every common divisor of a and b divides d). If $c \mid a$ and $c \mid b$, there exist integers q_1 and q_2 such that $a = cq_1$ and $b = cq_2$. Substituting,

$$d = am + bn = cq_1m + cq_2n = c(q_1m + q_2n),$$

which implies $c \mid d$. \square

Remark 8.22. Changing the sign of a and/or b has no effect on H , so $\gcd(-a, b) = \gcd(a, -b) = \gcd(-a, -b) = \gcd(a, b)$. In practice, we may as well assume a and b are both non-negative.

Theorem 8.21 says that for each pair of integers a and b (not both zero), there exists a greatest common divisor $d = \gcd(a, b)$. Moreover, there exist integers m and n such that $\gcd(a, b) = am + bn$, and $\gcd(a, b)$ is the smallest positive integer that can be written in this form.

This “smallest positive linear combination” characterization leads to an efficient algorithm for computing $\gcd(a, b)$: Divide the smaller number into the larger, take the remainder (if non-zero), and repeat using the remainder and smaller divisor, stopping if the division at some stage has remainder zero. This process must terminate after finitely many steps, and the last non-zero remainder is the gcd. Let’s see how the algorithm works in practice before stating a formal theorem.

Example 8.23. Find $d = \gcd(68, 20)$, and write d as a linear combination of 68 and 20.

Repeated long division gives

$$\begin{aligned} 68 &= 3 \cdot 20 + 8, \\ 20 &= 2 \cdot 8 + 4, \\ 8 &= 2 \cdot 4 + 0. \end{aligned}$$

Thus $\gcd(68, 20) = 4$, the last non-zero remainder.

To write 4 in terms of 68 and 20, start with the second-to-last equation just found, and substitute backward up the chain:

$$\begin{aligned} 4 &= 20 - 2 \cdot 8 \\ &= 20 - 2 \cdot (68 - 3 \cdot 20) = 20 - 2 \cdot 68 + 6 \cdot 20 \\ &= 7 \cdot 20 - 2 \cdot 68. \end{aligned}$$

As a check, this reads $4 = 140 - 136$.

Theorem 8.24 (Euclid's algorithm). *Let $0 < b < a$ be integers, and recursively define sequences of quotients and remainders as follows:*

$$\begin{aligned} a &= bq_0 + r_1, & 0 \leq r_1 < b \\ b &= r_1q_1 + r_2, & 0 \leq r_2 < r_1 \\ r_1 &= r_2q_2 + r_3, & 0 \leq r_3 < r_2 \\ &\vdots \\ r_k &= r_{k+1}q_{k+1} + r_{k+2}, & 0 \leq r_{k+2} < r_{k+1}, \quad \text{etc.} \end{aligned}$$

If $r_n \neq 0$ and $r_{n+1} = 0$, then $r_n = \gcd(a, b)$.

Proof. The key fact is simple: If $a = bq + r$, then $\gcd(a, b) = \gcd(b, r)$. This is proven below. Applying this abstract relationship to the lines of the algorithm in turn, we have

$$\begin{aligned} \gcd(a, b) &= \gcd(b, r_1) = \gcd(r_1, r_2) \\ &= \vdots \\ &= \gcd(r_{n-1}, r_n) \\ &= \gcd(r_n, r_{n+1}) = \gcd(r_n, 0). \end{aligned}$$

But for every positive integer c , $\gcd(c, 0) = c$, so indeed $\gcd(a, b) = r_n$, as claimed.

To complete the proof, it suffices to show that if $a = bq + r$, then $\langle a, b \rangle = \langle b, r \rangle$ as subgroups of $(\mathbf{Z}, +)$. That is, every integer linear combination of a and b may be expressed as an integer linear combination of b and r , and *vice versa*.

(Inclusion $\langle a, b \rangle \subseteq \langle b, r \rangle$). The general element of $\langle a, b \rangle$ is $ak + b\ell$. If $a = bq + r$, then for arbitrary integers k and ℓ ,

$$ak + b\ell = (bq + r)k + b\ell = b(qk + \ell) + rk,$$

which is an element of $\langle b, r \rangle$.

(Inclusion $\langle b, r \rangle \subseteq \langle a, b \rangle$). The general element of $\langle b, r \rangle$ is $bm + rn$. Since $a = bq + r$, we have $r = a - bq$, so if m and n are arbitrary, then

$$bm + rn = bm + (a - bq)n = b(m - qn) + an,$$

which is in $\langle a, b \rangle$.

Since $\langle a, b \rangle = \langle b, r \rangle$, these sets have the same smallest positive element. Theorem 8.21 implies $\gcd(a, b) = \gcd(b, r)$. \square

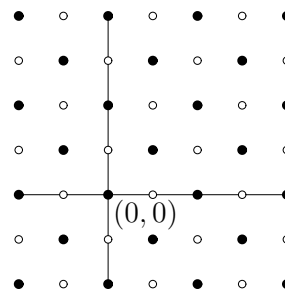
Exercises

Exercise 8.1. Calculate the following:

(a) $3^{487} \pmod{28}$. (b) $3^{120} \cdot 5^{531} \pmod{26}$. (c) $7^{97} \pmod{11}$.

Exercise 8.2. Let $\mathbf{Z}_7^\times = \{1, 2, 3, 4, 5, 6\}$ be the set of non-zero residue classes mod 7. Write out the Cayley table for $(\mathbf{Z}_7^\times, \cdot)$, prove $(\mathbf{Z}_7^\times, \cdot)$ is a cyclic group, and find the generators.

Exercise 8.3. Recall the geometric representation of the parity relation at right, in which the pair (m, n) is filled in if and only if $m \equiv n \pmod{2}$. Similarly, sketch the relations “congruent mod 3” and “congruent mod 5”. (For the latter you’ll need a larger grid than shown.)



Exercise 8.4. Let a and b be integers such that $a \mid b$. Prove $b \mid a$ if and only if $b = \pm a$.

Exercise 8.5. Prove that the “divides” relation is reflexive and transitive. (By Exercise 8.4, “divides” is not symmetric.)

Exercise 8.6. Prove Theorem 8.5.

Exercise 8.7. (a) Prove Theorem 8.10. (b) Prove Corollary 8.11.

Exercise 8.8. Let a and b be non-zero integers, and let $M_{ab} = \langle a \rangle \cap \langle b \rangle$ be the set of common multiples of a and b .

(a) Prove M_{ab} is a subgroup of $(\mathbf{Z}, +)$. The smallest positive element $\text{lcm}(a, b)$ of M_{ab} is called the *least common multiple* of a and b .

(b) Give conditions analogous to those in Definition 8.19, and prove $\text{lcm}(a, b)$ is the unique integer satisfying these conditions.

(c) Prove that $\text{gcd}(a, b) \text{lcm}(a, b) = ab$.

Suggestion: First show that if m and d are positive integers with $ab = md$, then d is a common divisor of a and b if and only if m is a common multiple.

Exercise 8.9. Fix $n > 1$, and let \mathbf{Z}_n be the set of residue classes mod n , equipped with the binary operation $+$, addition mod n , described in the text. Prove $(\mathbf{Z}_n, +)$ is a group. Your write-up should include a proof that $+$ is associative.

Exercise 8.10. Write out the Cayley table for the following cyclic groups, and use your tables to find all the generators.

- (a) $(\mathbf{Z}_4, +)$. (b) $(\mathbf{Z}_5, +)$. (c) $(\mathbf{Z}_8, +)$.

Exercise 8.11. Each part of this question concerns the direct product $(\mathbf{Z}_3 \times \mathbf{Z}_2, +)$, shown at right.

- (a) List the six elements of $\mathbf{Z}_3 \times \mathbf{Z}_2$.
- (b) Write out the Cayley table. Is this group cyclic? If so, which elements are generators?

Exercise 8.12. Write out the Cayley table for $(\mathbf{Z}_2 \times \mathbf{Z}_2, +)$. Is this group cyclic? If so, which elements are generators?

Exercise 8.13. Write out the Cayley table for $(\mathbf{Z}_4 \times \mathbf{Z}_2, +)$. Is this group cyclic? If so, which elements are generators?

Exercise 8.14. Represent the group $(\mathbf{Z}_4 \times \mathbf{Z}_3, +)$ as a rectangular array of points in the plane (compare Exercise 8.11), and use this picture to depict elements in the cyclic subgroup generated by $(1, 1)$.

Exercise 8.15. As in the preceding exercise, represent $(\mathbf{Z}_6 \times \mathbf{Z}_4, +)$ as a rectangular array, and use this picture to depict elements in the cyclic subgroup generated by $(1, 1)$.

Exercise 8.16. Repeat the preceding exercise for $(\mathbf{Z}_6 \times \mathbf{Z}_3, +)$.

Exercise 8.17. Let m and n be integers greater than 1, and consider the element $(1, 1)$ of the direct product $(\mathbf{Z}_m \times \mathbf{Z}_n, +)$.

- (a) Show $(r, r) = r \cdot (1, 1)$ is the identity element if and only if $m \mid r$ and $n \mid r$.
- (b) Show the cyclic subgroup generated by $(1, 1)$ contains $\text{lcm}(m, n)$ elements.
- (c) Show the direct product $(\mathbf{Z}_m \times \mathbf{Z}_n, +)$ is cyclic if $\text{gcd}(m, n) = 1$.

Exercise 8.18. List the eight elements of $(\mathbf{Z}_2 \times \mathbf{Z}_2 \times \mathbf{Z}_2, +)$, and show each non-identity element has order 2. Represent the elements of this group as vertices of a cube, and illustrate with a sketch.

Chapter 9

Primes

Every integer $a > 1$ has at least two positive divisors: 1 and a itself.

Definition 9.1. An integer $p > 1$ is *prime* if its *only* positive divisors are 1 and p . A non-prime integer $n > 1$ is *composite*.

Remark 9.2. The integer 1 is a *unit*, i.e., has a multiplicative inverse. In this chapter, invertibility conveys special status on 1, neither prime nor composite. Our goal is to factor composite integers uniquely into primes. If 1 were prime, uniqueness would be lost; if 1 were composite, existence of a factorization would be lost.

Example 9.3. The primes smaller than 20 are 2, 3, 5, 7, 11, 13, 17, and 19. Of these, only 2 is prime immediately from the definition: The *only* positive integers not exceeding 2 are 1 and 2, so 2 cannot have positive divisors other than 1 and 2!

Remark 9.4. It was known to Euclid around 300 BCE that there are infinitely many primes, a fact we prove below. At this writing, by contrast, it is unknown whether or not there exist infinitely many *twin primes*, pairs of primes differing by 2, such as 3 and 5 or 101 and 103.

In April 2013, Y. Zhang announced the existence of infinitely many pairs of primes differing by no more than 70 million, the first “bounded gap” result. By November 2013, J. Maynard had reduced the bound to 600. By April 2014, the PolyMath project had reduced the bound to 246, the best known bound at this writing (July 2015).

The primes are the “multiplicative building blocks” of the positive integers. This principle culminates in the “Fundamental Theorem of

Arithmetic”, Theorem 9.18 below. For now we are content to prove a technical result, later carried to its logical conclusion.

Theorem 9.5. *Let $N > 1$ be an integer. There exists a prime p such that $p \mid N$.*

Proof. The proof proceeds by mathematical induction on the following statement:

$P(N)$: For every integer n with $2 \leq n \leq N$, there exists a prime p (depending on n) such that $p \mid n$.

Informally, $P(5)$ says “Each of the integers 2, 3, 4, and 5 has a prime factor”.

The statement $P(2)$ is true; $p = 2$ is a divisor of $N = 2$. This establishes the base case.

Assume inductively that $P(k)$ is true for some $k > 1$, namely that every integer n with $2 \leq n \leq k$ has a prime factor.

The integer $k+1 > 1$ is either prime or composite. If $k+1$ is prime, then $p = k+1$ is a prime factor; together with $P(k)$, this implies every integer n with $2 \leq n \leq k+1$ has a prime factor, proving $P(k+1)$ in this case.

On the other hand, if $k+1$ is composite, there exist integers n and m , both greater than 1, such that $k+1 = nm$. It follows that $1 < n < k+1$, for if $k+1 \leq n$, then $(k+1)m \leq nm = k+1$, contrary to the inequality $1 < m$. Since $n < k+1$, we have $n \leq k$ by Theorem 3.12 (iv). By $P(k)$, there exists a prime p such that $p \mid n$. Now, $n \mid (k+1)$ and divisibility is transitive, so $p \mid (k+1)$ as well. Thus $k+1$ has a prime divisor, proving $P(k+1)$ in this case.

Since $P(2)$ is true and $P(k)$ implies $P(k+1)$ for each $k > 1$, $P(N)$ is true for all $N > 1$ by the principle of mathematical induction. \square

9.1 Coprimality

Definition 9.6. Integers a and b are *coprime* if $\gcd(a, b) = 1$.

By Theorem 8.21, a and b are coprime if and only if there exist integers m and n such that $am + bn = 1$.

Example 9.7. $a = 14 = 2 \cdot 7$ and $b = 15 = 3 \cdot 5$ are coprime.

Example 9.8. $a = 111 = 3 \cdot 37$ and $768 = 2^8 \cdot 3$ are not coprime.

Example 9.9. Among integers between 1 and 11 inclusive, 1, 5, 7, and 11 are coprime to 12. Mod 12, these four integers comprise two pairs: $1 \equiv -11$ and $5 \equiv -7$. By Theorem 8.21, an arbitrary integer N is coprime to 12 if and only if N is congruent mod 12 to 1, 5, 7, or 11.

Proposition 9.10. *If p is prime, then*

- (i) $\gcd(a, p) = p$ if and only if $p \mid a$.
- (ii) $\gcd(a, p) = 1$ if and only if $p \nmid a$. In particular, if $0 < a < p$, then a is coprime to p .

Proof. (i) By definition, $\gcd(a, p) \mid a$, so if $\gcd(a, p) = p$, then $p \mid a$.

Conversely, if $p \mid a$, then $\gcd(a, p) = p$ since $p \mid p$ always.

(ii) If p is prime and a in \mathbf{Z} is arbitrary, then *a priori* $\gcd(a, p)$ is either p or 1, since those are the only positive divisors of p . The second assertion follows immediately from (i). \square

Theorem 9.11. *Suppose $\gcd(a, b) = 1$. If $a \mid bc$, then $a \mid c$.*

Proof. By hypothesis, there exists an integer q such that $aq = bc$. Since $\gcd(a, b) = 1$, there exist integers m and n such that $am + bn = 1$. Multiplying by c and substituting,

$$c = c(am + bn) = acm + (bc)n = acm + (aq)n = a(cm + qn).$$

But since $cm + qn$ is an integer, $a \mid c$. \square

Theorem 9.12 (Euclid's lemma). *Let a and b be integers. If p is prime and $p \mid ab$, then $p \mid a$ or $p \mid b$.*

Proof. If $p \mid a$ there is nothing to prove. Otherwise, $\gcd(a, p) = 1$ by Proposition 9.10, so $p \mid b$ by Theorem 9.11. \square

Corollary 9.13. *If a_1, \dots, a_n are integers, p is prime, and $p \mid a_1 a_2 \dots a_n$, then there exists an index i such that $p \mid a_i$.*

Proof. The corollary follows by mathematical induction on the number of factors. Euclid's lemma is the base case, for two factors. The details are left as an exercise. \square

Remark 9.14. The hypothesis of coprimality cannot be dropped in Theorem 9.11: If $a = 6$, $b = 4$, and $p = 12$, then $p \mid ab$, but $p \nmid a$ and $p \nmid b$.

Of course, each prime factor of 12, namely 2 or 3, *is* a divisor of either a or b (or both), in accordance with Corollary 9.13.

9.2 Prime Factorizations

We are aiming for the *Fundamental Theorem of Arithmetic*: Every integer greater than 1 factors “uniquely” into primes.

The word “uniquely” requires explanation here. The integer $n = 60$ factors as $p_1 p_2 p_3 p_4 = 2 \cdot 2 \cdot 3 \cdot 5$, a product of four factors. Since multiplication is commutative, an arbitrary ordering of the same factors gives the same value for the product. Technically, however, $p_1 p_2 p_3 p_4$ and $p_2 p_1 p_3 p_4$ are “distinct products” (since the factors occur in different orders), even though they are *identical* as products of integers (because $p_1 = p_2$).

To avoid this purely linguistic issue, let us agree to organize products of primes so that (i) all occurrences of a given prime are gathered into a single prime power, and (ii) these prime powers are listed with the primes in increasing order.

We say that a product of prime powers satisfying (i) and (ii) is in *standard form*. Symbolically, a product of prime powers

$$N = p_1^{\nu_1} p_2^{\nu_2} \cdots p_m^{\nu_m} = \prod_{i=1}^m p_i^{\nu_i}, \quad \nu_i > 0 \text{ for } i = 1, \dots, m,$$

is in standard form if $p_1 < p_2 < \cdots < p_m$. To say the integer N factors “uniquely” into primes means any two representations of N as products of primes have *identical* standard forms.

Example 9.15. The products $60 = 2^2 \cdot 3 \cdot 5$ and $2352 = 2^4 \cdot 3 \cdot 7^2$ are in standard form, while $2 \cdot 2 \cdot 3 \cdot 5$ (condition (i) unmet) and $3 \cdot 2^2 \cdot 5$ (condition (ii) unmet) are not.

The proof of the Fundamental Theorem is broken into “existence” and “uniqueness”, since the required techniques are so different.

Theorem 9.16 (Existence of prime factorizations). *For every integer $N \geq 2$, there exist primes $p_1 < p_2 < \cdots < p_m$ and positive integers $\nu_1, \nu_2, \dots, \nu_m$ such that*

$$N = p_1^{\nu_1} p_2^{\nu_2} \cdots p_m^{\nu_m} = \prod_{i=1}^m p_i^{\nu_i}.$$

Briefly, every integer $N \geq 2$ factors into primes.

Proof. The proof proceeds by mathematical induction on the following statement:

$P(N)$ For every integer n with $2 \leq n \leq N$, n factors into primes.

For example, $P(4)$ asserts “2, 3, and 4 all factor into primes”.

The statement $P(2)$, “2 factors into primes”, is true because 2 is prime. This establishes the base case.

Assume inductively that $P(k)$ is true for some $k \geq 2$, that is, every integer n with $2 \leq n \leq k$ factors into primes. We wish to show that $(k+1)$ factors into primes, so that every integer n with $2 \leq n \leq k+1$ factors into primes.

By Theorem 9.5, $k+1$ has a prime divisor p . That is, there exists an integer q , $1 \leq q \leq k$, such that $(k+1) = pq$.

If $q = 1$, then $k+1 = p$ is itself prime. Otherwise, we have $2 \leq q \leq k$, so q factors into primes by the inductive hypothesis. Since p is prime, $k+1$ itself factors into primes.

In either case, $P(k)$ implies $P(k+1)$ for arbitrary $k \geq 2$, so by the principle of mathematical induction, $P(N)$ is true for all $N \geq 2$. \square

Theorem 9.17 (Uniqueness of prime factorization). *Suppose there exist primes $p_1 < p_2 < \cdots < p_m$ and $q_1 < q_2 < \cdots < q_\ell$, and there exist positive integers $\nu_1, \nu_2, \dots, \nu_m$ and $\mu_1, \mu_2, \dots, \mu_\ell$, such that*

$$p_1^{\nu_1} p_2^{\nu_2} \cdots p_m^{\nu_m} = q_1^{\mu_1} q_2^{\mu_2} \cdots q_\ell^{\mu_\ell}.$$

Then $\ell = m$, and for all $i = 1, \dots, m$, we have $p_i = q_i$ and $\nu_i = \mu_i$.

Proof. The proof is an increasingly-familiar refrain: Proceed by induction on the statement

$P(N)$ For every integer n with $2 \leq n \leq N$, n has a unique prime factorization (in standard form).

A complete proof is left as an exercise, but here is the key step: Assume inductively that every integer n , $2 \leq n \leq k$, factors uniquely into primes, and suppose

$$k+1 = p_1^{\nu_1} p_2^{\nu_2} \cdots p_m^{\nu_m} = q_1^{\mu_1} q_2^{\mu_2} \cdots q_\ell^{\mu_\ell}.$$

Consider the smallest prime factors p_1 and q_1 in the respective products. If $p_1 < q_1$, then $p_1 < q_i$ for all i (the prime factors are listed in increasing order). However, Corollary 9.13 implies $p_1 \mid q_i$ for some i , i.e.,

q_i is not prime. Contrapositively, if the q_i are all prime, then $q_1 \leq p_1$. A similar argument proves $p_1 \leq q_1$; thus $p_1 = q_1$.

Writing $k + 1 = np_1$ and cancelling $p_1 = q_1$ from the prime factorization of $k + 1$ gives

$$p_1^{\nu_1-1} p_2^{\nu_2} \cdots p_m^{\nu_m} = q_1^{\mu_1-1} q_2^{\mu_2} \cdots q_\ell^{\mu_\ell} = n \leq k.$$

By the inductive hypothesis, these factorizations are identical, so the factorizations of $k + 1$ are identical as well. \square

Respectively, Theorems 9.16 and 9.17 establish the the existence and uniqueness portions of the following basic result.

Theorem 9.18 (The Fundamental Theorem of Arithmetic). *Let $N \geq 2$ be an integer. There exist primes $p_1 < p_2 < \cdots < p_m$ and positive integers $\nu_1, \nu_2, \dots, \nu_m$, uniquely defined by N , such that*

$$N = p_1^{\nu_1} p_2^{\nu_2} \cdots p_m^{\nu_m} = \prod_{i=1}^m p_i^{\nu_i}.$$

Applications of the Fundamental Theorem

Among the many applications of the Fundamental Theorem of Arithmetic is Euclid's theorem on the infinitude of primes.

Theorem 9.19. *There exist infinitely many primes.*

Proof. Let $S = \{p_1, \dots, p_n\}$ be an arbitrary finite collection of primes, and let $N = p_1 p_2 \cdots p_n + 1$. By the Fundamental Theorem of Arithmetic, N can be factored into primes. However, no prime factor of N is an element of S , since by construction N leaves a remainder of 1 on division by each element of S . It follows that there exists a prime not in S , which means S is not the set of all primes. Since S was arbitrary, the set of primes is not finite. \square

Theorem 9.20. *Let $p_1 < p_2 < \cdots < p_m$ be primes. If*

$$N = p_1^{\nu_1} p_2^{\nu_2} \cdots p_m^{\nu_m} = \prod_{i=1}^m p_i^{\nu_i}, \quad \nu_i > 0 \text{ for } i = 1, \dots, m,$$

then the divisors of N are precisely the integers expressible in the form

$$a = p_1^{\mu_1} p_2^{\mu_2} \cdots p_m^{\mu_m} = \prod_{i=1}^m p_i^{\mu_i}, \quad 0 \leq \mu_i \leq \nu_i \text{ for } i = 1, \dots, m.$$

Corollary 9.21. *With notation as in the theorem, N has exactly*

$$(\nu_1 + 1)(\nu_2 + 1) \cdots (\nu_m + 1)$$

positive divisors.

Example 9.22. The integer $18 = 2 \cdot 3^2$ has exactly $(1 + 1)(2 + 1) = 6$ positive divisors. These can be listed by finding all ordered pairs of non-negative exponents (μ_1, μ_2) with $\mu_1 \leq 1$ and $\mu_2 \leq 2$:

$$\begin{array}{cccccc} (\mu_1, \mu_2) : & (0, 0) & (1, 0) & (0, 1) & (1, 1) & (0, 2) & (1, 2) \\ 2^{\mu_1} \cdot 3^{\mu_2} : & 1 & 2 & 3 & 6 & 9 & 18 \end{array}$$

Example 9.23. $26000 = 2^4 \cdot 5^3 \cdot 13$ has $(4 + 1)(3 + 1)(1 + 1) = 40$ positive divisors.

Example 9.24. Let $p_1 < p_2 < p_3 < \dots$ be the listing of *all* primes, taken in increasing order. To each sequence (m_1, m_2, m_3, \dots) of non-negative integers with at most finitely many non-zero terms, associate the positive integer

$$N = \prod_{i=1}^{\infty} p_i^{m_i}.$$

The product has only finitely many factors different from 1, so may be viewed as a finite product.

By the Fundamental Theorem of Arithmetic, the mapping f from the set of sequences of non-negative exponents to the positive integers is a *bijection*!

The mapping f satisfies a property reminiscent of the law of exponents: If $\mathbf{m} = (m_1, m_2, \dots)$ and $\mathbf{m}' = (m'_1, m'_2, \dots)$ are sequences, then

$$f(\mathbf{m} + \mathbf{m}') = f(\mathbf{m}) \cdot f(\mathbf{m}').$$

In principle, one can find the gcd of a and b by finding the associated sequences of exponents, taking the smaller exponent for each prime, and forming the resulting number. Analogously, the lcm is found by taking the larger exponent for each prime. For example,

$$\begin{aligned} 120 &= 2^3 \cdot 3^1 \cdot 5^1 \cdot 7^0 \leftrightarrow (3, 1, 1, 0, \dots) \\ 126 &= 2^1 \cdot 3^2 \cdot 5^0 \cdot 7^1 \leftrightarrow (1, 2, 0, 1, \dots) \\ \gcd(120, 126) &= 2^1 \cdot 3^1 \cdot 5^0 \cdot 7^0 \leftrightarrow (1, 1, 0, 0, \dots) \\ \text{lcm}(120, 126) &= 2^3 \cdot 3^2 \cdot 5^1 \cdot 7^1 \leftrightarrow (3, 2, 1, 1, \dots). \end{aligned}$$

This gives a second proof that $\gcd(a, b) \operatorname{lcm}(a, b) = ab$, see Exercise 8.8.

In practice, using prime factorization to find greatest common divisors is monumentally inefficient, while Theorem 8.24 (Euclid's algorithm) is computationally feasible for any integers small enough to represent conveniently in binary (millions of digits, say).

Exercises

Exercise 9.1. (a) Factor 2754 into primes, and determine the number of positive divisors.

(b) Factor 20400 into primes, and determine the number of positive divisors.

(c) Use the results of parts (a) and (b) to find the prime factorizations of $\gcd(2754, 20400)$ and $\operatorname{lcm}(2754, 20400)$.

Exercise 9.2. Prove Corollary 9.13.

Exercise 9.3. Prove Theorem 9.17.

Exercise 9.4. Pick an arbitrary three-digit number and write it down twice. Divide this integer by 7, then divide by 11, and finally divide by 13. The final quotient is the original number.

For example, starting with 456 yields, successively,

$$456456 \xrightarrow{\div 7} 65208 \xrightarrow{\div 11} 5928 \xrightarrow{\div 13} 456.$$

(a) Explain why this trick works.

(b) Give the details of the analogous trick if a two-digit number is written down three times, as in 424242.

Exercise 9.5. (a) Prove an integer N is divisible by 4 if and only if the number comprising the last two digits of N is divisible by 4.

(b) Prove an integer is divisible by 8 if and only if the number comprising the last three digits is divisible by 8.

Exercise 9.6. Prove an integer is divisible by 9 if and only if the sum of its digits is divisible by 9.

Hint: If $N = a_n \dots a_3 a_2 a_1$ is the decimal representation of N , then

$$N = a_1 + 10 a_2 + 100 a_3 + \cdots + 10^{n-1} a_n = \sum_{k=1}^n a_k 10^{k-1},$$

while the sum of the digits is

$$S = a_1 + a_2 + a_3 + \cdots + a_n = \sum_{k=1}^n a_k.$$

What can you say about $N - S$?

Exercise 9.7. Let \mathbf{Z}_7^\times be the set of non-zero residue classes mod 7, equipped with the operation of multiplication mod 7. Make a Cayley table, and show each element is invertible by exhibiting its inverse.

Exercise 9.8. Let p be a prime, \mathbf{Z}_p the set of residue classes mod p , equipped with the operation of multiplication mod p .

- (a) Show that if $[a]$ is a non-zero residue class mod p , there exists a residue class $[b]$ such that $[a][b] = [1]$. That is, every non-zero residue class mod p is invertible.
- (b) Show that \mathbf{Z}_p^\times , the set of non-zero residue classes mod p , is closed under multiplication mod p .

Exercise 9.9. Let p be a prime. Prove that $(p-1)! \equiv -1 \pmod{p}$.
Hints: By the preceding exercise, $(\mathbf{Z}_p^\times, \cdot)$ is a group. In the product $(p-1)!$, pair up residue classes with their multiplicative inverses. Handle the case $p = 2$ separately.

Exercise 9.10. Find all primes p such that $p+2$ and $p+4$ are also prime.

Exercise 9.11. Let N be an integer greater than 1.

- (a) Prove that if N is composite, there exists a divisor k of N such that $k^2 \leq N$.
- (b) Prove that if $2 \leq N \leq 120$, and if N is not divisible by any of 2, 3, 5, or 7, then N is prime.

Exercise 9.12. Recall that $10! = 3\,628\,800$ ends with two 0's, while $15! = 1\,307\,674\,368\,000$ ends with three 0's. Without using a computer, determine the number of 0's at the end of $1000!$ (the factorial of 1000).

Exercise 9.13. Determine (with proof) the number of primes in the sequence 101, 10101, 1010101,

Exercise 9.14. Let N be a positive integer. Prove there exist N consecutive composite integers.

Exercise 9.15. This question requires facts from calculus about infinite series. Let $s > 1$ be a real number.

(a) Show that if p is prime, then

$$\sum_{k=0}^{\infty} \frac{1}{(p^s)^k} = 1 + \frac{1}{p^s} + \frac{1}{(p^s)^2} + \frac{1}{(p^s)^3} + \cdots = \frac{1}{1 - p^{-s}}.$$

(b) Argue formally that

$$\sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}}.$$

Hint: Multiply the identities in (a) over all primes p , and use the Fundamental Theorem of Arithmetic.

In part (b), the infinite sum on the left makes sense for all *complex* s with real part greater than 1. Moreover, the resulting function of s has a unique complex-differentiable extension to the complex plane with the point $s = 1$ removed. This extension, the *Riemann ζ function*, has been the subject of intensive mathematical study for nearly three centuries, and possesses a vast literature.

The ζ function turns out to be equal to zero at each negative even integer. The *Riemann hypothesis*, one of the seven Clay Mathematics Institute “Millennium Problems”, asserts that all other zeros of the ζ function lie on the line $\operatorname{Re} s = \frac{1}{2}$.

The identities $\zeta(0) = -\frac{1}{2}$ and $\zeta(-1) = -\frac{1}{12}$ are often expressed, by formal substitution in (b), as

$$1 + 1 + 1 + \cdots = -\frac{1}{2}, \quad 1 + 2 + 3 + \cdots = -\frac{1}{12}.$$

The infinite series diverge, of course. The values are said to be obtained by “ ζ regularization”.

Chapter 10

Multiplicative Inverses in \mathbf{Z}_n

Fix an integer $n \geq 2$, and let \mathbf{Z}_n be the set of residue classes mod n :

$$\mathbf{Z}_n = \{[0], [1], \dots, [n-1] = [-1]\}.$$

By Theorem 8.12, there exist well-defined binary operations $+$ and \cdot on \mathbf{Z}_n satisfying

$$[a] + [b] = [a + b], \quad [a] \cdot [b] = [a \cdot b]$$

for all integers a and b . These operations are both associative and commutative, since the corresponding integer operations enjoy these properties. For example, if a , b , and c are integers, then

$$\begin{aligned} ([a] + [b]) + [c] &= [a + b] + [c] \\ &= [(a + b) + c] \\ &= [a + (b + c)] \\ &= [a] + [b + c] = [a] + ([b] + [c]). \end{aligned}$$

Replacing each “ $+$ ” by “ \cdot ” proves multiplication is associative. An entirely similar argument shows multiplication distributes over addition. Each operation has an identity element. The residue class $[0]$ is an identity element for addition, while $[1]$ is an identity element for multiplication.

Up to this point, the structures $(\mathbf{Z}_n, +)$ and (\mathbf{Z}_n, \cdot) have completely parallel properties. When we ask about *inverses*, however, the stories quickly diverge. Under addition, matters are trivial: Every element $[a]$ of \mathbf{Z}_n has an additive inverse, $-[a] = [n - a]$. In other words, $(\mathbf{Z}_n, +)$ is an Abelian group. Multiplicative inverses, by contrast, occupy the rest of this chapter.

10.1 Invertibility

Definition 10.1. A residue class $[a]$ is *invertible* in \mathbf{Z}_n if there exists a class $[x]$ such that $[a][x] = [1]$. The class $[x]$ is the *inverse* of $[a]$ in \mathbf{Z}_n .

Remark 10.2. For general reasons discussed in Chapter 6, an invertible class $[a]$ has a unique inverse, customarily denoted $[a]^{-1}$.

Remark 10.3. Applied to residue classes in \mathbf{Z}_n , the term “invertible” *always* refers to multiplication. The additive inverse of a residue class is its *negative*.

Example 10.4. In \mathbf{Z}_7 , $[2][4] = [1]$, so $[2]$ and $[4]$ are invertible in \mathbf{Z}_7 , and each is the inverse of the other.

Example 10.5. The class $[0]$ is *never* invertible in \mathbf{Z}_n : $[0][x] = [0] \neq [1]$ regardless of x .

In \mathbf{Z}_6 , the classes $[2]$, $[3]$, and $[4]$ are not invertible.

Definition 10.6. An invertible class $[a]$ in \mathbf{Z}_n is a *unit* (mod n). The set of units (mod n) is denoted \mathbf{Z}_n^\times .

Example 10.7. The Cayley tables for addition and multiplication mod 5 are

+	[0]	[1]	[2]	[3]	[4]
[0]	[0]	[1]	[2]	[3]	[4]
[1]	[1]	[2]	[3]	[4]	[0]
[2]	[2]	[3]	[4]	[0]	[1]
[3]	[3]	[4]	[0]	[1]	[2]
[4]	[4]	[0]	[1]	[2]	[3]

·	[0]	[1]	[2]	[3]	[4]
[0]	[0]	[0]	[0]	[0]	[0]
[1]	[0]	[1]	[2]	[3]	[4]
[2]	[0]	[2]	[4]	[1]	[3]
[3]	[0]	[3]	[1]	[4]	[2]
[4]	[0]	[4]	[3]	[2]	[1]

As expected because $(\mathbf{Z}_5, +)$ is a group, each element of \mathbf{Z}_5 appears exactly once in each row and column of the addition table.

The element $[0]$ appears more than once in the first row of the multiplication table, confirming (\mathbf{Z}_5, \cdot) is *not* a group.

By inspection, every *non-zero* class mod 5 has a multiplicative inverse, and a product of invertible classes is invertible. Reading from the table, $[1]$ and $[4]$ are their own inverses, while $[2]^{-1} = [3]$ and $[3]^{-1} = [2]$. The set $\mathbf{Z}_5^\times = \{[1], [2], [3], [4]\}$ is therefore a group under multiplication (mod 5), compare Corollaries 10.9 and 10.11.

Proposition 10.8. Let $n > 1$. If $[a]$ and $[b]$ are invertible classes in \mathbf{Z}_n , then $[a]^{-1}$ and $[a][b]$ are invertible in \mathbf{Z}_n . That is, \mathbf{Z}_n^\times is closed under multiplication and under inversion.

Proof. Let $[a]$ in \mathbf{Z}_n be invertible. The condition $[x] = [a]^{-1}$ says $[a][x] = [1]$. The relationship between an invertible class and its inverse is reciprocal, in the sense that $[x] = [a]^{-1}$ if and only if $[x]^{-1} = [a]$. Particularly, the inverse of $[a]$ is invertible.

Further, if $[b]$ is invertible with $[y] = [b]^{-1}$, then

$$([a][b])([y][x]) = [a]([b][y])[x] = [a][1][x] = [a][x] = [1].$$

This means the class $[a][b]$ is invertible in \mathbf{Z}_n , and its inverse is the product (in either order) of the inverses of $[a]$ and $[b]$. \square

Corollary 10.9. *Let $n > 1$. The pair $(\mathbf{Z}_n^\times, \cdot)$ is an Abelian group.*

Proof. For general reasons, multiplication (mod n) is an associative and commutative binary operation on \mathbf{Z}_n , $[1]$ is an identity element, and $[1]$ is invertible. By Proposition 10.8, \mathbf{Z}_n^\times is closed under multiplication (mod n), i.e. multiplication (mod n) is a binary operation on \mathbf{Z}_n^\times , and every element of \mathbf{Z}_n^\times is invertible. \square

Our next task is to find a computational criterion for invertibility.

Theorem 10.10. *If $a \in \mathbf{Z}$, then the residue class $[a] \in \mathbf{Z}_n$ is invertible if and only if $\gcd(a, n) = 1$.*

Proof. By definition, $[a]$ is invertible in \mathbf{Z}_n if and only if there exists an integer x such that $[a][x] = [1]$, namely, $ax \equiv 1 \pmod{n}$. This holds if and only if there exist integers x and y such that $ax + ny = 1$. By Theorem 8.21, this last condition is equivalent to $\gcd(a, n) = 1$. \square

Corollary 10.11. *If p is prime and $[a]$ is non-zero in \mathbf{Z}_p , then $[a]$ is invertible. That is, $\mathbf{Z}_p^\times = \mathbf{Z}_p \setminus \{[0]\}$.*

Proof. By Proposition 9.10, if $a \in \mathbf{Z}$, then $\gcd(a, p) = p$ if and only if $p \mid a$, and $\gcd(a, p) = 1$ otherwise. \square

Corollary 10.12. *The residue class $[a]$ is invertible in \mathbf{Z}_n if and only if $[-a] = [n - a]$ is invertible in \mathbf{Z}_n .*

Proof. By Remark 8.22, $\gcd(-a, n) = \gcd(a, n)$. \square

Example 10.13. We have $\mathbf{Z}_{12}^\times = \{[1], [5], [7], [11]\}$ (see also Example 9.9). Invertible classes occur in pairs of the form $[a]$ and $[-a]$, as guaranteed by Corollary 10.12. Since $[1]^2 = [1]$ and $[5]^2 = [1]$, we have $[7]^2 = [-5]^2 = [1]$ and $[11]^2 = [-1]^2 = [1]$. In particular, the group $(\mathbf{Z}_{12}^\times, \cdot)$ is not cyclic; every element has order 2.

Example 10.14. Since $14 = 2 \cdot 7$, we have

$$\mathbf{Z}_{14}^\times = \{[1], [3], [5], [9], [11], [13]\} = \{[1], [3], [5], [-5], [-3], [-1]\}.$$

Note that $[3]^2 = [9] = [-5]$, so $[3]^3 = [-15] = [-1]$. It follows that $[3]$ has order 6 in \mathbf{Z}_{14}^\times . The arithmetic properties of the cyclic group $(\mathbf{Z}_{14}^\times, \cdot)$ are easily understood by writing the elements as powers of $[3]$: $\{[3]^0, [3], [3]^2, [3]^3, [3]^4, [3]^5\} = \{[1], [3], [9], [13], [11], [5]\}$. For example, since $[11] = [3]^4$, we have $[11]^2 = [3]^8 = [3]^2 = [9]$.

Example 10.15. Since $30 = 2 \cdot 3 \cdot 5$, we have

$$\mathbf{Z}_{30}^\times = \{[1], [7], [11], [13], [17], [19], [23], [29]\}.$$

To study the group structure of $(\mathbf{Z}_{30}^\times, \cdot)$, we can pair off elements and their negatives to avoid multiplying numbers of absolute value greater than $30/2 = 15$.

Direct calculation gives $[7]^2 = [19]$, $[11]^2 = [1]$, and $[13]^2 = [19]$. It follows immediately that

$$[23]^2 = [-7]^2 = [19], \quad [19]^2 = [-11]^2 = [1], \quad [17]^2 = [-13]^2 = [19].$$

Putting these conclusions together, $[7]^4 = [19]^2 = [1]$, $[13]^4 = [1]$, $[17]^4 = [1]$, and $[23]^4 = [1]$. In particular, $(\mathbf{Z}_{30}^\times, \cdot)$ is *not* a cyclic group, since no element has order 8.

Each element of order 2 is its own inverse. To invert elements of order 4, argue as follows:

$$[7]^{-1} = [7]^3 = [19][7] = [-11][7] = [-17] = [13].$$

The inverses of the other elements may be deduced with no additional effort:

$$[13]^{-1} = [7], \quad [17]^{-1} = [-13]^{-1} = [-7] = [23], \quad [23]^{-1} = [17].$$

To emphasize, these calculations have been carried out without computing any products larger than $13^2 = 169$.

For larger n , it may be inconvenient to list the elements of \mathbf{Z}_n^\times . Nonetheless, we can easily test individual residue classes for membership, and can calculate inverses.

Example 10.16. Determine whether the following are invertible in \mathbf{Z}_{105} , and if so find the inverse.

[51]: By the Euclidean algorithm,

$$\begin{aligned} 105 &= 2 \cdot 51 + 3 \\ 51 &= 17 \cdot 3 + 0, \end{aligned}$$

so $\gcd(105, 51) = 3 \neq 1$. Thus [51] is not invertible in \mathbf{Z}_{105} .

[32]: The Euclidean algorithm gives

$$\begin{aligned} 105 &= 3 \cdot 32 + 9 \\ 32 &= 3 \cdot 9 + 5 \\ 9 &= 1 \cdot 5 + 4 \\ 5 &= 1 \cdot 4 + 1. \end{aligned}$$

Thus $\gcd(105, 32) = 1$, so [32] is invertible in \mathbf{Z}_{105} . To find the inverse, write 1 as a linear combination of 105 and 32:

$$\begin{aligned} 1 &= 5 - 4 \\ &= 5 - (9 - 5) = 2 \cdot 5 - 9 \\ &= 2 \cdot (32 - 3 \cdot 9) - 9 = 2 \cdot 32 - 7 \cdot 9 \\ &= 2 \cdot 32 - 7 \cdot (105 - 3 \cdot 32) = 23 \cdot 32 - 7 \cdot 105. \end{aligned}$$

Reducing mod 105, $[23][32] = [1]$, or $[32]^{-1} = [23]$.

As noted earlier, [0] never has a multiplicative inverse in \mathbf{Z}_n . To characterize non-invertible elements generally, we introduce the following definition.

Definition 10.17. A residue class $[a]$ in \mathbf{Z}_n is a *zero divisor* if there exists a non-zero class $[b]$ such that $[a][b] = [0]$.

Theorem 10.18. *The class $[a]$ in \mathbf{Z}_n is a zero divisor if and only if $\gcd(a, n) > 1$.*

Remark 10.19. Combining with Theorem 10.10, each class $[a]$ in \mathbf{Z}_n is either invertible or a zero divisor, but not both.

Proof. (If $[a]$ is a zero divisor, then $\gcd(a, n) > 1$). Contrapositively, suppose $d = \gcd(a, n) = 1$, and that $[a][b] = [0]$ for some $[b]$. We want to show $[b] = [0]$, which will prove $[a]$ is not a zero divisor. By Theorem 10.10, the class $[a]$ is invertible in \mathbf{Z}_n , so there exists an $[x]$ with $[x][a] = [1]$. Multiplying $[a][b] = [0]$ by $[x]$ on the left,

$$[0] = [x][0] = [x]([a][b]) = ([x][a])[b] = [1][b] = [b].$$

(If $\gcd(a, n) > 1$, then $[a]$ is a zero divisor). Suppose $\gcd(a, n) > 1$. If $a \equiv 0 \pmod{n}$, namely if $[a] = [0]$, there is nothing to prove. Otherwise, we may divide a and n by $d = \gcd(a, n)$ and deduce there exist non-zero integers a' and n' such that $a = da'$ and $n = dn'$. Since $1 < d$, we have $n' < n$, which implies $[n'] \neq [0]$ in \mathbf{Z}_n . However,

$$[a][n'] = [(a'd)n'] = [a'(dn')] = [a'n] = [0].$$

By definition, $[a]$ is a zero divisor in \mathbf{Z}_n . □

Example 10.20. It may help to follow the preceding proof with specific numbers. Let $n = 8$ and $a = 6$. Here, $d = \gcd(6, 8) = 2$, so $a' = a/2 = 3$ and $n' = n/2 = 4$. As expected, $[a][n'] = [6][4] = [0]$ in \mathbf{Z}_8 , which proves $[a] = [6]$ is a zero divisor.

10.2 The Geometry of Multiplication

The elements of the additive group $(\mathbf{Z}_n, +)$ may be viewed as a set of n evenly-spaced points on a circle. For each integer a , there is a “multiplication” mapping $\phi_a : \mathbf{Z}_n \rightarrow \mathbf{Z}_n$ defined by $\phi_a([x]) = [ax]$. Without loss of generality, the “slope” a may be reduced mod n , namely, may be regarded as an element of \mathbf{Z}_n . Starting from $[0]$, the successive values of ϕ_a are $[0], [a], [2a], [3a], \dots$. These may be visualized as follows: Place a pencil at $[0]$ and count off a spaces at a time around the circle, joining successive values with line segments, see Figure 10.1. The corners on such a diagram correspond to elements of the image of ϕ_a , namely the classes $[k]$ such that $[a][x] = [k]$ has a solution $[x]$ in \mathbf{Z}_n .

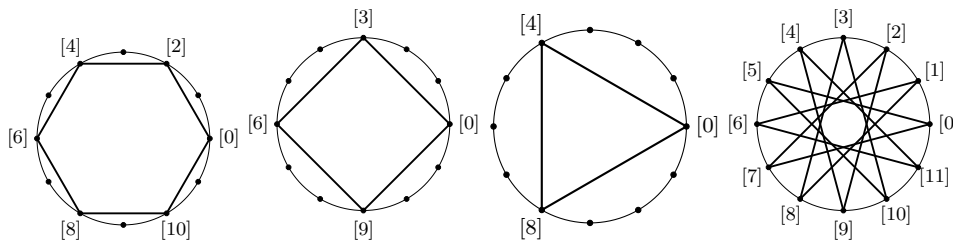


Figure 10.1: Multiplication diagrams for ϕ_2 , ϕ_3 , ϕ_4 , and ϕ_5 on \mathbf{Z}_{12} .

Algebraically, $ax \equiv k \pmod{n}$ if and only if there exists an integer y with $ax + ny = k$. Geometrically, $ax = k$ is the location reached by counting off a spaces x times. Adding or subtracting an integer multiple

of n corresponds to discarding traversals of the entire circle, which have no effect on the location of a corner.

By Theorem 8.21, there exist integers x and y with $ax + ny = k$ if and only if $\gcd(a, n) \mid k$. In other words, consecutive corners on the diagram are separated by $\gcd(a, n)$ spaces in \mathbf{Z}_n , and the number of corners in the diagram is $n / \gcd(a, n)$.

The dichotomy implied by Theorems 10.10 and 10.18 consequently has a mapping interpretation: $\gcd(a, n) = 1$ if and only if there are no “gaps” between consecutive corners, if and only if $\phi_a : \mathbf{Z}_n \rightarrow \mathbf{Z}_n$ is surjective (every element of \mathbf{Z}_n is a corner of the diagram).

This mathematics is the basis of the ingenious toy Spirograph, which consists of circular plastic rings of inner circumference n and disk-shaped gears of varying circumference a , Figure 10.2. Both the rings

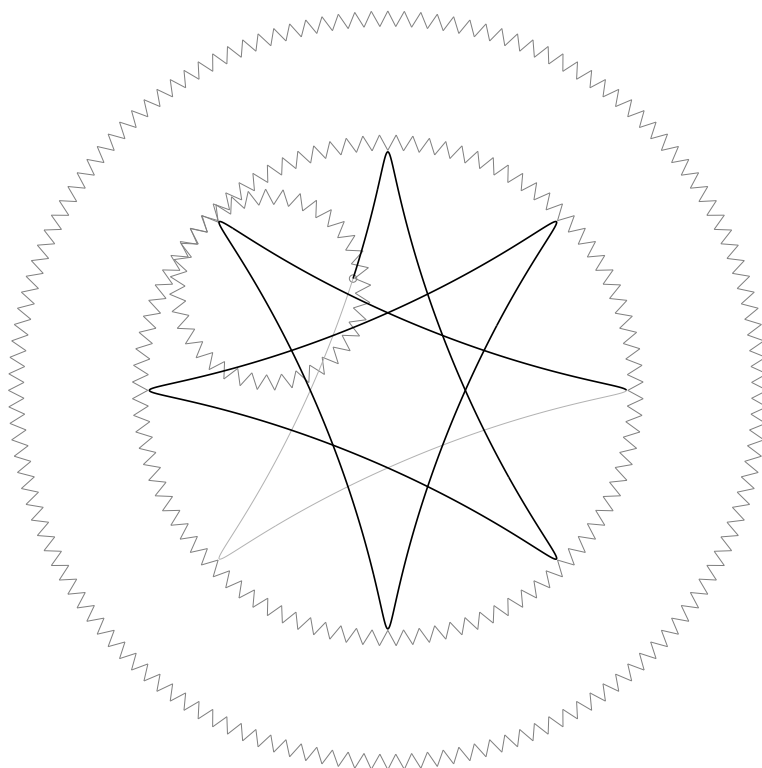


Figure 10.2: Spirograph; a 36-tooth gear in a 96-tooth ring.

and gears have teeth to ensure the gears roll without slipping. The gears have holes through which a pen fits. By tacking the ring to a sheet of paper and rolling a gear around the inside of a ring until the

pattern closes, you produce smooth mathematical curves having the overall geometry of the mapping $\phi_a : \mathbf{Z}_n \rightarrow \mathbf{Z}_n$.

Example 10.21. In Figure 10.2, $n = 96 = 8 \cdot 12$ and $a = 36 = 3 \cdot 12$. Since $\gcd(36, 96) = 12$, there are 12 teeth between consecutive points of the pattern, and the pattern has $96/12 = 8$ points.

Counting by 36 (mod 96), we obtain 0, 36, 72, 12, 48, 84, 24, 60 in succession before returning to 0. These numbers are precisely the multiples of 12 (mod 96) listed in the order they're visited when following the Spirograph pattern.

10.3 Linear Congruences

Example 8.15 introduced a simple linear congruence and solved it by essentially *ad hoc* tricks. We now have the algebraic tools to solve the general linear congruence

$$(10.1) \quad ax \equiv b \pmod{n},$$

in which a , b , and n are given and x is unknown.

Theorem 10.22. *Let a , n , and b be integers.*

- (i) *If $\gcd(a, n) \nmid b$, then (10.1) has no solution in \mathbf{Z}_n .*
- (ii) *If $\gcd(a, n) = 1$, then (10.1) has a unique solution in \mathbf{Z}_n .*
- (iii) *If $\gcd(a, n) \mid b$, then (10.1) has precisely $\gcd(a, n)$ solutions in \mathbf{Z}_n .*

Proof. (i) Assume contrapositively that $ax \equiv b \pmod{n}$ has a solution. There exists an integer y such that $ax = b + ny$, or $ax - ny = b$. This expresses b as a linear combination of a and n . By Theorem 8.21, $\gcd(a, n) \mid b$.

(ii) Assume $\gcd(a, n) = 1$. We must prove (10.1) has a solution (existence), and any two solutions are equal (uniqueness).

We *construct* a solution of (10.1). There exist integers s and t such that $as + nt = 1$. Multiply by b to get $a(sb) + n(tb) = b$. Setting $x = sb$, the preceding equation says $ax \equiv b \pmod{n}$. This establishes existence.

Still assuming $\gcd(a, n) = 1$, if x_1 and x_2 are solutions of (10.1), then $ax_1 \equiv b$ and $ax_2 \equiv b$. Subtracting, $a(x_1 - x_2) \equiv 0 \pmod{n}$, or

$n \mid a(x_1 - x_2)$. Since $\gcd(a, n) = 1$, Theorem 9.11 implies $n \mid (x_1 - x_2)$, so $x_1 \equiv x_2 \pmod{n}$. This proves uniqueness of solutions mod n .

(iii) For convenience, write $d = \gcd(a, n)$. By Theorem 8.21, there exist integers s and t such that $as + nt = d$, and there exist integers a' and n' such that $a = a'd$ and $n = n'd$. Further, if $d \mid b$, there exists an integer b' such that $b = b'd$.

Dividing $as + nt = d$ by d gives $a's + n't = 1$, which implies $\gcd(a', n') = 1$. Part (ii) of this theorem guarantees that the congruence $a'x \equiv b' \pmod{n'}$ has a unique solution $x_0 = sb'$. Multiplying $a'x_0 \equiv b' \pmod{n'}$ by d shows x_0 is a solution of (10.1).

It remains to show (10.1) has d solutions. For $i = 0, 1, 2, \dots, d-1$, let $x_i = x_0 + in'$, see Figure 10.3. Since $an' = a'dn' = a'n$, each x_i is a solution:

$$ax_i = ax_0 + i(an') = ax_0 + i(a'n) \equiv ax_0 \equiv b \pmod{n}.$$

Moreover, these d numbers are distinct mod n : If $0 \leq j \leq i < d$, then $x_i \equiv x_j \pmod{n}$ if and only if

$$n = n'd \mid (x_i - x_j) = (in' - jn') = (i - j)n',$$

if and only if $d \mid (i - j)$, if and only if $i = j$. We have therefore constructed d distinct solutions of (10.1).

There are no other solutions. If x is an arbitrary solution of $ax \equiv b \pmod{n}$, dividing through by d proves $a'x \equiv b' \pmod{n'}$, so $x \equiv x_0 \pmod{n'}$ by uniqueness. In other words, there exists an integer i such that $x = x_0 + in'$, so x is one of the solutions found above. \square

Remark 10.23. Theorem 10.22 algebraically recapitulates our earlier geometric discussion of multiplication in \mathbf{Z}_n .

Part (iii) gives an additional geometric refinement. If $\gcd(a, n) = d$, then the mapping $\phi_a : \mathbf{Z}_n \rightarrow \mathbf{Z}_n$ achieves each of its values precisely d times. Putting $n' = n/d$, the set of values of ϕ_a is the cyclic subgroup $d\mathbf{Z}_{n'} \subseteq (\mathbf{Z}_n, +)$. Viewing \mathbf{Z}_n as a bracelet with n beads and $\mathbf{Z}_{n'}$ as a bracelet with $n' = n/d$ beads, the mapping ϕ_a wraps \mathbf{Z}_n exactly d times around $\mathbf{Z}_{n'}$, see Figure 10.3.

Example 10.24. Solve the congruence $30x \equiv 18 \pmod{216}$.

Here $a = 30$ and $n = 216$, so $d = \gcd(a, n) = 6$. To find x_0 , divide through by 6 and solve $5x_0 \equiv 3 \pmod{36}$. Our earlier method using Euclid's algorithm gives $x_0 = 15$. There are $d = 6$ solutions in total, any two differing by a multiple of $n' = 36$: $x_1 = 15 + 36 = 51$, $x_2 = 87$, $x_3 = 123$, $x_4 = 159$, and $x_5 = 195$.

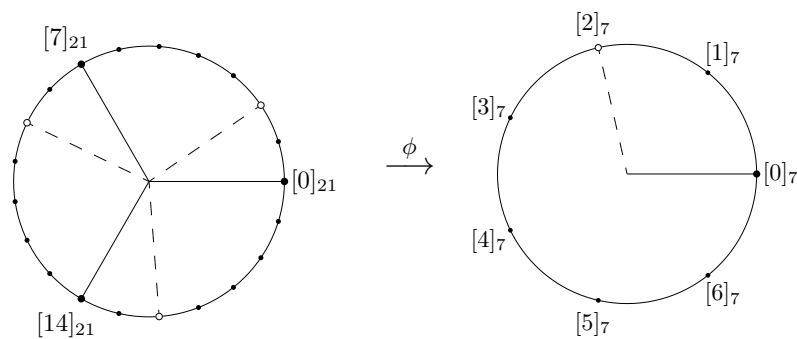


Figure 10.3: Wrapping \mathbf{Z}_n around $\mathbf{Z}_{n'}$ ($n = 21$, $n' = 7$, $d = 3$).

Exercises

Exercise 10.1. For the residue classes specified, determine whether or not each class is invertible in \mathbf{Z}_n , and if so, find the inverse.

(a) In \mathbf{Z}_{48} : $[a] = [17]$, $[a] = [21]$, $[a] = [25]$.

(b) In \mathbf{Z}_{140} : $[a] = [35]$, $[a] = [33]$, $[a] = [81]$.

Exercise 10.2. List the elements of \mathbf{Z}_9^\times and write out the Cayley table for multiplication. For $[a] = [2], [3], [4]$ in \mathbf{Z}_9 , sketch the multiplication diagram for ϕ_a .

Exercise 10.3. List the elements of \mathbf{Z}_{10}^\times and write out the Cayley table for multiplication. For $[a] = [2], [3], [4]$ in \mathbf{Z}_{10} , sketch the multiplication diagram for ϕ_a .

Exercise 10.4. List the elements of \mathbf{Z}_7^\times , write out the Cayley table for multiplication, and determine the inverse of each element.

Exercise 10.5. List the elements of \mathbf{Z}_{18}^\times , write out the Cayley table for multiplication, and determine the inverse of each element.

Exercise 10.6. Solve the congruence $18x \equiv 6 \pmod{24}$.

Exercise 10.7. Solve the congruence $18x \equiv 6 \pmod{28}$.

Exercise 10.8. Solve the congruence $150x \equiv 84 \pmod{567}$.

Exercise 10.9. Solve the congruence $150x \equiv 84 \pmod{210}$.

Chapter 11

Linear Transformations

Groups and other algebraic phenomena arise in geometry. This chapter introduces a simple but rich class of mappings of the Cartesian space \mathbf{R}^n , particularly mappings of the plane \mathbf{R}^2 . We begin with dimension-independent generalities.

11.1 The Cartesian Vector Space

In mathematics, “vectors” are objects that can be added to each other (satisfying the axioms of an Abelian group), and that can be multiplied by numerical “scalars” (satisfying axioms analogous to the associative and distributive laws, see Remark 11.2 below).

The elements of \mathbf{R}^n are ordered n -tuples of real numbers. We denote the general element of \mathbf{R}^n by $\mathbf{x} = (x^1, x^2, \dots, x^n)$. The superscripts are indices, not exponents.

Componentwise addition is a binary operation on \mathbf{R}^n , and $(\mathbf{R}^n, +)$ is an Abelian group with identity element $\mathbf{0} = (0, \dots, 0)$, and with $-\mathbf{x} = (-x^1, \dots, -x^n)$ the additive inverse of $\mathbf{x} = (x^1, \dots, x^n)$.

Scalar multiplication on \mathbf{R}^n is the mapping $\cdot : \mathbf{R} \times \mathbf{R}^n \mapsto \mathbf{R}^n$ defined by $t \cdot \mathbf{x} = t \cdot (x^1, \dots, x^n) = (tx^1, \dots, tx^n)$.

Definition 11.1. The data $(\mathbf{R}^n, +, \cdot)$ comprise the *vector space* \mathbf{R}^n . If $\mathbf{x} = (x^1, \dots, x^n)$ is a vector in \mathbf{R}^n , the number x^j is the *jth component* of \mathbf{x} . The special elements $\mathbf{e}_1 = (1, 0, \dots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \dots, 0)$, \dots , $\mathbf{e}_n = (0, \dots, 0, 1)$, are collectively called the *standard basis* of \mathbf{R}^n .

Remark 11.2. If \mathbf{x} , \mathbf{x}_1 , and \mathbf{x}_2 are vectors in \mathbf{R}^n and α and β are real scalars, the following properties hold, as you should check:

- (Associativity) $(\alpha\beta) \cdot \mathbf{x} = \alpha \cdot (\beta \cdot \mathbf{x})$.
- (Left-distributivity) $(\alpha + \beta) \cdot \mathbf{x} = (\alpha \cdot \mathbf{x}) + (\beta \cdot \mathbf{x})$.
- (Right-distributivity) $\alpha \cdot (\mathbf{x}_1 + \mathbf{x}_2) = (\alpha \cdot \mathbf{x}_1) + (\alpha \cdot \mathbf{x}_2)$.
- (Normalization) $1 \cdot \mathbf{x} = \mathbf{x}$.

Loosely, “the usual rules of algebra hold” when working with vectors. In practice, the dot denoting scalar multiplication is often omitted.

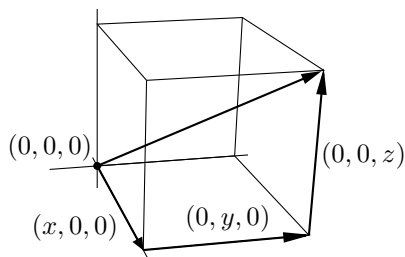
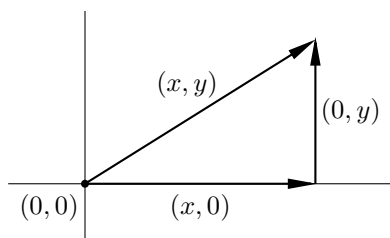
Remark 11.3. Denoting vectors as columns, we may use addition and scalar multiplication to decompose an arbitrary vector as

$$\begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^n \end{bmatrix} = \begin{bmatrix} x^1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ x^2 \\ \vdots \\ 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ x^n \end{bmatrix} = x^1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x^2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + x^n \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

More concisely, an arbitrary vector decomposes as a *linear combination* of standard basis vectors:

$$(11.1) \quad \mathbf{x} = x^1 \mathbf{e}_1 + x^2 \mathbf{e}_2 + \cdots + x^n \mathbf{e}_n = \sum_{j=1}^n x^j \mathbf{e}_j.$$

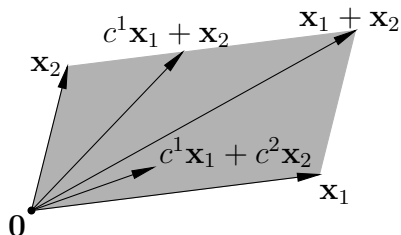
Remark 11.4. A vector \mathbf{x} may be viewed as a displacement, depicted as an arrow from $\mathbf{0}$ to \mathbf{x} , or more generally from an arbitrary point \mathbf{x}_0 to $\mathbf{x}_0 + \mathbf{x}$. Addition in \mathbf{R}^n may be viewed as concatenation of displacements. As with complex numbers, addition may also be interpreted as a parallelogram law.



Scalar multiplication may be interpreted as “stretching”, preserving (or diametrically reversing) the direction of displacement but changing its magnitude.

Example 11.5. The *parallelogram spanned by* vectors \mathbf{x}_1 and \mathbf{x}_2 in \mathbf{R}^n is the set of linear combinations of the form $c^1\mathbf{x}_1 + c^2\mathbf{x}_2$ with $0 \leq c^1 \leq 1$ and $0 \leq c^2 \leq 1$.

In particular, the *unit square* in \mathbf{R}^2 , the parallelogram spanned by the standard basis vectors $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$, consists of all points (c^1, c^2) whose Cartesian coordinates are between 0 and 1.

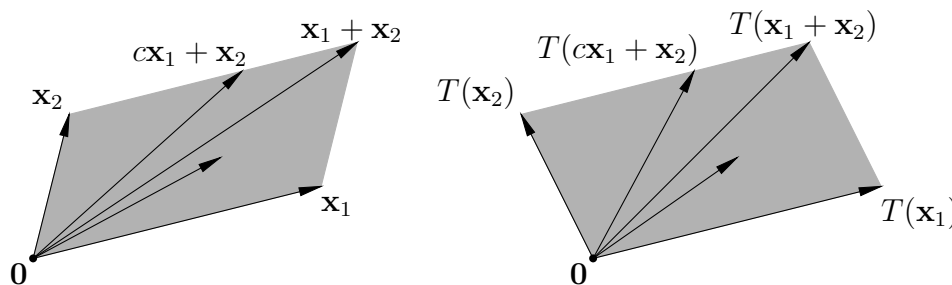


Within the family of mappings from the vector space \mathbf{R}^n to itself, among the simplest and most interesting are those that “respect” the vector space structure in the sense of “distributing” through linear combinations.

Definition 11.6. Let n and m be arbitrary positive integers. A mapping $T : \mathbf{R}^n \rightarrow \mathbf{R}^m$ is a *linear transformation* if

$$T(c \cdot \mathbf{x}_1 + \mathbf{x}_2) = c \cdot T(\mathbf{x}_1) + T(\mathbf{x}_2) \quad \text{for all } \mathbf{x}_1, \mathbf{x}_2 \text{ in } \mathbf{R}^n, \text{ all real } c.$$

A linear transformation “respects parallelograms” in the sense that the vertices \mathbf{x}_1 , \mathbf{x}_2 , and $\mathbf{x}_1 + \mathbf{x}_2$ of an arbitrary parallelogram “based at $\mathbf{0}$ ” map to the vertices $T(\mathbf{x}_1)$, $T(\mathbf{x}_2)$, and $T(\mathbf{x}_1) + T(\mathbf{x}_2)$ of a parallelogram based at $\mathbf{0}$. (For a general mapping, $T(\mathbf{x}_1 + \mathbf{x}_2) \neq T(\mathbf{x}_1) + T(\mathbf{x}_2)$.)



Remark 11.7. We are concerned with groups under mapping composition, and for the most part assume $m = n$; that is we restrict our attention to maps from \mathbf{R}^n to itself.

Mathematical induction on the number of summands shows that a linear transformation distributes over an arbitrary linear combination:

Lemma 11.8. Assume $T : \mathbf{R}^n \rightarrow \mathbf{R}^m$ is a linear transformation. If $(\mathbf{x}_i)_{i=1}^N$ are arbitrary vectors in \mathbf{R}^n and $(c^i)_{i=1}^N$ are scalars, then

$$T\left(\sum_{i=1}^N c^i \mathbf{x}_i\right) = \sum_{i=1}^N c^i T(\mathbf{x}_i).$$

Theorem 11.9. If $(\mathbf{y}_j)_{j=1}^n$ are elements of \mathbf{R}^m , there exists a unique linear transformation $T : \mathbf{R}^n \rightarrow \mathbf{R}^m$ satisfying $T(\mathbf{e}_j) = \mathbf{y}_j$ for all j .

Proof. (Uniqueness). By equation (11.1) and Lemma 11.8, a linear transformation is completely determined by its values on the standard basis:

$$T(\mathbf{x}) = T\left(\sum_{j=1}^n x^j \mathbf{e}_j\right) = \sum_{j=1}^n x^j T(\mathbf{e}_j) \quad \text{for all } \mathbf{x} \text{ in } \mathbf{R}^n.$$

(Existence). If $(\mathbf{y}_j)_{j=1}^n$ are arbitrary vectors in \mathbf{R}^m , the formula

$$T(\mathbf{x}) = \sum_{j=1}^n x^j \mathbf{y}_j$$

defines a mapping $T : \mathbf{R}^n \rightarrow \mathbf{R}^m$ satisfying $T(\mathbf{e}_j) = \mathbf{y}_j$ for all j . To prove T is a linear transformation, note that if $\mathbf{x}_1 = (x_1^1, \dots, x_1^n)$ and $\mathbf{x}_2 = (x_2^1, \dots, x_2^n)$ are arbitrary elements of \mathbf{R}^n and if c is real, then

$$\begin{aligned} T(c \cdot \mathbf{x}_1 + \mathbf{x}_2) &= \sum_{j=1}^n (cx_1^j + x_2^j) \mathbf{y}_j \\ &= c \sum_{j=1}^n x_1^j \mathbf{y}_j + \sum_{j=1}^n x_2^j \mathbf{y}_j = c \cdot T(\mathbf{x}_1) + T(\mathbf{x}_2). \quad \square \end{aligned}$$

Theorem 11.10. Let $T_1, T_2 : \mathbf{R}^n \rightarrow \mathbf{R}^m$ be linear transformations.

(i) If α is a real number, the transformation $\alpha \cdot T_1 + T_2$ defined by

$$(\alpha \cdot T_1 + T_2)(\mathbf{x}) = \alpha \cdot T_1(\mathbf{x}) + T_2(\mathbf{x}) \quad \text{for all } \mathbf{x} \text{ in } \mathbf{R}^n$$

is linear.

(ii) The composition $T_2 \circ T_1 : \mathbf{R}^n \rightarrow \mathbf{R}^m$ is linear.

Proof. See Exercise 11.4. □

11.2 Plane Transformations

By Theorem 11.9, if $\mathbf{y}_1 = (a, c)$ and $\mathbf{y}_2 = (b, d)$ are arbitrary elements of \mathbf{R}^2 , there exists a unique linear transformation $T : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ such that $T(1, 0) = (a, c)$ and $T(0, 1) = (b, d)$, and every linear transformation of the plane has this form. In fact, the proof of the theorem gives the explicit formula

$$\begin{aligned} T \begin{bmatrix} x \\ y \end{bmatrix} &= x T \begin{bmatrix} 1 \\ 0 \end{bmatrix} + y T \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= x \begin{bmatrix} a \\ c \end{bmatrix} + y \begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix} \quad \text{for all } \begin{bmatrix} x \\ y \end{bmatrix} \text{ in } \mathbf{R}^2. \end{aligned}$$

Example 11.11. Let c be a real number. The linear transformation $S_c(x, y) = (cx, cy)$ is called *scaling* with scale factor c . If $c \neq 0$, scaling is invertible, and the inverse map is $S_{1/c}$, scaling with factor $1/c$. In particular, $S_1 = I$ is the identity transformation and $S_0 = Z$ is the *zero transformation*, defined by $Z(x, y) = (0, 0)$. The transformation S_{-1} is called a *half-turn* about the origin, or *reflection about the origin*.

We turn next to a notational calculus for evaluating and composing linear transformations of the plane. To highlight algebraic symmetries, denote Cartesian coordinates by (x^1, x^2) rather than by (x, y) .

Definition 11.12. Let A_1^1, A_2^1, A_1^2 , and A_2^2 be real numbers. The array

$$A = \begin{bmatrix} A_1^1 & A_2^1 \\ A_1^2 & A_2^2 \end{bmatrix} = [A_j^i]$$

is called a 2×2 (*real*) *matrix*. The *product* of A with $\mathbf{x} = \begin{bmatrix} x^1 \\ x^2 \end{bmatrix}$ is

$$A\mathbf{x} = \begin{bmatrix} A_1^1 & A_2^1 \\ A_1^2 & A_2^2 \end{bmatrix} \begin{bmatrix} x^1 \\ x^2 \end{bmatrix} = \begin{bmatrix} A_1^1 x^1 + A_2^1 x^2 \\ A_1^2 x^1 + A_2^2 x^2 \end{bmatrix}.$$

The transformation $T_A : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ defined by $T_A(\mathbf{x}) = A\mathbf{x}$ is the *linear transformation with matrix A* .

Remark 11.13. The number A_j^i in the i th row and j th column of A is called the (i, j) *entry* of A . The columns of A are the values

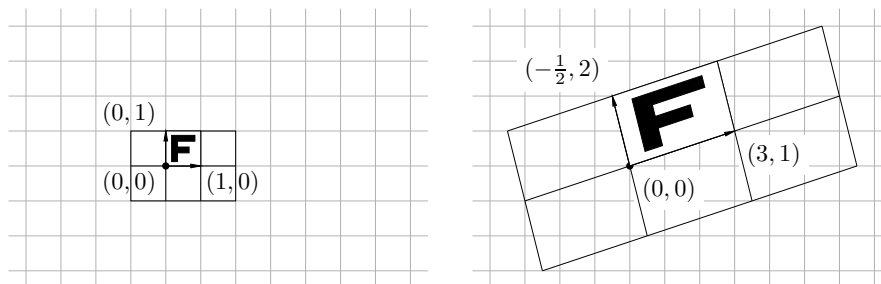
$$T_A \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} A_1^1 \\ A_1^2 \end{bmatrix}, \quad T_A \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} A_2^1 \\ A_2^2 \end{bmatrix}.$$

Example 11.14. The scaling transformation $S_c(\mathbf{x}) = c\mathbf{x}$, the identity transformation $S_1 = I$, and the zero transformation $S_0 = Z$ have respective matrices

$$\begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix}, \quad I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{0}^{2 \times 2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

We call I_2 the (2×2) *identity matrix* and $\mathbf{0}^{2 \times 2}$ the *zero matrix*.

Example 11.15. The matrix $A = \begin{bmatrix} 3 & -\frac{1}{2} \\ 1 & 2 \end{bmatrix}$ acts on the plane as shown:



The unit square (on the left, containing the F) maps to the parallelogram with edges $(3, 1)$ and $(-\frac{1}{2}, 2)$ on the right. With respect to Cartesian coordinates, the mapping T is given by

$$\begin{bmatrix} y^1 \\ y^2 \end{bmatrix} = \begin{bmatrix} 3x^1 - \frac{1}{2}x^2 \\ x^1 + 2x^2 \end{bmatrix}.$$

Example 11.16. Rotation about the origin through an angle θ preserves parallelograms, and is therefore a linear transformation Rot_θ . Since the standard basis vectors map to

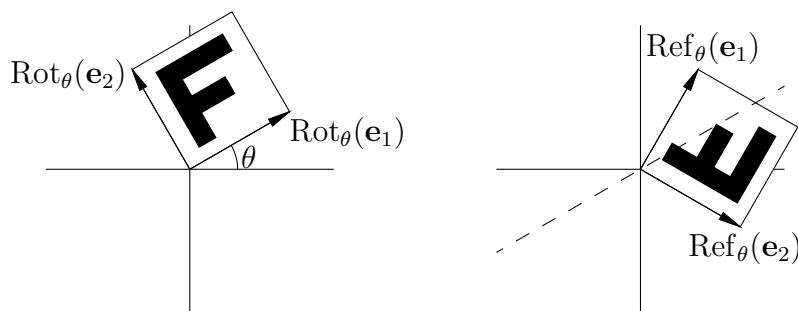
$$\text{Rot}_\theta(\mathbf{e}_1) = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix}, \quad \text{Rot}_\theta(\mathbf{e}_2) = \begin{bmatrix} \cos(\theta + \frac{\pi}{2}) \\ \sin(\theta + \frac{\pi}{2}) \end{bmatrix} = \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix},$$

the transformation Rot_θ has matrix $\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$.

Example 11.17. Let θ be a real number, and let Ref_θ denote reflection across the line through the origin making angle θ with the positive x -axis. Since the standard basis vectors map to

$$\text{Ref}_\theta(\mathbf{e}_1) = \begin{bmatrix} \cos 2\theta \\ \sin 2\theta \end{bmatrix}, \quad \text{Ref}_\theta(\mathbf{e}_2) = \begin{bmatrix} \cos(2\theta - \frac{\pi}{2}) \\ \sin(2\theta - \frac{\pi}{2}) \end{bmatrix} = \begin{bmatrix} \sin 2\theta \\ -\cos 2\theta \end{bmatrix},$$

the transformation Ref_θ has matrix $\begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}$.



Definition 11.18. If $B = [B_k^i]$ and $A = [A_j^k]$ are 2×2 matrices, their *product* is the 2×2 matrix BA defined by

$$(BA)_j^i = \sum_{k=1}^2 B_k^i A_j^k = B_1^i A_j^1 + B_2^i A_j^2.$$

Remark 11.19. The product of 2×2 matrices may be visualized by “partitioning” the left-hand factor into rows and the right-hand factor into columns and “pairing up” entries as shown:

$$\begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a'a + b'c & a'b + b'd \\ c'a + d'c & c'b + d'd \end{bmatrix}.$$

Proposition 11.20. Let $B = [B_k^i]$ and $A = [A_j^k]$ be 2×2 real matrices. The matrix of the composition $T_B \circ T_A$ is the product BA .

Proof. By definition, the equations $\mathbf{z} = T_B(\mathbf{y})$ and $\mathbf{y} = T_A(\mathbf{x})$ mean

$$z^i = \sum_{k=1}^2 B_k^i y^k \quad \text{for } i = 1, 2, \quad y^k = \sum_{j=1}^2 A_j^k x^j \quad \text{for } k = 1, 2.$$

Substituting the second into the first,

$$z^i = \sum_{k=1}^2 B_k^i \left(\sum_{j=1}^2 A_j^k x^j \right) = \sum_{j=1}^2 \left(\sum_{k=1}^2 B_k^i A_j^k \right) x^j = \sum_{j=1}^2 (BA)_j^i x^j. \quad \square$$

Remark 11.21. Written without summation notation, the preceding computation is

$$\begin{aligned} \begin{bmatrix} z^1 \\ z^2 \end{bmatrix} &= \begin{bmatrix} B_1^1 & B_2^1 \\ B_1^2 & B_2^2 \end{bmatrix} \begin{bmatrix} y^1 \\ y^2 \end{bmatrix} = \begin{bmatrix} B_1^1 y^1 + B_2^1 y^2 \\ B_1^2 y^1 + B_2^2 y^2 \end{bmatrix}, \\ \begin{bmatrix} y^1 \\ y^2 \end{bmatrix} &= \begin{bmatrix} A_1^1 & A_2^1 \\ A_1^2 & A_2^2 \end{bmatrix} \begin{bmatrix} x^1 \\ x^2 \end{bmatrix} = \begin{bmatrix} A_1^1 x^1 + A_2^1 x^2 \\ A_1^2 x^1 + A_2^2 x^2 \end{bmatrix}. \end{aligned}$$

Substituting the second into the first,

$$\begin{aligned} \begin{bmatrix} z^1 \\ z^2 \end{bmatrix} &= \begin{bmatrix} B_1^1(A_1^1x^1 + A_2^1x^2) + B_2^1(A_1^2x^1 + A_2^2x^2) \\ B_1^2(A_1^1x^1 + A_2^1x^2) + B_2^2(A_1^2x^1 + A_2^2x^2) \end{bmatrix} \\ &= \begin{bmatrix} (B_1^1A_1^1 + B_2^1A_1^2)x^1 + (B_1^1A_2^1 + B_2^1A_2^2)x^2 \\ (B_1^2A_1^1 + B_2^2A_1^2)x^1 + (B_1^2A_2^1 + B_2^2A_2^2)x^2 \end{bmatrix} \\ &= \begin{bmatrix} B_1^1A_1^1 + B_2^1A_1^2 & B_1^1A_2^1 + B_2^1A_2^2 \\ B_1^2A_1^1 + B_2^2A_1^2 & B_1^2A_2^1 + B_2^2A_2^2 \end{bmatrix} \begin{bmatrix} x^1 \\ x^2 \end{bmatrix}. \end{aligned}$$

The relative messiness should convince you of the power of summation notation, and the importance of being able to use it.

Corollary 11.22. *Matrix multiplication is associative as a binary operation on the set $\mathbf{R}^{2 \times 2}$ of real 2×2 matrices.*

Proof. Let C , B , and A be 2×2 matrices. Mapping composition is associative by Proposition 4.31, so $(T_C \circ T_B) \circ T_A = T_C \circ (T_B \circ T_A)$. By the theorem, these transformations have matrices $(CB)A$ and $C(BA)$, respectively. \square

Corollary 11.23. *If θ is a real number, then*

$$\begin{aligned} \cos(2\theta) &= \cos^2 \theta - \sin^2 \theta, \\ \sin(2\theta) &= 2 \cos \theta \sin \theta. \end{aligned}$$

Proof. The composition of the rotation Rot_θ with itself is $\text{Rot}_{2\theta}$. Multiplying matrices and equating entries,

$$\begin{aligned} \begin{bmatrix} \cos(2\theta) & -\sin(2\theta) \\ \sin(2\theta) & \cos(2\theta) \end{bmatrix} &= \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \\ &= \begin{bmatrix} \cos^2 \theta - \sin^2 \theta & -2 \cos \theta \sin \theta \\ 2 \cos \theta \sin \theta & \cos^2 \theta - \sin^2 \theta \end{bmatrix}. \quad \square \end{aligned}$$

Remark 11.24. The entries of a matrix may be residue classes, complex numbers, or generally any entities that can be “added” and “multiplied”. The operation of matrix multiplication is associative provided addition and multiplication are associative and multiplication distributes over addition.

Example 11.25. Every complex matrix can be written uniquely as a real matrix plus i times a real matrix. For example,

$$\begin{bmatrix} 1+i & \sqrt{3}i \\ -\sqrt{3}i & 1-i \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} i & \sqrt{3}i \\ -\sqrt{3}i & -i \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + i \begin{bmatrix} 1 & \sqrt{3} \\ -\sqrt{3} & -1 \end{bmatrix}.$$

Definition 11.26. The *determinant* of a 2×2 matrix is the expression

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc.$$

Remark 11.27. A 2×2 real or complex matrix A has non-zero determinant if and only if the columns of A are not proportional.

Proposition 11.28. If A, B are 2×2 , then $\det(AB) = (\det A)(\det B)$.

Proof. Exercise 11.10. □

Definition 11.29. A 2×2 matrix A is *invertible* if there exists a 2×2 matrix B satisfying $AB = I_2$ and $BA = I_2$.

Remark 11.30. Since matrix multiplication is associative, an invertible matrix has a unique inverse, denoted A^{-1} .

Proposition 11.31. A 2×2 real or complex matrix A is invertible if and only if $\det A = ad - bc \neq 0$, and

$$A^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Proof. If A is invertible, there exists a matrix B such that $AB = I_2$. By Proposition 11.28, $(\det A)(\det B) = \det I_2 = 1$, so $\det A \neq 0$.

Conversely, if $\det A \neq 0$, it suffices to observe that in the notation of the proposition, $AA^{-1} = I_2$ and $A^{-1}A = I_2$. □

Definition 11.32. The set of invertible real 2×2 matrices, denoted $GL(2, \mathbf{R})$, forms a group under matrix multiplication, the *general linear group* of the plane.

By Proposition 11.28, the set of real 2×2 matrices with determinant 1, denoted $SL(2, \mathbf{R})$, forms a subgroup, the *special linear group*.

Remark 11.33. Similarly, the set $GL(2, \mathbf{C})$ of invertible 2×2 complex matrices and the set $SL(2, \mathbf{C})$ of complex matrices of unit determinant are groups under matrix multiplication.

Example 11.34. Let a and b be real numbers. The 2×2 real matrix

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} = a \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + b \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = aI + bJ$$

“represents” the complex number $a+bi$, see Exercise 11.16. This matrix is invertible if and only if $\det A = a^2 + b^2 \neq 0$ (the superscripts denote squaring), i.e., a and b are not both zero, and the inverse is

$$A^{-1} = \frac{1}{a^2 + b^2} \begin{bmatrix} a & b \\ -b & a \end{bmatrix}.$$

This is the matrix version of the formula for the reciprocal of a non-zero complex number, compare Example 2.29.

The set of invertible matrices of this type forms an Abelian subgroup of the general linear group.

11.3 Cartesian Transformations

Much of the preceding section generalizes immediately to linear transformations from \mathbf{R}^n to \mathbf{R}^m ; the only extra burden is to keep track of the sizes of matrices.

By Theorem 11.9, a linear transformation $T : \mathbf{R}^n \rightarrow \mathbf{R}^m$ is uniquely determined by an ordered set of n vectors in \mathbf{R}^m .

Definition 11.35. Let m and n be positive integers, and let A_j^i be real numbers for $i = 1, \dots, m$ and $j = 1, 2, \dots, n$. The array

$$A = \begin{bmatrix} A_1^1 & A_2^1 & A_3^1 & \cdots & A_n^1 \\ A_1^2 & A_2^2 & A_3^2 & \cdots & A_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_1^m & A_2^m & A_3^m & \cdots & A_n^m \end{bmatrix} = [A_j^i]$$

is called a *real matrix* of size $m \times n$. The *product* of A with a vector $\mathbf{x} = [x^j]$ in \mathbf{R}^n is the vector $\mathbf{y} = [y^i]$ in \mathbf{R}^m with components

$$y^i = \sum_{j=1}^n A_j^i x^j = A_1^i x^1 + A_2^i x^2 + \cdots + A_n^i x^n, \quad i = 1, 2, \dots, m.$$

The transformation $T_A : \mathbf{R}^n \rightarrow \mathbf{R}^m$ defined by $T_A(\mathbf{x}) = A\mathbf{x}$ is the *linear transformation with matrix A* .

Remark 11.36. The (i, j) entry A_j^i is the i th component of $T(\mathbf{e}_j)$. That is, the j th column of A is $A_j = T(\mathbf{e}_j)$, and for all \mathbf{x} in \mathbf{R}^n ,

$$T(\mathbf{x}) = \sum_{j=1}^n x^j A_j = \sum_{j=1}^n x^j \left(\sum_{i=1}^m A_j^i \mathbf{e}_i \right) = \sum_{i=1}^m \left(\sum_{j=1}^n A_j^i x^j \right) \mathbf{e}_i.$$

Remark 11.37. The following *computational procedure* is useful in practice: Mentally divide the matrix A into rows. To compute the i th row of $A\mathbf{x}$, multiply each entry in the i th row of A by the corresponding component of \mathbf{x} , and sum the products to get y^i :

$$\begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^m \end{bmatrix} = \begin{bmatrix} \hline A_1^1 & A_2^1 & A_3^1 & \cdots & A_n^1 \\ A_1^2 & A_2^2 & A_3^2 & \cdots & A_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline A_1^m & A_2^m & A_3^m & \cdots & A_n^m \end{bmatrix} \begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^n \end{bmatrix}$$

By contrast, the *geometric meaning* of matrix multiplication is to take the linear combination of the columns of A using the x^j as coefficients.

Remark 11.38. If $\mathbf{y} = A\mathbf{x}$ expresses the variables y^i as functions of the x^j , the entry A_j^i in the i th row and j th column measures the change $\Delta y^i = A_j^i \Delta x^j$ in the i th output upon making a small change Δx^j in the j th input while holding the remaining variables constant. In the language of calculus, A_j^i is the partial derivative of y^i with respect to x^j .

Definition 11.39. Let m and n be positive integers.

An $m \times 1$ matrix is a *column matrix*. A $1 \times n$ matrix is a *row matrix*. An $n \times n$ matrix (having as many rows as columns) is *square*.

The set of all $m \times n$ real matrices is denoted $\mathbf{R}^{m \times n}$. The binary operation of *matrix addition* on $\mathbf{R}^{m \times n}$ is defined “entrywise”, by

$$A + B = [A_j^i] + [B_j^i] = [A_j^i + B_j^i].$$

If c is a real number, *scalar multiplication* (by c) sends A to the $m \times n$ matrix $cA = [cA_j^i]$ obtained by multiplying each entry by c .

Remark 11.40. A matrix sum is undefined if the summands do not have the same size.

Remark 11.41. Since addition of numbers is associative and commutative, matrix addition is associative and commutative as well.

Matrix addition on $\mathbf{R}^{m \times n}$ has an identity element, the $m \times n$ *zero matrix*, whose entries are all 0. Every $m \times n$ matrix $A = [A_j^i]$ has an *additive inverse* $-A = -[A_j^i] = [-A_j^i]$, obtained by taking the negative of each entry. The pair $(\mathbf{R}^{m \times n}, +)$ is therefore an Abelian group.

Remark 11.42. The set $\mathbf{C}^{m \times n}$ of $m \times n$ *complex* matrices similarly admits an addition operation and multiplication by *complex* scalars.

Definition 11.43. If $B = [B_k^i] \in \mathbf{R}^{m \times p}$ and $A = [A_j^k] \in \mathbf{R}^{p \times n}$ are matrices, their *product* is the $m \times n$ matrix BA defined by

$$(11.2) \quad (BA)_j^i = \sum_{k=1}^p B_k^i A_j^k = B_1^i A_j^1 + B_2^i A_j^2 + \cdots + B_p^i A_j^p.$$

Proposition 11.44. Let $B = [B_k^i]$ be a real $m \times p$ matrix and $A = [A_j^k]$ a real $p \times n$ matrix. The matrix of the composition $T_B \circ T_A$ is BA .

Proof. By definition, the equations $\mathbf{z} = T_B(\mathbf{y})$ and $\mathbf{y} = T_A(\mathbf{x})$ mean

$$z^i = \sum_{k=1}^p B_k^i y^k, \quad 1 \leq i \leq m, \quad y^k = \sum_{j=1}^n A_j^k x^j, \quad 1 \leq k \leq p.$$

Substituting the second into the first,

$$z^i = \sum_{k=1}^p B_k^i \left(\sum_{j=1}^n A_j^k x^j \right) = \sum_{j=1}^n \left(\sum_{k=1}^p B_k^i A_j^k \right) x^j = \sum_{j=1}^n (BA)_j^i x^j. \quad \square$$

Corollary 11.45. Matrix multiplication is associative. Precisely, if $C \in \mathbf{R}^{m \times p}$, $B \in \mathbf{R}^{p \times q}$, and $A \in \mathbf{R}^{q \times n}$, then $(CB)A = C(BA)$.

Remark 11.46. The hypotheses on the sizes merely ensure the products CB and BA are defined. A computational proof using Equation (11.2) is left to you, see Exercise 11.6.

Proof. Let $T_A : \mathbf{R}^n \rightarrow \mathbf{R}^q$, $T_B : \mathbf{R}^q \rightarrow \mathbf{R}^p$, and $T_C : \mathbf{R}^p \rightarrow \mathbf{R}^m$ be the linear transformations with respective matrices A , B , and C . Since composition of mappings is associative, $(T_C \circ T_B) \circ T_A = T_C \circ (T_B \circ T_A)$ as linear transformations from \mathbf{R}^n to \mathbf{R}^m . But the respective matrices are $(CB)A$ and $C(BA)$. \square

Definition 11.47. The *transpose* of an $m \times n$ matrix $A = [A_j^i]$ is the $n \times m$ matrix A^T whose (i, j) entry is the (j, i) entry of A : $(A^T)_j^i = A_i^j$.

Example 11.48. $\begin{bmatrix} a^1 & a^2 & a^3 \\ b^1 & b^2 & b^3 \end{bmatrix}^T = \begin{bmatrix} a^1 & b^1 \\ a^2 & b^2 \\ a^3 & b^3 \end{bmatrix}$, and *vice versa*.

Remark 11.49. The transpose of a column is a row and *vice versa*.

Example 11.50. If $B = \begin{bmatrix} \text{repo} & \text{robo} \end{bmatrix}$ and $A = \begin{bmatrix} \text{man} & \text{cop} \end{bmatrix}$, then

$$B(A^\top) = \begin{bmatrix} \text{repo} & \text{robo} \end{bmatrix} \begin{bmatrix} \text{man} \\ \text{cop} \end{bmatrix} = \begin{bmatrix} \text{repoman} + \text{robocop} \end{bmatrix},$$

$$(B^\top)A = \begin{bmatrix} \text{repo} \\ \text{robo} \end{bmatrix} \begin{bmatrix} \text{man} & \text{cop} \end{bmatrix} = \begin{bmatrix} \text{repoman} & \text{repocop} \\ \text{roboman} & \text{robocop} \end{bmatrix}.$$

One product is a double feature, the other is an array of four movies, three of which exist at this writing.

Proposition 11.51. If B and A are $n \times n$, then $(BA)^\top = A^\top B^\top$.

Proof. Exercise 11.8. □

Definition 11.52. A square matrix A is *symmetric* if $A^\top = A$, and is *skew-symmetric* if $A^\top = -A$.

Example 11.53. If $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then $A^\top = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$. The matrices

$$\frac{1}{2}(A + A^\top) = \begin{bmatrix} a & \frac{1}{2}(b + c) \\ \frac{1}{2}(c + b) & d \end{bmatrix}, \quad \frac{1}{2}(A - A^\top) = \begin{bmatrix} 0 & \frac{1}{2}(b - c) \\ \frac{1}{2}(c - b) & 0 \end{bmatrix}$$

are symmetric and skew-symmetric, respectively.

The General Linear Group

Since a product of $n \times n$ matrices has size $n \times n$, matrix multiplication defines an associative binary operation on $\mathbf{R}^{n \times n}$. Exercises 11.20 ff. explore the extent to which matrix multiplication is not commutative.

Definition 11.54. The matrix of the identity map $I : \mathbf{R}^n \rightarrow \mathbf{R}^n$ is the $n \times n$ *identity matrix*

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

The (i, j) entry of I_n is the *Kronecker delta symbol*

$$\delta_j^i = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Proposition 11.55. *The identity matrix I_n is the identity element for matrix multiplication on $\mathbf{R}^{n \times n}$.*

Proof. This is immediate from Proposition 11.44. For illustration, we give a proof using the formula for matrix multiplication.

Let $A = [A_j^k]$ be an arbitrary $n \times n$ matrix. By equation (11.2) with $B_k^i = I_k^i = \delta_k^i$, the (i, j) entry of $I_n A$ is

$$(I_n A)_j^i = \sum_{k=1}^n \delta_k^i A_j^k.$$

If $k \neq i$, the summand $\delta_k^i A_j^k$ is zero. The term with $k = i$ is equal to $\delta_i^i A_j^i = A_j^i$ since $\delta_i^i = 1$. Summing over k , we find the (i, j) entry of $I_n A$ is A_j^i , the (i, j) entry of A : $I_n A = A$ for all A in $\mathbf{R}^{n \times n}$. An entirely similar argument shows $A I_n = A$ for all A . \square

Definition 11.56. An $n \times n$ matrix A is *invertible* if there exists an $n \times n$ matrix B such that $BA = I_n$ and $AB = I_n$.

The set $GL(n, \mathbf{R})$ of *invertible* $n \times n$ real matrices forms a group under matrix multiplication, the *general linear group*.

Remark 11.57. An $n \times n$ matrix A has a determinant, a polynomial in the entries of A . It turns out that $\det(BA) = (\det B)(\det A)$ for $n \times n$ matrices, and a matrix A is invertible if and only if its determinant is non-zero. Finally, there is a formula for A^{-1} , with each entry expressed as a rational function of the entries of A .

Though these facts parallel the situation for 2×2 matrices, their proofs are not straightforward generalizations: $\det A$ is a polynomial with $n!$ summands, each a monomial of degree n , so the corresponding formulas are too complex to be much practical use. Instead, one develops computationally efficient *algorithms* for finding the determinant and inverse of a square matrix.

Definition 11.58. Let \mathbf{x}_0 be an arbitrary element of \mathbf{R}^n , and let \mathbf{v} be a non-zero element. The set of elements of the form $\mathbf{x}_0 + t\mathbf{v}$ for some real number t is the *line* through \mathbf{x}_0 with direction \mathbf{v} .

Lemma 11.59. *Let \mathbf{x}_0 and \mathbf{x}_1 be elements of \mathbf{R}^n , and let \mathbf{v}_0 and \mathbf{v}_1 be non-zero.*

- (i) *The line ℓ_0 through \mathbf{x}_0 with direction \mathbf{v}_0 is equal to the line ℓ_1 through \mathbf{x}_1 with direction \mathbf{v}_1 if and only if \mathbf{x}_1 lies on ℓ_0 and there exists a non-zero real number s such that $\mathbf{v}_1 = s\mathbf{v}_0$.*

(ii) If $\mathbf{x}_1 \neq \mathbf{x}_0$, there exists a unique line ℓ containing both points.

Proof. (i). If $\ell_0 = \ell_1$, then in particular, the points $\mathbf{x}_1 + \mathbf{v}_1$ and \mathbf{x}_1 are on ℓ_0 , so there exist real numbers t and t_0 such that

$$\begin{aligned}\mathbf{x}_1 + \mathbf{v}_1 &= \mathbf{x}_0 + t\mathbf{v}_0, \\ \mathbf{x}_1 &= \mathbf{x}_0 + t_0\mathbf{v}_0.\end{aligned}$$

Subtracting the second from the first, $\mathbf{v}_1 = (t - t_0)\mathbf{v}_0$, and since $\mathbf{v}_1 \neq \mathbf{0}$, it must be that $s := t - t_0 \neq 0$.

Conversely, assume \mathbf{x}_1 lies on ℓ_0 and there exists a non-zero real number s such that $\mathbf{v}_1 = s\mathbf{v}_0$. By definition there exists a real number t_0 such that $\mathbf{x}_1 = \mathbf{x}_0 + t_0\mathbf{v}_0$, so for all real t ,

$$\mathbf{x}_1 + t\mathbf{v}_1 = (\mathbf{x}_0 + t_0\mathbf{v}_0) + t(s\mathbf{v}_0) = \mathbf{x}_0 + (t_0 + ts)\mathbf{v}_0 \in \ell_0.$$

By definition, $\ell_1 \subseteq \ell_0$. Further, $\mathbf{x}_0 = \mathbf{x}_1 - t_0\mathbf{v}_0$ and $\mathbf{v}_0 = \frac{1}{s}\mathbf{v}_1$, so

$$\mathbf{x}_0 + t\mathbf{v}_0 = (\mathbf{x}_1 - t_0\mathbf{v}_0) + t\mathbf{v}_0 = \mathbf{x}_1 + \frac{1}{s}(t - t_0)\mathbf{v}_1 \quad \text{for all real } t,$$

proving $\ell_0 \subseteq \ell_1$. This completes the proof of (i).

(ii). If $\mathbf{x}_1 \neq \mathbf{x}_0$, then $\mathbf{v}_0 = \mathbf{x}_1 - \mathbf{x}_0$ is non-zero, and \mathbf{x}_1 lies on the line ℓ_0 through \mathbf{x}_0 with direction \mathbf{v}_0 . To prove this line is unique, suppose \mathbf{v} is a non-zero vector such that \mathbf{x}_1 lies on the line ℓ through \mathbf{x}_0 with direction \mathbf{v} . Particularly, $\mathbf{x}_1 = \mathbf{x}_0 + s\mathbf{v}$ for some real number s , necessarily non-zero, which means $\mathbf{v}_0 = s\mathbf{v}$. By (i), $\ell = \ell_0$. \square

Remark 11.60. A general linear transformation induces a bijection of the set of lines in \mathbf{R}^n , see Exercise 11.19.

Exercises

All matrices are assumed to have real entries unless specified otherwise.

Exercise 11.1. If $B = \begin{bmatrix} \text{basic} & \text{fatal} \end{bmatrix}$ and $A = \begin{bmatrix} \text{instinct} & \text{attraction} \end{bmatrix}$, compute the matrix products $B(A^\top)$ and $(B^\top)A$.

Exercise 11.2. Find a 2×3 matrix A and a 3×2 matrix B such that $AB = I_2$ but $BA \neq I_3$. Suggestion: Start with 2×2 identity matrices, and “pad” them suitably.

Exercise 11.3. Let A and B be matrices such that the products AB and BA are both defined. What can you say about the sizes of A and B ?

Exercise 11.4. Prove Theorem 11.10. That is, suppose T_1 and T_2 are linear transformations of \mathbf{R}^n .

- (a) If α is a real number, the transformation $\alpha \cdot T_1 + T_2 : \mathbf{R}^n \rightarrow \mathbf{R}^n$ defined by

$$(\alpha \cdot T_1 + T_2)(\mathbf{x}) = \alpha \cdot T_1(\mathbf{x}) + T_2(\mathbf{x}) \quad \text{for all } \mathbf{x} \text{ in } \mathbf{R}^n$$

is linear.

- (b) The composition $T_2 \circ T_1 : \mathbf{R}^n \rightarrow \mathbf{R}^n$ is linear.

Exercise 11.5. Let $I = I_n$ be the $n \times n$ identity matrix. Modify the proof of Proposition 11.55 to prove:

- (a) $AI = A$ for all A in $\mathbf{R}^{m \times n}$.
 (b) $IA = A$ for all A in $\mathbf{R}^{n \times p}$.

Exercise 11.6. Use Equation (11.2) to prove Corollary 11.45.

Hint: Write $C = [C_k^i]$, $B = [B_\ell^k]$, and $A = [A_j^\ell]$, and show the (i, j) entries of $(CB)A$ and $C(BA)$ are each equal to

$$\sum_{k=1}^p \sum_{\ell=1}^q C_k^i B_\ell^k A_j^\ell.$$

Exercise 11.7. (a) Prove matrix multiplication distributes over addition: If A is $m \times p$ and B, C are $p \times n$, then $A(B + C) = AB + AC$.

- (b) Formulate and prove the analogous claim for multiplication on the right.

Exercise 11.8. Prove Proposition 11.51: If B and A are $n \times n$ matrices, then $(BA)^\top = A^\top B^\top$.

Exercise 11.9. Let A and B be *symmetric* $n \times n$ matrices. Prove BA is symmetric if and only if A and B commute.

Formulate and prove analogous assertions if one or both matrices are instead *skew-symmetric*.

Exercise 11.10. Let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} w & x \\ y & z \end{bmatrix}.$$

Prove $\det(AB) = \det A \det B$ by direct calculation.

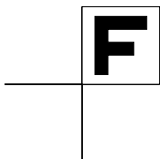
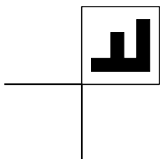
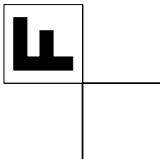
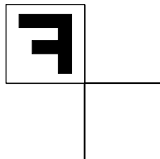
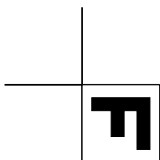
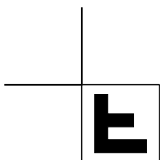
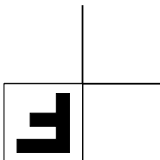
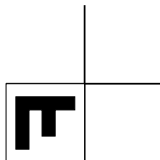
Exercise 11.11. Use Exercise 11.10 to prove the following sets of 2×2 matrices are groups under matrix multiplication:

- (a) The set $GL(2, \mathbf{R})$ of real matrices with non-zero determinant.
- (b) The set $SL(2, \mathbf{R})$ of real matrices with determinant 1.
- (c) The set $SL(2, \mathbf{Z})$ of integer matrices with determinant 1.

Exercise 11.12. Prove that an integer matrix A in $\mathbf{Z}^{2 \times 2}$ is invertible if and only if $\det A = \pm 1$.

Exercise 11.13. Eight matrices and eight transformations are given below. Match each matrix with the corresponding image of the unit F , and compute $T_A(\mathbf{x})$ for each A .

(i) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	(ii) $\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$	(iii) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$	(iv) $\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$
(v) $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$	(vi) $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$	(vii) $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$	(viii) $\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$

Exercise 11.14. Let J be the 2×2 real matrix inducing the linear transformation $T_J(x, y) = (-y, x)$. Describe the action of T_J geometrically, and show that J generates a cyclic group of order 4 under matrix multiplication.

Exercise 11.15. Consider the linear transformations $H(x, y) = (x, -y)$ and $V(x, y) = (-x, y)$. Describe the actions of H and V geometrically, and show that $\{I, H, V, H \circ V\}$ is a non-cyclic group under mapping composition.

Exercise 11.16. Consider the 2×2 matrices $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, and let $M\mathbf{C}$ be the set of matrices of the form

$$A = aI + a'J = \begin{bmatrix} a & -a' \\ a' & a \end{bmatrix}, \quad a, a' \text{ real.}$$

- (a) Show $J^2 = -I$.
- (b) Show $M\mathbf{C}$ is closed under addition.
- (c) Show $M\mathbf{C}$ is closed under matrix multiplication. Give two proofs: One by direct calculation, and one using the distributive law for matrix multiplication.
- (d) Show $AB = BA$ for all A and B in $M\mathbf{C}$. Again, give two proofs.
- (e) Let $\alpha = a + a'i$ and $\beta = b + b'i$ be complex numbers, and let $A = aI + a'J$ and $B = bI + b'J$ be the corresponding matrices. Compute the sum and product of the complex numbers α and β , and the sum and product of the matrices A and B . What do you observe about the respective sums and products?

Define $\phi : \mathbf{C} \rightarrow M\mathbf{C}$ by the formula $\phi(a + a'i) = aI + a'J$, and show that your conclusions can be expressed as

$$\phi(\alpha + \beta) = \phi(\alpha) + \phi(\beta), \quad \phi(\alpha \cdot \beta) = \phi(\alpha) \cdot \phi(\beta).$$

Exercise 11.17. On $\mathbf{R}^{2 \times 2}$, the transpose operator takes the form

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix}.$$

This question investigates *orthogonal* 2×2 matrices, whose inverse is equal to their transpose. Let $O(2)$ be the set of 2×2 real matrices A such that $A^{-1} = A^T$, and let \mathcal{R} denote the set of real matrices of either form

$$\text{Rot}_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \quad \text{Ref}_\theta = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}.$$

- (a) Show $\det \text{Rot}_\theta = 1$ and $\det \text{Ref}_\theta = -1$.

- (b) Prove that if $A \in \mathcal{R}$, then $A^{-1} = A^T$, i.e., that $A^T A = A A^T = I_2$.
- (c) Conversely, suppose $A^{-1} = A^T$. Prove $\det A = \pm 1$.
Hint: Use Exercise 11.10.
- (d) Suppose $A^{-1} = A^T$. Prove $A \in \mathcal{R}$.
Suggestion: Set $A A^T = I_2$, and use the fact that if $a^2 + b^2 = 1$, there exists a real number θ such that $a = \cos \theta$ and $b = \sin \theta$.
- (e) For each matrix A in Exercise 11.13, determine whether A is of the form Rot_θ or Ref_θ , and find the corresponding value of θ .

Exercise 11.18. Show that if θ and ϕ are arbitrary real numbers, then

$$\begin{aligned}\cos(\theta + \phi) &= \cos \theta \cos \phi - \sin \theta \sin \phi, \\ \sin(\theta + \phi) &= \cos \theta \sin \phi + \sin \theta \cos \phi.\end{aligned}$$

Suggestion: Follow the proof of Corollary 11.23.

Exercise 11.19. Let $T : \mathbf{R}^n \rightarrow \mathbf{R}^n$ be an invertible linear transformation. Show that T maps lines to lines, and that the induced map on lines is bijective.

Exercise 11.20. Let A and B be square matrices of the same size. Define the *commutator* of A and B to be the matrix $[A, B] = AB - BA$.

- (a) Prove $[B, A] = -[A, B]$.
- (b) Prove A and B commute if and only if $[A, B] = 0$.

Exercise 11.21. Let $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$.

- (a) Compute AB , BA , and $[A, B] = AB - BA$.
- (b) Describe the set of matrices commuting with A .

Suggestion: Write $C = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, compute $[A, C]$, and eliminate as many variables as you can.

Exercise 11.22. Let a and b be numbers, and introduce the *horizontal shear* and *vertical shear* matrices

$$H_a = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}, \quad V_b = \begin{bmatrix} 1 & 0 \\ b & 1 \end{bmatrix}.$$

- (a) Show that the set $\{H_a : a \text{ real}\}$ is an Abelian group under matrix multiplication, and sketch the effect of H_a on the unit square.

(b) Calculate the commutator $[H_a, V_b]$.

Exercise 11.23. In each part, let $J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$.

- (a) Describe the set of 2×2 real matrices A that commute with J .
- (b) Describe the set of 2×2 real matrices A that anti-commute with J , i.e., that satisfy $AJ + JA = \mathbf{0}^{2 \times 2}$.

Exercise 11.24. Assume $A \in \mathbf{R}^{2 \times 2}$, and suppose $AB = BA$ for every B in $\mathbf{R}^{2 \times 2}$. Prove there exists a real number α such that $A = \alpha I_2$.
Hint: By hypothesis, A commutes with the four matrices

$$E_1^1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad E_2^1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad E_1^2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad E_2^2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Exercise 11.25. An $n \times n$ matrix A is a *scalar matrix* if there exists a real number c such that $A = cI_n$.

- (a) Prove that a scalar matrix commutes with every $n \times n$ matrix under matrix multiplication.
- (b) Conversely, suppose $AB = BA$ for every B in $\mathbf{R}^{n \times n}$. Prove A is a scalar matrix.
Hint: Fix indices k and ℓ , and consider the commutator of A with the matrix $E_{k\ell}$ whose (k, ℓ) entry is 1 and whose other entries are zero, namely, whose (i, j) entry is $\delta_k^i \delta_\ell^j$ (cf. Exercise 11.24).

Exercise 11.26. Let A , B , and C be square matrices of the same size. (This ensures the product of any two is defined.)

- (a) Expand the double commutator $[A, [B, C]]$. (See Exercise 11.20.)
- (b) Prove the *Jacobi identity*:

$$[A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0.$$

Suggestion: Re-use the result of part (a).

- (c) If $A \in \mathbf{R}^{n \times n}$, define $D_A(B) = [A, B]$. Prove the *Leibniz rule*

$$D_A[B, C] = [D_A(B), C] + [B, D_A(C)].$$

Suggestion: Rearrange the Jacobi identity and use Exercise 11.20.

Exercise 11.27. An $n \times n$ matrix $A = [A_j^i]$ is *diagonal* if $A_j^i = 0$ when $i \neq j$, that is, if there exist numbers A_1, A_2, \dots, A_n such that $A_j^i = A_i \delta_j^i$. For brevity, we write $A = \text{diag}[A_1, A_2, \dots, A_n]$.

- Write out the general 2×2 and 3×3 diagonal matrices.
- Show the set of diagonal $n \times n$ matrices is closed under addition.
- Show the set of diagonal $n \times n$ matrices is closed under matrix multiplication, and find a formula for the product of two diagonal matrices of the same size.
- Determine when a diagonal matrix A in $\mathbf{R}^{n \times n}$ is invertible, and find a formula for the inverse.

Exercise 11.28. If θ is a real number, the matrix

$$\text{Rot}_{\hat{\theta}}^z = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

corresponds to counterclockwise rotation by θ about the z -axis. (That is, the z -axis is fixed, and the rotation is counterclockwise when looking “down” along the positive z -axis.)

- Write down matrices for counterclockwise rotation through an angle θ about the y -axis, or about the x -axis. (A sketch of the coordinate axes should help you decide which axis rotates “toward” which axis.)
- Calculate the commutator of the matrices you found in part (a), and the commutator of each with $\text{Rot}_{\hat{\theta}}^z$.
- Let $R^x = \text{Rot}_{\pi/2}^x$, $R^y = \text{Rot}_{\pi/2}^y$, and $R^z = \text{Rot}_{\pi/2}^z$ be the quarter-turns about the coordinate axes. Calculate the nine products of pairs of these matrices.

Exercise 11.29. An $n \times n$ real matrix A is *orthogonal* if $A^T = A^{-1}$, compare Exercise 11.17. Prove that the set $O(n)$ of $n \times n$ real orthogonal matrices is a group under matrix multiplication (the *orthogonal group*).

Exercise 11.30. Let \mathcal{H} be the set of real 3×3 matrices of the form

$$\begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix}, \quad a, b, c \text{ real.}$$

- (a) Show \mathcal{H} is a group under matrix multiplication (the *Heisenberg group*).
- (b) Show that the set of elements of \mathcal{H} having *integer* entries is a subgroup.

Exercise 11.31. In each part, let

$$A = \begin{bmatrix} [1] & [1] \\ [1] & [2] \end{bmatrix}, \quad B = \begin{bmatrix} [2] & [1] \\ [0] & [2] \end{bmatrix},$$

noting carefully that the modulus has not yet been specified.

(a) Calculate the products AB and BA assuming $n = 3$ (i.e., the entries are elements of \mathbf{Z}_3). Is either matrix A or B invertible? If so, find the inverse(s).

(b) Calculate the products AB and BA assuming $n = 4$. Is either matrix A or B invertible? If so, find the inverse(s).

Exercise 11.32. Let $n > 1$ be an integer, and assume $A \in \mathbf{Z}_n^{2 \times 2}$ is a 2×2 matrix with entries in \mathbf{Z}_n .

(a) If n is prime, show A is invertible if and only if $\det A \neq [0]$, and find a formula for A^{-1} .

(b) Find a necessary and sufficient condition for invertibility of a matrix A in $\mathbf{Z}_n^{2 \times 2}$ in terms of $\det A$ and n . Give a formula for A^{-1} if your condition is satisfied, and find an example of a non-invertible matrix A in $\mathbf{Z}_6^{2 \times 2}$ such that $\det A \neq [0]$.

Exercise 11.33. Let $GL(2, \mathbf{Z}_2)$ denote the set of invertible 2×2 matrices with entries in $\mathbf{Z}_2 = \{0, 1\}$.

(a) Prove $GL(2, \mathbf{Z}_2)$ is a group under matrix multiplication. (Suggestion: Use Exercises 11.32 (a) and 11.10.)

(b) List the elements of $GL(2, \mathbf{Z}_2)$, and find the order of each. Hints: How many elements does $\mathbf{Z}_2^{2 \times 2}$ have? How few/many “1” entries can an invertible matrix have?

Exercise 11.34. This exercise introduces a particularly interesting class of 2×2 complex matrices representing the *quaternions* in the

same way the class of 2×2 real matrices in Example 11.34 represented complex numbers.

Consider the 2×2 complex matrices*

$$\mathbf{1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{i} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{j} = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}, \quad \mathbf{k} = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}.$$

For real numbers a , b , c , and d , define

$$A = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} = \begin{bmatrix} a + ci & -b + di \\ b + di & a - ci \end{bmatrix}.$$

Let $\mathbf{H} \subseteq \mathbf{C}^{2 \times 2}$ be the set of all such matrices.

- Compute $\det A$ (use the same formula as for a 2×2 real matrix) and show $\det A$ is real and non-negative.
- Find a formula for A^{-1} provided $A \neq 0$.
- Show $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -\mathbf{1}$ and $\mathbf{i} \cdot \mathbf{j} = \mathbf{k}$ by direct calculation. (Thus

$$a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} = a\mathbf{1} + b\mathbf{i} + (c\mathbf{1} + d\mathbf{i})\mathbf{j},$$

and $\mathbf{i}^2 \cdot \mathbf{j}^2 \neq \mathbf{k}^2$, which in turn means $\mathbf{i} \cdot \mathbf{j} \neq \mathbf{j} \cdot \mathbf{i}$ by Example 7.17.)

- As we saw when we divided a matrix into columns or into rows, a matrix can have other matrices as entries, provided the “sub-matrices” stack up into a rectangular array. By writing complex numbers as 2×2 real matrices, we may represent the Pauli spin matrices as 4×4 real matrices. For example,

$$\mathbf{i} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \longleftrightarrow \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Find the 4×4 real matrices corresponding to $\mathbf{1}$, \mathbf{j} , \mathbf{k} , and A .

Exercise 11.35. If $A = [A_{ij}^i]$ in $\mathbf{C}^{n \times n}$ is a complex matrix, we define the *conjugate* of A to be the matrix $\overline{A} = [\overline{A_{ij}^i}]$ whose entries are the complex conjugates of the entries of A . We say A in $\mathbf{C}^{n \times n}$ is *unitary* if $(\overline{A})^T = A^{-1}$, i.e., if the conjugate transpose of A is the inverse of A .

- Prove that $U(n) = \{A \text{ in } \mathbf{C}^{n \times n} : A \text{ is unitary}\}$ is a group under matrix multiplication.

Hint: Note that $(\overline{A})^T = \overline{(A^T)}$, and use Exercise 11.8.

*Physicists know these as the *Pauli spin matrices*.

- (b) Prove that if $A \in U(2)$, then $|\det A| = 1$, cf. Exercise 11.17 (c).
(In other words, $\det A = e^{i\theta}$ for some real θ .)

Exercise 11.36. In the notation of Exercise 11.34, let $u = a + ci$ and $v = b + di$ be complex numbers, so that

$$A = a\mathbf{1} + b\mathbf{i} + (c\mathbf{1} + d\mathbf{i})\mathbf{j} = \begin{bmatrix} a + ci & -b + di \\ b + di & a - ci \end{bmatrix} = \begin{bmatrix} u & -\bar{v} \\ v & \bar{u} \end{bmatrix}.$$

- (a) Show that $SU(2) = \{A \text{ in } \mathbf{H} : \det A = 1\}$ is a group under matrix multiplication. (This group is the *special unitary group*.)
- (b) Prove $SU(2) = U(2) \cap \mathbf{H}$. ($U(2)$ is defined in Exercise 11.35.)
- (c) Show that $T = \{A \text{ in } SU(2) : A \text{ is diagonal}\}$ is a group under matrix multiplication. Give two proofs: One based on matrix properties alone (i.e., not considering individual elements of A), and one by explicitly writing out the general element of T .
- (d) Let $P = \begin{bmatrix} 1 & i \\ -i & -1 \end{bmatrix}$. Calculate the general element of PTP^{-1} .

Exercise 11.37. In the notation of Exercise 11.34, show that the eight-element set $\{\pm\mathbf{1}, \pm\mathbf{i}, \pm\mathbf{j}, \pm\mathbf{k}\} \subseteq \mathbf{H}$ is closed under matrix multiplication, and write out the Cayley table.

Suggestion: Use Exercise 11.34 (c) rather than multiplying out matrices. For example, $\mathbf{i}^{-1} = -\mathbf{i}$ (etc.), and $(AB)^{-1} = B^{-1}A^{-1}$ generally under an associative operation (see Exercise 7.16), so

$$(\mathbf{j} \cdot \mathbf{i})^{-1} = \mathbf{i}^{-1} \cdot \mathbf{j}^{-1} = -\mathbf{i} \cdot -\mathbf{j} = \mathbf{i} \cdot \mathbf{j} = \mathbf{k};$$

thus $\mathbf{j} \cdot \mathbf{i} = \mathbf{k}^{-1} = -\mathbf{k}$.

Chapter 12

Isomorphisms

Examples 6.8 and 6.9 present three pairs of “abstractly equivalent” binary operations, such as

$$\begin{array}{c|ccc} \cdot & a & b & c \\ \hline a & a & c & b \\ b & b & a & c \\ c & c & b & a \end{array} \quad \text{and} \quad \begin{array}{c|ccc} * & 0 & 1 & 2 \\ \hline 0 & 0 & 2 & 1 \\ 1 & 1 & 0 & 2 \\ 2 & 2 & 1 & 0 \end{array}$$

The binary operation \cdot acts on the set $A = \{a, b, c\}$, while $*$ acts on $B = \{0, 1, 2\}$, so these operations are not literally identical. However, when we “rename” elements in an obvious way, the *entries of the Cayley table correspond under the same renaming*. We regard the pairs (A, \cdot) and $(B, *)$ as “the same” insofar as properties of binary operations are concerned. Indeed, if we distinguish them mathematically, we must admit that American mathematicians, Japanese mathematicians, and Greek mathematicians are engaged in different endeavors, since they use different names to express their concepts!

Formally, a “renaming” of elements of a set A with “labels” from a set B is a bijection $\phi : A \rightarrow B$. The condition of the Cayley table “having corresponding entries” is expressed by the concept of “isomorphism”, see also Figure 6.1.

Definition 12.1. Let (G_1, \cdot) and $(G_2, *)$ be groups. We say (G_1, \cdot) is *isomorphic* to $(G_2, *)$ if there exists a bijection $\phi : G_1 \rightarrow G_2$ satisfying the *morphism condition*

$$\phi(a \cdot b) = \phi(a) * \phi(b) \quad \text{for all } a \text{ and } b \text{ in } G_1.$$

An isomorphism $\phi : (G, \cdot) \rightarrow (G, \cdot)$ is an *automorphism* of (G, \cdot) .

A few general properties of isomorphisms will be useful to record before looking at examples.

Theorem 12.2. *Let $\phi : G_1 \rightarrow G_2$ be an isomorphism of groups, and let e_i in G_i be the respective identity elements.*

- (i) $\phi(e_1) = e_2$.
- (ii) For all a in G_1 , $\phi(a^{-1}) = \phi(a)^{-1}$.
- (iii) $\phi^{-1} : G_2 \rightarrow G_1$ is an isomorphism.
- (iv) For each a in G_1 , $a^m = e_1$ if and only if $\phi(a)^m = e_2$.
- (v) $a \cdot b = b \cdot a$ in G_1 if and only if $\phi(a) * \phi(b) = \phi(b) * \phi(a)$ in G_2 .

Remark 12.3. If (G_1, \cdot) and $(G_2, *)$ are isomorphic, then both groups have the same numbers of elements of order k for each positive integer k , and both groups are Abelian or both are non-Abelian.

Proof. (i) Applying ϕ to the identity $e_1 \cdot e_1 = e_1$ gives

$$\phi(e_1) * \phi(e_1) = \phi(e_1 \cdot e_1) = \phi(e_1) = \phi(e_1) * e_2.$$

The cancellation law in $(G_2, *)$ implies $\phi(e_1) = e_2$.

(ii) Let a in G_1 be arbitrary. Applying ϕ to $e_1 = a \cdot a^{-1}$ and using part (i) gives

$$e_2 = \phi(e_1) = \phi(a \cdot a^{-1}) = \phi(a) * \phi(a^{-1}).$$

By uniqueness of inverses, $\phi(a^{-1}) = \phi(a)^{-1}$.

(iii) Let $a_2 = \phi(a_1)$ and $b_2 = \phi(b_1)$ be arbitrary elements of G_2 . By the morphism condition,

$$a_2 * b_2 = \phi(a_1) * \phi(b_1) = \phi(a_1 \cdot b_1).$$

Applying ϕ^{-1} gives

$$\phi^{-1}(a_2 * b_2) = \phi^{-1}(\phi(a_1 \cdot b_1)) = a_1 \cdot b_1 = \phi^{-1}(a_2) \cdot \phi^{-1}(b_2).$$

This proves (iii) since a_2 and b_2 were arbitrary.

(iv) A straightforward induction argument shows $\phi(a^k) = \phi(a)^k$ for all integers k . By parts (i) and (iii) of this theorem, $a^m = e_1$ if and only if $\phi(a)^m = \phi(a^m) = \phi(e_1) = e_2$.

(v) If $a \cdot b = b \cdot a$, applying ϕ gives $\phi(a) * \phi(b) = \phi(b) * \phi(a)$ by the morphism condition. The converse follows immediately from (iii). \square

Example 12.4. You have studied isomorphisms since high school, probably without knowing it! Consider the additive group of real numbers $(\mathbf{R}, +)$ and the multiplicative group of positive reals (\mathbf{R}^+, \cdot) . The mapping $\phi : \mathbf{R} \rightarrow \mathbf{R}^+$ defined by $\phi(x) = e^x$ is a bijection, and satisfies the morphism condition by the law of exponents: If x and y are arbitrary real numbers, then

$$\phi(x + y) = e^{x+y} = e^x \cdot e^y = \phi(x) \cdot \phi(y).$$

The inverse mapping, a.k.a. the natural logarithm function \log , is also an isomorphism. If $a = e^x$ and $b = e^y$, the morphism condition becomes the logarithm identity

$$\log(a \cdot b) = \phi^{-1}(a \cdot b) = \phi^{-1}(a) + \phi^{-1}(b) = \log a + \log b.$$

Example 12.5. The additive group $(\mathbf{Z}_2, +)$ of integers mod 2 and the multiplicative group $(\{1, -1\}, \cdot)$ of signs are isomorphic under the mapping $\phi([0]) = 1$, $\phi([1]) = -1$. The respective Cayley tables are

$$\begin{array}{c|cc} + & [0] & [1] \\ \hline [0] & [0] & [1] \\ [1] & [1] & [0] \end{array} \quad \text{and} \quad \begin{array}{c|cc} \cdot & 1 & -1 \\ \hline 1 & 1 & -1 \\ -1 & -1 & 1 \end{array}$$

The isomorphism in this example is given by a formula, $\phi(x) = (-1)^x$. The morphism condition is again the law of exponents:

$$\phi(a + b) = (-1)^{a+b} = (-1)^a \cdot (-1)^b = \phi(a) \cdot \phi(b)$$

if a and b are arbitrary elements of \mathbf{Z}_2 .

Example 12.6. Fix an integer $n \geq 2$, and put $\zeta = e^{2\pi i/n}$, an n th root of unity. Let $G = (\mathbf{Z}_n, +)$ be the additive group of residue classes mod n , and let $G' = (\langle \zeta \rangle, \cdot)$ be the cyclic group of order n generated by ζ in the multiplicative group of non-zero complex numbers.

The mapping $\phi([k]) = e^{2\pi i k/n}$ is an isomorphism from G to G' . First, ϕ is well-defined and bijective since for arbitrary integers k_1 and k_2 , we have $e^{2\pi i k_1/n} = e^{2\pi i k_2/n}$ if and only if $e^{2\pi i (k_1 - k_2)/n} = 1$, if and only if $n \mid (k_1 - k_2)$, if and only if $[k_1] = [k_2]$ in G .

Once again, the morphism condition follows immediately from the law of exponents:

$$\phi([k_1] + [k_2]) = e^{2\pi i (k_1 + k_2)/n} = e^{2\pi i k_1/n} \cdot e^{2\pi i k_2/n} = \phi([k_1]) \cdot \phi([k_2]).$$

When $n = 2$, this reduces to the preceding example.

Example 12.7. Consider the three groups of order four shown in Table 12.1: The additive group $(\mathbf{Z}_4, +)$ of integers mod 4; the multiplicative group $(\mathbf{Z}_5^\times, \cdot)$ of units mod 5, and the group $(\mathbf{Z}_8^\times, \cdot)$ of units mod 8.

$+_4$	$[0]_4$	$[1]_4$	$[2]_4$	$[3]_4$	\cdot_5	$[1]_5$	$[2]_5$	$[3]_5$	$[4]_5$
$[0]_4$	$[0]_4$	$[1]_4$	$[2]_4$	$[3]_4$	$[1]_5$	$[1]_5$	$[2]_5$	$[3]_5$	$[4]_5$
$[1]_4$	$[1]_4$	$[2]_4$	$[3]_4$	$[0]_4$	$[2]_5$	$[2]_5$	$[4]_5$	$[1]_5$	$[3]_5$
$[2]_4$	$[2]_4$	$[3]_4$	$[0]_4$	$[1]_4$	$[3]_5$	$[3]_5$	$[1]_5$	$[4]_5$	$[2]_5$
$[3]_4$	$[3]_4$	$[0]_4$	$[1]_4$	$[2]_4$	$[4]_5$	$[4]_5$	$[3]_5$	$[2]_5$	$[1]_5$

\cdot_8	$[1]_8$	$[3]_8$	$[5]_8$	$[7]_8$
$[1]_8$	$[1]_8$	$[3]_8$	$[5]_8$	$[7]_8$
$[3]_8$	$[3]_8$	$[1]_8$	$[7]_8$	$[5]_8$
$[5]_8$	$[5]_8$	$[7]_8$	$[1]_8$	$[3]_8$
$[7]_8$	$[7]_8$	$[5]_8$	$[3]_8$	$[1]_8$

Table 12.1: Cayley tables for $(\mathbf{Z}_4, +)$, $(\mathbf{Z}_5^\times, \cdot)$, and $(\mathbf{Z}_8^\times, \cdot)$.

A glance at the three Cayley tables may suggest no two of these groups are isomorphic, since the table entries do not correspond in an obvious way. However, this assessment may be incorrect; the structure of a Cayley table can be obscured by the order in which the elements are written in the table. Instead, for each pair of groups we must either find an isomorphism or prove no isomorphism exists.

As a first approach, we'll compute the order of each element in each group and attempt to use Theorem 12.2 (iv).

In $(\mathbf{Z}_4, +)$, $[x] + [x] + [x] + [x] = [4x] = [0]$ for all $[x]$; clearly, $[1]_4$ and $[3]_4 = [-1]_4$ have order 4, while $[2]_4$ has order 2.

In $(\mathbf{Z}_5^\times, \cdot)$, reading products out of the Cayley table enables us to compute powers easily: $[2]_5^2 = [4]_5$, $[2]_5^3 = [2]_5[4]_5 = [3]_5$, and $[2]_5^4 = [1]_5$; thus $[2]_5$ has order 4, $[3]_5 = [2]_5^{-1}$ has order 4, and $[4]_5$ has order 2.

By contrast, the group $(\mathbf{Z}_8^\times, \cdot)$ contains three elements of order 2; that is, $x^2 = [1]_8$ for all x in \mathbf{Z}_8^\times . Theorem 12.2 (iv) implies $(\mathbf{Z}_8^\times, \cdot)$,

which contains no element of order 4, is not isomorphic to either $(\mathbf{Z}_4, +)$ or $(\mathbf{Z}_5^\times, \cdot)$.

Are $(\mathbf{Z}_4, +)$ and $(\mathbf{Z}_5^\times, \cdot)$ isomorphic? If there exists an isomorphism $\phi : (\mathbf{Z}_4, +) \rightarrow (\mathbf{Z}_5^\times, \cdot)$, then Theorem 12.2 implies $\phi([0]_4) = [1]_5$ (the identity maps to the identity) and $\phi([2]_4) = [4]_5$ (an element of order 2 maps to an element of order 2). Since ϕ must be a bijection, we can only have $\phi([1]_4) = [2]_5$ or $\phi([1]_4) = [3]_5$. We'll examine these possibilities one at a time.

If $\phi([1]_4) = [2]_5$, the morphism condition completely determines the mapping:

$$\begin{aligned}\phi([2]_4) &= \phi([1]_4 + [1]_4) = \phi([1]_4) \cdot \phi([1]_4) = [2]_5 \cdot [2]_5 = [4]_5, \\ \phi([3]_4) &= \phi([1]_4 + [2]_4) = \phi([1]_4) \cdot \phi([2]_4) = [2]_5 \cdot [4]_5 = [3]_5, \\ \phi([0]_4) &= \phi([1]_4 + [3]_4) = \phi([1]_4) \cdot \phi([3]_4) = [2]_5 \cdot [3]_5 = [1]_5.\end{aligned}$$

This mapping is induced by the formula $\phi(x) = ([2]_5)^x$ for x in \mathbf{Z} . Indeed, if $x \equiv y \pmod{4}$, then $[2]_5^x = [2]_5^y$ because $[2]_5^4 = [1]_5$. The morphism condition follows from the law of exponents:

$$\phi([x]_4 + [y]_4) = \phi([x + y]_4) = [2]_5^{x+y} = [2]_5^x \cdot [2]_5^y = \phi([x]_4) \cdot \phi([y]_4).$$

As an exercise, write out the Cayley table for $(\mathbf{Z}_5^\times, \cdot)$ with the elements in the order $\{[1], [2], [4], [3]\}$ and verify that this rearranged table corresponds entry-for-entry with the table for $(\mathbf{Z}_4, +)$.

If $\psi([1]_4) = [3]_5$ instead, similar considerations give

$$\psi([2]_4) = [4]_5, \quad \psi([3]_4) = [2]_5, \quad \psi([0]_4) = [1]_5.$$

This bijection is induced by $\psi(x) = ([3]_5)^x$, and as before is an isomorphism from $(\mathbf{Z}_4, +)$ to $(\mathbf{Z}_5^\times, \cdot)$. These isomorphisms are closely related. Since $[3]_5 = [2]_5^{-1}$, we have $\psi(x) = ([3]_5)^x = ([2]_5)^{-x} = \phi(-x)$.

Example 12.8. Consider the group $G = (\mathbf{Z}_8, +)$. For each integer k , there is a mapping $\phi_k : \mathbf{Z}_8 \rightarrow \mathbf{Z}_8$ defined by the formula $\phi_k([a]_8) = [ka]_8$. Without loss of generality we may reduce k mod 8. These mappings satisfy the morphism condition by the distributive law:

$$\phi_k([a + b]_8) = k[a + b]_8 = k[a]_8 + k[b]_8 = \phi_k([a]_8) + \phi_k([b]_8).$$

Consequently, ϕ_k is an automorphism of G if and only if ϕ_k is bijective, if and only if ϕ_k is invertible as a mapping, if and only if $[k]_8$ is multiplicatively invertible in \mathbf{Z}_8 , if and only if $\gcd(k, 8) = 1$.

The mappings ϕ_k are the *only* mappings from \mathbf{Z}_8 to itself satisfying the morphism condition. The group G is cyclic and $1 = [1]_8$ is a generator, so just as in Example 12.7, the value $k = \phi(1)$ completely determines the mapping ϕ , and $\phi(a) = ka$.

To summarize, there exist exactly four automorphisms of $(\mathbf{Z}_8, +)$,

	[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	Formula
ϕ_1	[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	$\phi_1(x) = x$,
ϕ_3	[0]	[3]	[6]	[1]	[4]	[7]	[2]	[5]	$\phi_3(x) = 3x$,
ϕ_5	[0]	[5]	[2]	[7]	[4]	[1]	[6]	[3]	$\phi_5(x) = 5x$,
ϕ_7	[0]	[7]	[6]	[5]	[4]	[3]	[2]	[1]	$\phi_7(x) = 7x$.

12.1 Classification of Cyclic Groups

A cyclic group is classified up to isomorphism by its order. Before formally stating and proving a theorem, we establish a technical property of powers of elements in a group.

Proposition 12.9. *Let (G, \cdot) be a group with identity element e , and assume a in G has order n . If k_1 and k_2 are integers, then $a^{k_1} = a^{k_2}$ if and only if $k_1 \equiv k_2 \pmod{n}$.*

Proof. Since $a^{k_1} = a^{k_2}$ if and only if $a^{k_2-k_1} = e$, it suffices to prove $a^k = e$ if and only if $n \mid k$.

(If $n \mid k$ then $a^k = e$). If $n \mid k$, then $k = n\ell$ for some integer ℓ . By Example 7.16, $a^k = (a^n)^\ell = e^\ell = e$.

(If $a^k = e$, then $n \mid k$). By the division algorithm, there exist integers q and r , $0 \leq r < n$, such that $k = nq + r$. By the law of exponents,

$$e = a^k = a^{nq+r} = (a^n)^q \cdot a^r = e^q \cdot a^r = a^r.$$

Since n is the smallest positive exponent such that $a^n = e$, we have $r = 0$. Thus $k = nq$, i.e., $n \mid k$. \square

Theorem 12.10. *Let (G, \cdot) be a cyclic group, and assume $a \in G$ is a generator.*

If G is infinite, then the formula $\phi(k) = a^k$ defines an isomorphism $\phi : (\mathbf{Z}, +) \rightarrow (G, \cdot)$.

If G contains n elements, the formula $\bar{\phi}([k]) = a^k$ defines an isomorphism $\bar{\phi} : (\mathbf{Z}_n, +) \rightarrow (G, \cdot)$.

Proof. Let a be a generator of (G, \cdot) , and define a mapping $\phi : \mathbf{Z} \rightarrow G$ by $\phi(k) = a^k$. By hypothesis, ϕ is surjective. The law of exponents guarantees that ϕ satisfies the morphism condition:

$$\phi(k + \ell) = a^{k+\ell} = a^k \cdot a^\ell = \phi(k) \cdot \phi(\ell).$$

Finally, $\phi(k_1) = \phi(k_2)$ if and only if $a^{k_1} = a^{k_2}$, if and only if $a^{k_2-k_1} = e$.

Suppose G has infinite order. By Proposition 12.9, a has infinite order; that is, $a^k \neq e$ for all non-zero k . By the observation above, if $\phi(k_1) = \phi(k_2)$ for some integers k_1, k_2 , then $k = k_2 - k_1 = 0$, i.e., $k_1 = k_2$. This means ϕ is injective, hence bijective, hence an isomorphism.

Suppose instead that a has order n in (G, \cdot) . Let R be the equivalence relation “congruence mod n ” on \mathbf{Z} , whose set of equivalence classes is \mathbf{Z}_n .

By Proposition 12.9, $a^{k_1} = a^{k_2}$ if and only if $k_2 - k_1 \in n\mathbf{Z}$, if and only if $k_1 \equiv k_2 \pmod{n}$. This means the mapping ϕ is constant on equivalence classes of R , so there is a well-defined (i.e., single-valued) induced mapping $\bar{\phi} : \mathbf{Z}_n \rightarrow G$ defined by $\bar{\phi}([k]) = a^k$. This mapping is injective by the preceding observations, and satisfies the morphism condition by the law of exponents:

$$\bar{\phi}([k] + [\ell]) = \bar{\phi}([k + \ell]) = a^{k+\ell} = a^k \cdot a^\ell = \bar{\phi}([k]) \cdot \bar{\phi}([\ell]).$$

That is, $\bar{\phi} : (\mathbf{Z}_n, +) \rightarrow (G, \cdot)$ is an isomorphism. \square

Corollary 12.11. *Let (G, \cdot) be a group. If a has order n in G , then*

$$\langle a \rangle = \{a^k : 0 \leq k < n\} = \{e, a, a^2, \dots, a^{n-1}\}.$$

Proof. Since the element a has order n in (G, \cdot) , the cyclic group $\langle a \rangle$ consists of powers of a corresponding to distinct residue classes mod n by Proposition 12.9. \square

Corollary 12.12. *Let (G, \cdot) be a group, a in G arbitrary. The elements a and a^{-1} have the same order.*

Proof. Clearly $\langle a \rangle = \langle a^{-1} \rangle$, so these sets are either both infinite, or are both finite and contain the same number of elements. \square

Remark 12.13. In an arbitrary group, taking powers of an element a yields nothing worse than a cyclic subgroup, whose abstract structure is completely determined by the order of a .

Example 12.14. In $(\mathbf{R}^\times, \cdot)$, the multiplicative group of non-zero real numbers, the identity element 1 has order 1, the element -1 has order 2, and every other element has infinite order. (Why?)

Example 12.15. Let $G = SL(2, \mathbf{R})$ be the set of 2×2 real matrices of determinant 1, $n > 0$ an integer and consider the elements

$$a = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}, \quad \theta = 2\pi/n.$$

Geometrically, a shears the plane horizontally, while b is a rotation by $1/n$ of a turn. A brief calculation for a , or an invocation of Exercise 11.18 for b , shows

$$a^r = \begin{bmatrix} 1 & r \\ 0 & 1 \end{bmatrix}, \quad b^r = \begin{bmatrix} \cos r\theta & -\sin r\theta \\ \sin r\theta & \cos r\theta \end{bmatrix}$$

for all r in \mathbf{Z} . Thus a has infinite order, while b has order n .

Example 12.16. Again, let $G = SL(2, \mathbf{R})$, and consider

$$a = \begin{bmatrix} \cos 1 & -\sin 1 \\ \sin 1 & \cos 1 \end{bmatrix}, \quad \text{for which} \quad a^r = \begin{bmatrix} \cos r & -\sin r \\ \sin r & \cos r \end{bmatrix}.$$

We have $a^r = I$ if and only if $r = 2\pi s$ for some integer s . If $r \neq 0$, then $s \neq 0$, and this condition reads $2\pi = r/s$. Since 2π is irrational, there do not exist integers r and s such that $2\pi = r/s$. This means $a^r = I$ if and only if $r = 0$, which means a has infinite order. Generally, rotation by θ has infinite order if and only if θ is an irrational multiple of 2π .

Theorem 12.17. Let $(\langle a \rangle, \cdot)$ be a cyclic group of order n .

- (i) An element a^k is a generator of $\langle a \rangle$ if and only if $\gcd(k, n) = 1$.
- (ii) A mapping $\phi : \langle a \rangle \rightarrow \langle a \rangle$ is an automorphism if and only if there exists an integer k , coprime to n , such that $\phi(a^r) = (a^r)^k$ for all $r = 0, 1, \dots, n-1$, i.e., $\phi(x) = x^k$ for all x in $\langle a \rangle$.

Proof. (i) By Theorem 12.10, the correspondence associating $[k]$ in $(\mathbf{Z}_n, +)$ to a^k in $(\langle a \rangle, \cdot)$ is an isomorphism of groups. An element a^k is therefore a generator of $(\langle a \rangle, \cdot)$ if and only if $[k]$ is a generator of $(\mathbf{Z}_n, +)$. By Theorem 10.10, this is equivalent to $\gcd(k, n) = 1$.

(ii) A mapping $\phi : \langle a \rangle \rightarrow \langle a \rangle$ that satisfies the morphism condition is uniquely determined by $\phi(a) = a^k$, since $\phi(a^r) = \phi(a)^r = (a^k)^r = a^{kr}$. For k fixed, an arbitrary element of $\langle a \rangle$ can be written in the form a^{kr} if and only if a^k is a generator of $\langle a \rangle$, and by part (i) this is equivalent to saying k and n are coprime. \square

Exercises

Exercise 12.1. Find all automorphisms of $(\mathbf{Z}_5, +)$, and tabulate their values as in Example 12.8.

Exercise 12.2. Find all automorphisms of $(\mathbf{Z}_6, +)$, and tabulate their values as in Example 12.8.

Exercise 12.3. Find all automorphisms of $(\mathbf{Z}_{12}, +)$ and tabulate their values as in Example 12.8.

Exercise 12.4. Find all automorphisms of $(\mathbf{Z}_{14}^\times, \cdot)$ and tabulate their values as in Example 12.8.

Hint: First prove $(\mathbf{Z}_{14}^\times, \cdot)$ is a cyclic group.

Exercise 12.5. Find all automorphisms of $(\mathbf{Z}_2 \times \mathbf{Z}_3, +)$ and tabulate their values as in Example 12.8.

Hint: Show $(\mathbf{Z}_2 \times \mathbf{Z}_3, +)$ is cyclic.

Exercise 12.6. Show $(\mathbf{Z}_3 \times \mathbf{Z}_4, +)$ is cyclic. For each automorphism ϕ , find a formula of the type $\phi(x, y) = (kx, \ell y)$ with x in \mathbf{Z}_3 and y in \mathbf{Z}_4 .

Exercise 12.7. If a has order 9 in (G, \cdot) , find the order of each element of $\langle a \rangle$, and list the generators.

Exercise 12.8. If a has order 10 in (G, \cdot) , find the order of each element of $\langle a \rangle$, and list the generators.

Exercise 12.9. If a has order n in (G, \cdot) , prove that $(a^k)^{-1} = a^{n-k}$ for $0 < k < n$.

Exercise 12.10. Let $r > 0$ be a fixed non-zero real number, and define $\phi_r : \mathbf{R}^+ \rightarrow \mathbf{R}^+$ by $\phi_r(a) = a^r$. Prove ϕ_r is an automorphism of the multiplicative group (\mathbf{R}^+, \cdot) .

Exercise 12.11. In each part, $\phi : \mathbf{C} \rightarrow \mathbf{C}$ denotes complex conjugation, $\phi(a + bi) = a - bi$.

- (a) Let $(\mathbf{C}, +)$ be the additive group of complex numbers. Prove that ϕ is an automorphism.
- (b) Let $(\mathbf{C}^\times, \cdot)$ be the multiplicative group of non-zero complex numbers. Prove that ϕ is an automorphism.

Exercise 12.12. Let α be a non-zero complex number. Prove:

- (a) If $|\alpha| \neq 1$, then α has infinite order in $(\mathbf{C}^\times, \cdot)$.
- (b) If $\alpha = e^{i\theta}$ and $\theta/(2\pi)$ is not rational, then α has infinite order.
- (c) If $\alpha = e^{i\theta}$ with $\theta/(2\pi) = p/q$ in lowest terms and $q > 0$, then α has order q . (Compare Example 2.39 and Section 7.3.)

Exercise 12.13. Prove that every *finite* subgroup of $(\mathbf{C}^\times, \cdot)$ is cyclic. Hint: Use the preceding exercise.

Exercise 12.14. For each matrix given, list the elements of the cyclic subgroup generated by A in the multiplicative group $GL(2, \mathbf{R})$ of invertible matrices, and sketch the image of the standard \mathbf{F} under A .

$$(a) A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad (b) A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad (c) A = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}.$$

Exercise 12.15. For each matrix, determine the order, and if the order is finite list all powers of the matrix.

$$(a) A = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}, \quad (b) B = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad (c) C = AB.$$

Exercise 12.16. Let $(\langle a \rangle, \cdot)$ be a cyclic group.

- (a) If b is a generator, prove b^{-1} is also a generator.
- (b) If a has finite order greater than 2, prove $(\langle a \rangle, \cdot)$ has an even number of generators.

Hint: Use part (a).

Exercise 12.17. (a) Prove that a composition of isomorphisms is an isomorphism. (Your proof should include a formal statement.)

(b) Define a relation \simeq on the set of all groups by saying $G_1 \simeq G_2$ if and only if G_1 is isomorphic to G_2 . Prove \simeq is an equivalence relation.

Exercise 12.18. Let (G, \cdot) be a group, and fix a in G . Define the mapping $\phi_a : G \rightarrow G$ by $\phi_a(x) = a^{-1}xa$.

- (a) Prove ϕ_a is an automorphism of (G, \cdot) .
- (b) Prove $(\phi_a)^{-1} = \phi_{a^{-1}}$.
- (c) If (G, \cdot) is Abelian, prove ϕ_a is the identity mapping, regardless of a .

Exercise 12.19. Define $\phi : (\mathbf{R}, +) \rightarrow (SL(2, \mathbf{R}), \cdot)$ by

$$\phi(t) = \begin{bmatrix} e^t & 0 \\ 0 & e^{-t} \end{bmatrix}.$$

- (a) Prove the image of ϕ is a subgroup of $(SL(2, \mathbf{R}), \cdot)$
- (b) Prove ϕ is an isomorphism onto its image.

Exercise 12.20. Define $\phi : (\mathbf{R}, +) \rightarrow (SL(2, \mathbf{R}), \cdot)$ by

$$\phi(t) = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix}.$$

- (a) Does ϕ satisfy the morphism condition?
- (b) Is ϕ an isomorphism onto its image?

Exercise 12.21. For real t , define the *hyperbolic trig functions* by

$$\cosh t = \frac{1}{2}(e^t + e^{-t}), \quad \sinh t = \frac{1}{2}(e^t - e^{-t}).$$

Establish the following identities, with s and t arbitrary real numbers.

- (a) $\cosh t \pm \sinh t = e^{\pm t}$.
- (b) $\cosh^2 t - \sinh^2 t = 1$. Hint: Use part (a).
- (c) $\cosh(s + t) = \cosh s \cosh t + \sinh s \sinh t$.
- (d) $\sinh(s + t) = \cosh s \sinh t + \sinh s \cosh t$.
- (e) Prove $\sinh : \mathbf{R} \rightarrow \mathbf{R}$ is a bijection. Hint: Write $y = \sinh x$, and multiply through by e^x . Use the quadratic formula to solve for e^x .

Exercise 12.22. Define $\phi : (\mathbf{R}, +) \rightarrow (SL(2, \mathbf{R}), \cdot)$ by

$$\phi(t) = \begin{bmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{bmatrix}.$$

(See Exercise 12.21 for the definitions and properties of \cosh and \sinh .)

- (a) Does ϕ satisfy the morphism condition?
- (b) Is ϕ an isomorphism onto its image?

Exercise 12.23. Prove that $(\mathbf{Z}_{30}^\times, \cdot)$, the group of units mod 30, is isomorphic to $(\mathbf{Z}_4 \times \mathbf{Z}_2, +)$.

Hint: The structure of $(\mathbf{Z}_{30}^\times, \cdot)$ is partially worked out in Example 10.15.

Exercise 12.24. Let (G, \cdot) be a group. Let $\text{Aut}(G)$ be the set of automorphisms of (G, \cdot) . Prove $\text{Aut}(G)$ is a group under mapping composition.

Exercise 12.25. Prove $\text{Aut}(\mathbf{Z}_8)$, the group of automorphisms of $(\mathbf{Z}_8, +)$, is isomorphic to $(\mathbf{Z}_8^\times, \cdot)$, the multiplicative group of units mod 8.

Exercise 12.26. Let $n \geq 2$. Prove $\text{Aut}(\mathbf{Z}_n)$, the group of automorphisms of $(\mathbf{Z}_n, +)$, is isomorphic to $(\mathbf{Z}_n^\times, \cdot)$, the group of units mod n .

Exercise 12.27. Let (G, \cdot) be a group. Define a new binary operation $*$ by $a * b = b \cdot a$.

(a) Prove $(G, *)$ is a group.*

(b) Define $\phi : G \rightarrow G$ by $\phi(a) = a^{-1}$. Prove ϕ is an isomorphism from (G, \cdot) to $(G, *)$.

(c) Prove the inversion map ϕ in part (b) is an automorphism of (G, \cdot) if and only if (G, \cdot) is Abelian.

*It is called the *opposite group* of (G, \cdot) .

Chapter 13

The Symmetric Group

Let X be a finite, non-empty set. On $\mathcal{M}(X)$, the set of all mappings $f : X \rightarrow X$, composition is an associative binary operation.

Definition 13.1. A *permutation* of X is a bijection $f : X \rightarrow X$, i.e., an invertible element of $(\mathcal{M}(X), \circ)$.

The set S_X of all permutations of X forms a group under mapping composition, called the *symmetric group* of X .

Remark 13.2. If X is the set $\{1, 2, 3, \dots, n\}$, we call S_X the *symmetric group on n letters*, and denote it S_n .

Remark 13.3. Because X is finite, a mapping $f : X \rightarrow X$ is bijective if and only if it is injective, if and only if it is surjective, see Exercise 4.7.

Proposition 13.4. Let $X = \{1, 2, 3, \dots, n\}$. The set $\mathcal{M}(X)$ of mappings has n^n elements. The symmetric group S_n has order $n!$.

Proof. A mapping $f : X \rightarrow X$ is uniquely specified by its n values $f(1), \dots, f(n)$. There are n choices for each value and these choices are independent, so there are n^n distinct mappings.

If $f : X \rightarrow X$ is a bijection, by contrast, there are n choices for $f(1)$, $(n - 1)$ choices for $f(2) \neq f(1)$, $(n - 2)$ choices for $f(3)$, etc. In total, there are $n! = n(n - 1)(n - 2) \dots 3 \cdot 2 \cdot 1$ bijections of X . \square

Example 13.5. At a dinner party for six, a seating arrangement is a bijection from the sets of guests to the set of chairs. Consequently, there are $6! = 720$ distinct seatings, one for each permutation on six letters. Similarly, there are

$$\begin{aligned} 52! &= 80,658,175,170,943,878,571,660,636,856,403,766, \\ &\quad 975,289,505,440,883,277,824,000,000,000,000, \end{aligned}$$

or about 8.0658×10^{67} , distinct shufflings for a deck of playing cards.

13.1 Structure of the Symmetric Group

Our goals in this section and the next are to understand both the structure of the symmetric group S_n as well as the action of individual permutations. A “typical” permutation, Example 13.6, will convey the main qualitative phenomena.

Example 13.6. Consider the permutation on nine letters defined by

a	1	2	3	4	5	6	7	8	9
$f(a)$	4	8	2	9	7	6	1	3	5

The element 1 maps to 4 under f . In turn, 4 maps to 9, 9 maps to 5, 5 to 7, and 7 maps back to 1. The five elements $\{1, 4, 9, 5, 7\}$ are therefore *cyclically* permuted by f .

Because f is a bijection, the complementary set $\{2, 3, 6, 8\}$ is also permuted by f . Inspection shows the elements $\{2, 8, 3\}$ are cyclically permuted, while 6 is mapped to itself.

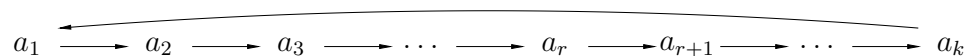
The domain of f , $\{1, 2, \dots, 9\} = \{1, 4, 5, 7, 9\} \cup \{2, 3, 8\} \cup \{6\}$, is partitioned into three subsets, each cyclically permuted by f .

Definition 13.7. Let a_1, a_2, \dots, a_k be distinct elements of X . The k -cycle $f = (a_1 \ a_2 \ \dots \ a_k)$ is the permutation defined by

$$\begin{aligned} f(a_r) &= a_{r+1} && \text{if } 1 \leq r < k, \\ f(a_k) &= a_1 \\ f(a) &= a && \text{if } a \neq a_r \text{ for } r = 1, 2, \dots, k. \end{aligned}$$

A 2-cycle $(a_1 \ a_2)$ is also called a *transposition*.

Remark 13.8. In words, $(a_1 \ a_2 \ \dots \ a_k)$ cycles the k letters a_1, \dots, a_k and fixes all other elements of X . To emphasize the cyclic nature of the permutation, the first two conditions in the definition can be combined into $f(a_r) = a_{(r+1) \bmod k}$, with residue classes $r = 1, \dots, k$. Pictorially, the k -cycle $(a_1 \ a_2 \ \dots \ a_k)$ should be visualized as



Remark 13.9. A k -cycle can be written exactly k different ways in cycle notation. Any of the k indices a_r may be listed first, but the remaining indices are completely determined. For example,

$$(1\ 4\ 3) = (4\ 3\ 1) = (3\ 1\ 4) \quad \text{and} \quad (2\ 5) = (5\ 2).$$

For definiteness, the smallest index is conventionally written first.

Example 13.10. The symmetric group S_2 contains two elements: The identity permutation $(1)(2)$ and the transposition $(1\ 2)$.

S_3 contains $3! = 6$ elements. To list them, enumerate the orderings of 1, 2, and 3 by viewing each ordering as an integer in standard notation, 123, 132, 213, 231, 312, and 321:

$$\begin{array}{lll} 1\ 2\ 3 \longleftrightarrow (1)(2)(3) & 2\ 1\ 3 \longleftrightarrow (1\ 2)(3) & 3\ 1\ 2 \longleftrightarrow (1\ 3\ 2), \\ 1\ 3\ 2 \longleftrightarrow (1)(2\ 3) & 2\ 3\ 1 \longleftrightarrow (1\ 2\ 3) & 3\ 2\ 1 \longleftrightarrow (1\ 3)(2). \end{array}$$

The “digits” on the left-hand side of each correspondence tabulate the respective values $f(1)$, $f(2)$, $f(3)$. The right-hand side gives the cycle decomposition.

Example 13.11. The permutation f of Example 13.6 is made up of the 5-cycle $(1\ 4\ 9\ 5\ 7)$, the 3-cycle $(2\ 8\ 3)$, and the 1-cycle (6) . With concatenation standing for mapping composition,

$$f = (1\ 4\ 9\ 5\ 7)(2\ 8\ 3)(6).$$

An arbitrary permutation similarly decomposes into “disjoint cycles”, see Theorem 13.21.

As we will see, useful properties can be read off this representation. For example, $f = (1\ 4\ 9\ 5\ 7)(2\ 8\ 3)(6)$ in S_9 generates a cyclic subgroup of order 15 (the least common multiple of the cycle lengths). Further, f can be written as the product of six transpositions (and no fewer) but not as the product of any odd number of transpositions.

In establishing properties of general permutations, several pieces of terminology will be useful.

Definition 13.12. Let X be a finite, non-empty set, and f in S_X .

A subset $A \subseteq X$ is *invariant under f* if $f(A) = A$, namely if $a \in A$ implies $f(a) \in A$.

A non-empty set A is *cyclical under f* if the elements of A are cyclically permuted by f .

If $f(a) = a$, we say a is a *fixed point* of f . The complement of the fixed set of f , namely $\{x \text{ in } X : f(x) \neq x\}$, is the *support* of f .

Example 13.13. The permutation $f = (1\ 4\ 9\ 5\ 7)(2\ 8\ 3)(6)$ in S_9 has three cyclical sets: $\{1, 4, 5, 7, 9\}$, $\{2, 3, 8\}$, and $\{6\}$, namely the supports of the disjoint cycles comprising f . Unions of these sets are also invariant.

The only fixed point of f is 6. The support of f is the eight-element set $X \setminus \{6\} = \{1, 2, 3, 4, 5, 7, 8, 9\}$.

Example 13.14. Let f in S_X be arbitrary. The following claims are easily verified, and left as exercises for practice.

Every cyclical set is invariant under f . A union or intersection of invariant sets is invariant, see Proposition 4.14 (i), (ii). The complement of an invariant set is invariant, see Exercise 4.1 (b).

If $f(a) = a$, then $\{a\}$ is cyclical, hence invariant. By the preceding paragraph, every set consisting entirely of fixed points of f is invariant, and the support of f (i.e., the complement of the fixed point set) is invariant under f .

Definition 13.15. Two permutations f and g are *disjointly supported* if their supports are disjoint sets, that is, if every element of X is fixed by either f or g (or both).

Theorem 13.16. If f, g in S_X are disjointly supported, then $gf = fg$.

Remark 13.17. Intuitively, disjointly-supported permutations act on “different universes” (their respective supports), so it does not matter which is performed first; the overall result is the same either way.

Example 13.18. Make tables of values to show that the transpositions $(1\ 2)$ and $(3\ 4)$ commute. By contrast, $(1\ 3)(1\ 2) = (1\ 2\ 3)$ while $(1\ 2)(1\ 3) = (1\ 3\ 2)$. (Remember, the right-hand factor acts *first*.)

Remark 13.19. The symmetric groups S_n with $n \geq 3$ are non-Abelian.

Remark 13.20. The converse of Theorem 13.16 is false; commuting permutations need not have disjoint support.

Proof of theorem. It suffices to show $fg(a) = gf(a)$ for all a in X .

Case 1. $f(a) = a$ and $g(a) = a$. Obviously $fg(a) = a = gf(a)$.

Case 2a. $f(a) = a$ but $g(a) \neq a$. In this event, $g(a)$, being in the support of g , is *not* in the support of f , namely is fixed by f . Consequently,

$$fg(a) = f(g(a)) = g(a) = g(f(a)) = gf(a).$$

Case 2b. $g(a) = a$ but $f(a) \neq a$. The same argument with the roles of f and g reversed shows $fg(a) = gf(a)$. \square

Theorem 13.21. *Let X be a finite, non-empty set. If $f \in S_X$, then the cyclical sets of f partition X ; i.e., f is a product of disjoint cycles.*

Proof. The proof proceeds by induction on n , using the statement

$P(n)$ If X is a set of m elements, $1 \leq m \leq n$, and if $f \in S_X$,
then the cyclical sets of f partition X .

(Base case). The only permutation on one letter is the identity, and the set X is cyclical.

(Inductive step). Assume inductively that $P(N)$ is true for some $N \geq 1$. Let X be a set of $(N+1)$ elements, f an arbitrary permutation in S_X .

Pick an element a_1 in X arbitrarily, and consider the successive values $a_2 = f(a_1)$, $a_3 = f(a_2)$, \dots , $a_{r+1} = f(a_r)$, \dots . After finitely many steps, this sequence must return to the value a_1 . Let k be the smallest positive index such that $a_{k+1} = a_1$.

The set $X_1 = \{a_1, a_2, \dots, a_k\}$ is cyclical under f , so the complement X_1^c of X_1 is invariant under f . If X_1^c is empty, then f acts cyclically on X . Otherwise, X_1^c has m elements, with $1 \leq m \leq N$. By the inductive hypothesis, X_1^c partitions into cyclical sets of f .

In either case, $X = X_1 \cup X_1^c$ is partitioned into cyclical sets of f , so we have proven $P(N+1)$. By the principle of mathematical induction, $P(n)$ is true for all $n \geq 1$. \square

13.2 Cycle Multiplication

By Theorem 13.21, every permutation of a finite set X is a product of disjoint cycles. We turn next to computational tools: how to compute powers of a cycle, how to find the order of a product of disjoint cycles, and how to simplify a product of arbitrary cycles to a product of disjoint cycles.

Assume $f \in S_X$. Recall that the *order* of f is the smallest positive integer k such that the k -fold composition f^k is the identity map on X , a.k.a. the order of the cyclic subgroup $\langle f \rangle \subseteq S_X$.

Example 13.22. Let $f = (a_1 \ a_2 \ \dots \ a_k)$ be a k -cycle, which is to say $f(a_r) = a_{(r+1) \bmod k}$. The inverse maps each index to its preimage under f , so $f^{-1} = (a_1 \ a_k \ a_{k-1} \ \dots \ a_2)$, or $f^{-1}(a_r) = a_{(r-1) \bmod k}$.

Example 13.23. The order of a k -cycle is k . Indeed, if s is an integer, then in the notation of the preceding example, $f^s(a_r) = a_{(r+s) \bmod k}$. This is the identity map if and only if $k \mid s$.

Powers of a cycle can be written as products of disjoint cycles. If $f = (1\ 2\ 3\ 4\ 5\ 6)$, for example, then by the formula of the preceding paragraph, the elements of the cyclic subgroup $\langle f \rangle$ are

$$\begin{aligned} f &= (1\ 2\ 3\ 4\ 5\ 6), \\ f^2 &= (1\ 3\ 5)(2\ 4\ 6), \\ f^3 &= (1\ 4)(2\ 5)(3\ 6), \\ f^4 &= (1\ 5\ 3)(2\ 6\ 4), \\ f^5 &= (1\ 6\ 5\ 4\ 3\ 2), \\ f^6 &= (1)(2)(3)(4)(5)(6). \end{aligned}$$

Since $f^6 = e$, we have $f^5 = f^{-1}$, $f^4 = f^{-2}$, and $f^3 = f^{-3}$.

These powers can also be verified geometrically by labelling the vertices of a regular hexagon with integers 1–6, then rotating through $s/6$ of a turn to read off the vertex permutation f^s .

We now develop computational techniques for multiplying arbitrary permutations. To review ideas, and to highlight the need for more compact notation, we first compute a composition “naively”.

Example 13.24. Consider the transpositions $f = (1\ 3)$ and $g = (2\ 3)$. To compute fg , we need to compute $fg(a)$ for $a = 1, 2, 3$:

$$\begin{aligned} g(1) &= 1 \text{ and } f(1) = 3, \text{ so } fg(1) = f(1) = 3, \\ g(3) &= 2 \text{ and } f(2) = 2, \text{ so } fg(3) = f(2) = 2, \\ g(2) &= 3 \text{ and } f(3) = 1, \text{ so } fg(2) = f(3) = 1. \end{aligned}$$

In cycle notation, $(1\ 3)(2\ 3) = (1\ 3\ 2)$.

Tabulating values in this way quickly becomes tiresome, particularly when the cycles are longer and/or there are multiple factors in the composition. In practice, we use the following algorithm to keep track of the intermediate values when multiplying cycles.

Example 13.25. Let $f = (1\ 2\ 4\ 3)$, $g = (1\ 2\ 4)$, and $h = (2\ 3\ 4)$. We wish to compute $fgh = (1\ 2\ 4\ 3)(1\ 2\ 4)(2\ 3\ 4)$.

Plug the first letter, 1, into the rightmost cycle, $h = (2\ 3\ 4)$, obtaining $h(1) = 1$. Plug this output value into g , the next cycle to the left, obtaining $gh(1) = 2$. Plus 2 into f , obtaining $fgh(1) = 4$.

Since $fgh(1) = 4$, append 4 to the running total and use the same procedure to calculate $fgh(4)$. Since $fgh(4) = 3$, append 3 to the running total and compute $fgh(3)$. Continue in this way until all letters have been handled.

Step	Reasoning	“Running total”
0	$(1\ 2\ 4\ 3)(1\ 2\ 4)(2\ 3\ 4) = (1$ $\quad\quad\quad (2 \rightarrow 4) \quad (1 \rightarrow 2) \quad (1 \rightarrow 1)$	
1	$\overbrace{(1\ 2\ 4\ 3)}^{(4 \rightarrow 3)} \overbrace{(1\ 2\ 4)}^{(2 \rightarrow 4)} \overbrace{(2\ 3\ 4)}^{(4 \rightarrow 2)} = (1\ 4$	
2	$\overbrace{(1\ 2\ 4\ 3)}^{(1 \rightarrow 2)} \overbrace{(1\ 2\ 4)}^{(4 \rightarrow 1)} \overbrace{(2\ 3\ 4)}^{(3 \rightarrow 4)} = (1\ 4\ 3$	
3	$\overbrace{(1\ 2\ 4\ 3)}^{(3 \rightarrow 1)} \overbrace{(1\ 2\ 4)}^{(3 \rightarrow 3)} \overbrace{(2\ 3\ 4)}^{(2 \rightarrow 3)} = (1\ 4\ 3\ 2$	
4	$\overbrace{(1\ 2\ 4\ 3)}^{(3 \rightarrow 1)} \overbrace{(1\ 2\ 4)}^{(3 \rightarrow 3)} \overbrace{(2\ 3\ 4)}^{(2 \rightarrow 3)} = (1\ 4\ 3\ 2).$	

Example 13.26. Generally, a product of cycles might not be a single cycle. Let $f = (1\ 2\ 4\ 5\ 3)$ and $g = (1\ 2\ 4\ 5)$. We wish to compute fg .

Step	Reasoning	“Running total”
0	$(1\ 2\ 4\ 5\ 3)(1\ 2\ 4\ 5) = (1$ $\quad\quad\quad (2 \rightarrow 4) \quad (1 \rightarrow 2)$	
1	$\overbrace{(1\ 2\ 4\ 5\ 3)}^{(5 \rightarrow 3)} \overbrace{(1\ 2\ 4\ 5)}^{(4 \rightarrow 5)} = (1\ 4$	
2	$\overbrace{(1\ 2\ 4\ 5\ 3)}^{(3 \rightarrow 1)} \overbrace{(1\ 2\ 4\ 5)}^{(3 \rightarrow 3)} = (1\ 4\ 3$	
3	$\overbrace{(1\ 2\ 4\ 5\ 3)}^{(4 \rightarrow 5)} \overbrace{(1\ 2\ 4\ 5)}^{(2 \rightarrow 4)} = (1\ 4\ 3)(2$	
4	$\overbrace{(1\ 2\ 4\ 5\ 3)}^{(1 \rightarrow 2)} \overbrace{(1\ 2\ 4\ 5)}^{(5 \rightarrow 1)} = (1\ 4\ 3)(2\ 5$	
5	$\overbrace{(1\ 2\ 4\ 5\ 3)}^{(1 \rightarrow 2)} \overbrace{(1\ 2\ 4\ 5)}^{(5 \rightarrow 1)} = (1\ 4\ 3)(2\ 5).$	

In this product, a cycle “closes up” before all the elements of X have been examined. In Step 3, pick the first element not seen so far (here 2, the first element not in the subset $\{1, 3, 4\}$) and continue from there.

It may happen that some element j is mapped to itself under a product of cycles. In that case, the 1-cycle (j) may be included as a factor in the product, or omitted, as dictated by convenience.

Example 13.27. Calculate the indicated products of cycles and check your results against the answers provided.

$$(3\ 4)(2\ 3\ 5)(3\ 4) \quad (1\ 2\ 3\ 4)(1\ 3\ 2\ 4) \quad (2\ 4\ 3)(2\ 5\ 3)(1\ 3\ 5\ 4)(1\ 2\ 5)$$

Answers: $(2\ 4\ 5)$, $(1\ 4\ 2)$, $(1\ 5\ 4)(2\ 3)$.

Theorem 13.28. Let $f = c_1 c_2 \dots c_\ell$ be a product of disjoint cycles in S_X . For every integer s ,

$$f^s = c_1^s c_2^s \dots c_\ell^s.$$

The order of f is the least common multiple of the lengths of the cycles c_1, \dots, c_ℓ .

Proof. By Theorem 13.16, disjoint cycles commute. By Exercise 7.21, $(ab)^s = a^s b^s$ if $ba = ab$; this result extends easily to products of more than two factors. For each s , the permutations c_r^s have disjoint supports, so the product f^s is the identity if and only if c_r^s is the identity for $r = 1, 2, \dots, \ell$. The smallest positive integer with this property is the least common multiple of the cycle lengths. \square

Example 13.29. The permutation $f = (1\ 3\ 2\ 7)(5\ 8\ 6)$ has order 12, the least common multiple of 4 and 3.

The permutation $g = (2\ 6\ 3\ 4)(5\ 7)$ has order 4, the least common multiple of 4 and 2.

Example 13.30. To compute the order of $f = (2\ 3)(1\ 3\ 5\ 4)(2\ 4\ 3)$, a product of *non-disjoint* cycles, first multiply out to get $f = (1\ 2)(3)(4\ 5)$; the order of f is 2, not 12.

Conjugacy

The names of “letters” (objects being permuted) do not affect the structure of a permutation. Viewed as elements of S_7 , the permutations

$$f = (1\ 2\ 3)(4\ 5) \quad \text{and} \quad f' = (4\ 7\ 6)(2\ 5)$$

each represent the composition of a 3-cycle and a disjoint transposition. The relationship between two such permutations is analogous to the relationship between isomorphic groups; the distinction is more a matter of notation than of structure. If we relabel letters suitably, we can “convert” f into f' .

Here, consider the permutation $g = (1\ 4\ 2\ 7)(3\ 6)$ in S_7 , which maps the letters in f to the corresponding letters in f' . We have $f' = gfg^{-1}$. This may be checked by direct computation, but can also be understood conceptually as a “relabeling principle”. Since g maps an “old” set of labels to a “new” set, the composition gfg^{-1} first performs g^{-1} (“replace new names by their old labels”), then f (“permute the old labels”), then g (“restore the original names”). The overall result is to “apply f to a renamed set of labels”.

Definition 13.31. Assume $f, g \in S_X$. The permutations f and gfg^{-1} are said to be *conjugate* in S_X .

Example 13.32. Let $f = (1\ 2\ 3\ 4\ 5)$ and $g = (2\ 6)(3\ 8)$. The permutation g exchanges the names 2 and 6, and 3 and 8, so by the relabeling principle,

$$gfg^{-1} = (2\ 6)(3\ 8)(1\ 2\ 3\ 4\ 5)(2\ 6)(3\ 8) = (1\ 6\ 8\ 4\ 5),$$

the cyclic permutation of five elements obtained by taking the original f and replacing 2 by 6 and 3 by 8. You should verify this claim by multiplying cycles.

Example 13.33. Generally, if $f = (a_1\ a_2\ \dots\ a_k)$ is a k -cycle and g is an arbitrary permutation, then

$$gfg^{-1} = (g(a_1)\ g(a_2)\ \dots\ g(a_k)),$$

the k -cycle permuting the *values* of g on the indices appearing in f .

Theorem 13.34. If $f, g \in S_X$, then $(gfg^{-1})^s = gf^sg^{-1}$ for all s .

Proof. The mapping $\phi : S_X \rightarrow S_X$ defined by $\phi(f) = gfg^{-1}$ is an automorphism by Exercise 12.18. The theorem follows immediately. (A direct proof using mathematical induction to handle the case $s \geq 0$ and the reverse order law for $s < 0$ can be easily given.) \square

Generators for the Symmetric Group

Let S_n be the group of permutations of $X = \{1, \dots, n\}$. In this section we give sets of permutations that generate S_n .

Lemma 13.35. *The k -cycle $(a_1 a_2 \dots a_k)$ may be written as a product of $(k - 1)$ transpositions, and no fewer.*

Proof. By the cycle multiplication algorithm,

$$(a_1 a_2 \dots a_k) = (a_1 a_k)(a_1 a_{k-1}) \dots (a_1 a_3)(a_1 a_2).$$

To see that this product is “optimal”, think of the letters $1, \dots, k$ as “vertices”, and a transposition (ℓm) as an “edge” joining ℓ and m . A set of transpositions cannot generate a k -cycle unless each vertex can be joined (by a contiguous path of edges) to every other vertex. This requires at least $(k - 1)$ edges. \square

Remark 13.36. Every transposition is its own inverse, i.e., has order 2 as a group element. By the reverse order law, Exercise 7.16, the inverse of a cycle may be written as the product of the same transpositions in the opposite order:

$$(a_1 a_2 \dots a_k)^{-1} = (a_1 a_2)(a_1 a_3) \dots (a_1 a_{k-1})(a_1 a_k).$$

Lemma 13.37. *An arbitrary transposition (ℓm) with $\ell \neq m$ may be factored as a product of an odd number of transpositions $(k k + 1)$ of adjacent letters.*

Proof. By the relabeling principle or direct calculation,

$$(\ell m) = (1 \ell)(1 m)(1 \ell).$$

But for each letter $k = 1, 2, \dots, n - 1$, we have

$$(1 k + 1) = (1 k)(k k + 1)(1 k).$$

A straightforward induction shows that $(1 k)$ can therefore be expressed as a product of $(2k - 3)$ transpositions of adjacent letters, so (ℓm) can be expressed as a product of $2(2\ell - 3) + (2m - 3)$ transpositions, and this number is odd. \square

Theorem 13.38. *The following sets of permutations generate S_n .*

- (i) *The set of all transpositions.*

- (ii) The set of transpositions $(1\ k)$ with $k = 2, 3, \dots, n$.
- (iii) The set of transpositions $(k\ k+1)$ with $k = 1, 2, \dots, n-1$.
- (iv) The set $\{(1\ 2\ \dots\ n), (1\ 2)\}$.

Proof. (i) Every permutation is a product of disjoint cycles by Theorem 13.21. By Lemma 13.35, every cycle is a product of transpositions.

(ii) and (iii) By the proof of Lemma 13.37, every transposition can be written as a product of transpositions of the indicated special form.

(iv) For convenience, set $\sigma = (1\ 2\ \dots\ n)$, and $\tau = (1\ 2)$. Direct calculation, see Example 13.33, shows

$$\sigma\tau\sigma^{-1} = (2\ 3), \quad \sigma^2\tau\sigma^{-2} = (3\ 4), \quad \dots, \quad \sigma^k\tau\sigma^{-k} = (k+1\ k+2)$$

for all integers k . In particular, $\langle\sigma, \tau\rangle$ contains every transposition of adjacent elements, so by (iii) $\langle\sigma, \tau\rangle = S_n$. \square

Remark 13.39. By (iii), every permutation on n letters can be effected by exchanging pairs of adjacent letters. This fact is sometimes called the *librarian's nightmare*.

Remark 13.40. The set of two permutations in (iv) is the smallest possible generating set if $n \geq 3$. A single permutation generates a cyclic (hence Abelian) subgroup, but S_n is non-Abelian for $n \geq 3$.

13.3 Parity and the Alternating Group

Let $n \geq 2$, and consider the space \mathbf{R}^n of ordered n -tuples of real numbers. A typical element of \mathbf{R}^n is written $x = (x_1, x_2, \dots, x_n)$. Define the “sign polynomial”

$$\begin{aligned} s(x) &= \prod_{i < j} (x_j - x_i) \\ &= (x_2 - x_1) \dots (x_n - x_1) (x_3 - x_2) \dots (x_n - x_2) \dots (x_n - x_{n-1}), \end{aligned}$$

the product of all differences of monomials whose first term has larger index than the second term.

A permutation f “acts on” x , or on a polynomial s , by “re-indexing variables”:

$$\begin{aligned} x_f &= (x_{f^{-1}(1)}, x_{f^{-1}(2)}, \dots, x_{f^{-1}(n)}), \\ s_f(x) &= s(x_f) = \prod_{i < j} (x_{f^{-1}(j)} - x_{f^{-1}(i)}). \end{aligned}$$

The reason for applying f^{-1} instead of f itself is technical, see Theorem 13.44 (i) below.

Example 13.41. Let $n = 3$. The sign polynomial is

$$s(x) = (x_2 - x_1)(x_3 - x_1)(x_3 - x_2),$$

and the action of each permutation is

$$\begin{aligned} f = (1)(2)(3) & \quad s_f(x) = (x_2 - x_1)(x_3 - x_1)(x_3 - x_2) = s(x), \\ f = (1\ 2)(3) & \quad s_f(x) = (x_1 - x_2)(x_3 - x_2)(x_3 - x_1) = -s(x), \\ f = (1\ 3)(2) & \quad s_f(x) = (x_2 - x_3)(x_1 - x_3)(x_1 - x_2) = -s(x), \\ f = (1)(2\ 3) & \quad s_f(x) = (x_3 - x_1)(x_2 - x_1)(x_2 - x_3) = -s(x), \\ f = (1\ 2\ 3) & \quad s_f(x) = (x_1 - x_3)(x_2 - x_3)(x_2 - x_1) = s(x), \\ f = (1\ 3\ 2) & \quad s_f(x) = (x_3 - x_2)(x_1 - x_2)(x_1 - x_3) = s(x). \end{aligned}$$

Each polynomial s_f differs from s in “at worst” a sign change.

The behavior of the preceding example, in which each permutation f changes the sign polynomial by at most a sign, is no accident.

Proposition 13.42. *Let s denote the sign polynomial in n variables. If f is an arbitrary permutation in S_n , then $s_f = \pm s$.*

Proof. The monomial factors of s are of the form $(x_j - x_i)$ with $i \neq j$. After applying f , the monomial factors of s_f are $(x_{f^{-1}(j)} - x_{f^{-1}(i)})$. Since f^{-1} is a bijection on the set of indices, f^{-1} is also a bijection on the set of pairs of distinct indices. In other words, s_f is a product of the same monomial factors as s , re-ordered and possibly multiplied by -1 . \square

Definition 13.43. The *sign* or *parity* of a permutation f is the number $\text{sgn}(f) = (-1)^f$, equal to 1 or -1 , satisfying $s_f(x) = \text{sgn}(f) s(x)$. We say f is an *even permutation* if $\text{sgn}(f) = 1$, and an *odd permutation* if $\text{sgn}(f) = -1$.

Theorem 13.44. *Let f and g be permutations on n letters.*

- (i) $x_{fg} = (x_f)_g$. In particular, $\text{sgn}(fg) = \text{sgn}(f) \text{sgn}(g)$.
- (ii) If f is a transposition, then $\text{sgn}(f) = -1$.
- (iii) If f can be written as a product of k transpositions, $\text{sgn}(f) = (-1)^k$.

Proof. (i) By the reverse order law, $(fg)^{-1} = g^{-1}f^{-1}$. Thus

$$\begin{aligned} xfg &= (x_{(fg)^{-1}(1)}, x_{(fg)^{-1}(2)}, \dots, x_{(fg)^{-1}(n)}) \\ &= (x_{g^{-1}f^{-1}(1)}, x_{g^{-1}f^{-1}(2)}, \dots, x_{g^{-1}f^{-1}(n)}) \\ &= (x_{f^{-1}(1)}, x_{f^{-1}(2)}, \dots, x_{f^{-1}(n)})_g \\ &= (x_f)_g. \end{aligned}$$

(ii) By Lemma 13.37, an arbitrary transposition $(\ell \ m)$ with $\ell \neq m$ may be factored as a product of an odd number of transpositions of the type $(k \ k+1)$. It therefore suffices to show each transposition $f = (k \ k+1)$ changes the sign of s . To this end, we analyze the effect of $f = f^{-1}$ on the monomials $(x_j - x_i)$ with $i < j$.

Figure 13.1 depicts the set of indices (i, j) with $1 \leq i < j \leq n$. The dot at (i, j) corresponds to the factor $(x_j - x_i)$ in the sign polynomial s . The circled dot is $(k, k+1)$, the pair of indices exchanged by f . The “shaded” dots are index pairs for which exactly one of i or j is equal to k or $k+1$.

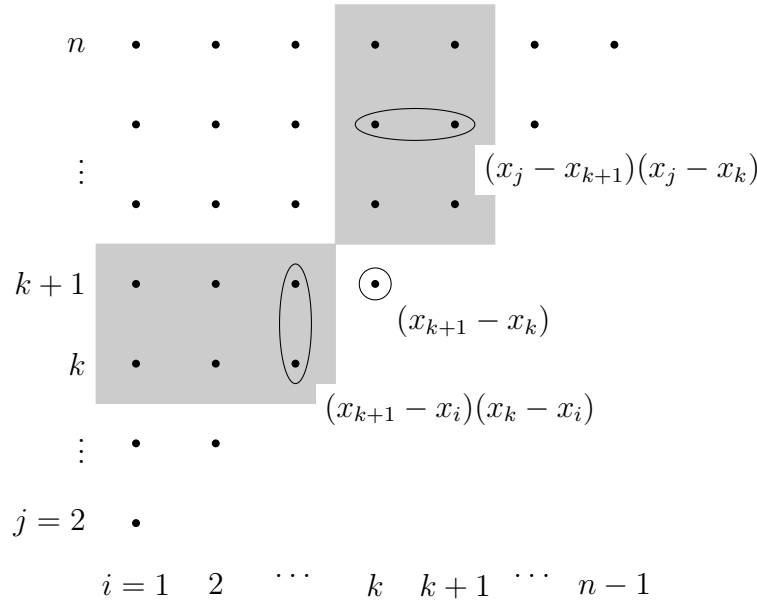


Figure 13.1: Index pairs (i, j) with $1 \leq i < j \leq n$.

The index pair $(k, k+1)$ corresponds to the factor $(x_{k+1} - x_k)$, whose sign is changed by f .

The leftmost shaded box corresponds to index pairs for which $i < k$ and $j = k$ or $k + 1$. When these factors of s are paired as indicated, the corresponding product of factors, $(x_{k+1} - x_i)(x_k - x_i)$, is unchanged by f . The product of all such terms is therefore unchanged by f .

The rightmost shaded box consists of terms for which $k + 1 < j$ and $i = k$ or $k + 1$. Again, the indicated products of pairs of terms are unchanged by f , so the entire product is unchanged by f .

Unshaded dots are index pairs for which neither i nor j is equal to k or $k + 1$. For these, $f(i) = i$ and $f(j) = j$, so $(x_j - x_i)$ is unchanged by f .

The total sign change among all the factors of s comes from the first case, the monomial factor $(x_{k+1} - x_k)$.

(iii) Follows immediately from (i) and (ii). \square

Corollary 13.45. *A permutation f in S_n can be written as a product of an even number of transpositions if and only if $\text{sgn}(f) = 1$.*

No permutation can be expressed as both a product of an even number and a product of an odd number of transpositions.

Corollary 13.46. *A k -cycle is an even permutation if and only if the length k is odd.*

Remark 13.47. This disparity, as it were, between the length of a cycle and its parity as a permutation, is an unavoidable fact of life.

Corollary 13.48. *The set A_n of even permutations on n letters is a subgroup of S_n that contains $\frac{1}{2}n!$ elements.*

Proof. (A_n is closed under composition). If $f, g \in A_n$, then by Theorem 13.44 (i), $\text{sgn}(fg) = \text{sgn}(f)\text{sgn}(g) = 1$. That is, $fg \in A_n$.

(A_n is closed under inversion). If an arbitrary permutation f is expressed as a product of transpositions, then f^{-1} is the product of the same transpositions taken in reverse order, so $\text{sgn}(f^{-1}) = \text{sgn}(f)$. In particular, if $f \in A_n$, then $f^{-1} \in A_n$.

By Theorem 7.32, A_n is a subgroup of S_n .

(Counting elements of A_n). Consider the transposition $\tau = (1\ 2)$. The mapping $f \mapsto \tau f$ is a bijection of S_n to itself, and changes the parity of every permutation, i.e., maps A_n to $S_n \setminus A_n$; consequently, A_n contains half as many elements as S_n . \square

Definition 13.49. The set A_n of even permutations on n letters is called the *alternating group* (on n letters).

Example 13.50. $A_2 = \{(1)\}$. $A_3 = \{(1)(2)(3), (1\ 2\ 3), (1\ 3\ 2)\}$.

Exercises

Exercise 13.1. On a set X with ten elements, what percentage of mappings are permutations? On a set with 100 elements?

Exercise 13.2. Express $(1\ 6\ 4\ 2\ 5\ 3)$ as a product of transpositions.

Exercise 13.3. Let $X = \{\alpha, \beta, \gamma\}$. Express each element of S_X as a product of disjoint cycles:

$$(a) (\alpha\ \beta)(\alpha\ \gamma) \qquad (b) (\alpha\ \gamma)(\alpha\ \beta) \qquad (c) (\alpha\ \beta)(\alpha\ \gamma)(\alpha\ \beta)$$

Exercise 13.4. In each part, $a = (1\ 2)(3\ 4)$, $b = (1\ 3)(2\ 4)$, and $c = ab$.

- (a) Express the products ab and ba as products of disjoint cycles. Do a and b commute?
- (b) Make a Cayley table for the set $G = \{e, a, b, c\}$ under composition.
- (c) Prove G is a group under composition, isomorphic to $(\mathbf{Z}_2 \times \mathbf{Z}_2, +)$, the Klein 4-group.

Exercise 13.5. Calculate the following:

$$(a) (3\ 4)(1\ 2\ 3)(3\ 4) \qquad (b) (2\ 4)(1\ 2\ 3)(2\ 4) \qquad (c) (2\ 4)(1\ 2\ 3)(3\ 4)$$

Exercise 13.6. In parts (a) and (b), symbols represent elements of some set.

- (a) Express $(3\ 4\ 1)(6\ 4\ 2\ 1)(1\ 2)$ as a product of disjoint cycles.
- (b) Express $(c\ d\ a)(f\ d\ b\ a)(a\ b)$ as a product of disjoint cycles.
- (c) How are the answers to parts (a) and (b) related?

Exercise 13.7. For each product of cycles, determine the order and parity.

- (a) $(1\ 2\ 3\ 4)(2\ 3\ 5\ 7)(3\ 4\ 6\ 7)$. (b) $(4\ 6)(1\ 2\ 3\ 4)(2\ 3\ 5\ 7)(3\ 4\ 6\ 7)$.
- (c) $(1\ 3\ 2\ 4)(2\ 3\ 5\ 7)(3\ 4\ 6\ 7)$. (d) $(5\ 7)(1\ 2\ 3\ 4)(2\ 3\ 5\ 7)(3\ 4\ 6\ 7)$.

Exercise 13.8. Show by example that the parity of a permutation (even or odd), and whether that permutation has even or odd order, are completely independent; knowledge of one datum implies nothing about the other.

Exercise 13.9. Following Example 13.23, compute the powers of cycles of length 3, 4, 5, and 8.

Exercise 13.10. Write out the sign polynomial s in four variables (there are six factors). By direct calculation determine the effect on s of the 3-cycle $f = (1\ 2\ 3)$, and of the 4-cycle $g = (1\ 2\ 3\ 4)$.

Exercise 13.11. Following the technique of Example 13.10, list the 24 elements of S_4 . Determine the cycle structure of each element; that is, determine whether each element is a transposition, 3-cycle, 4-cycle, or a product of disjoint transpositions. How many elements of each type are there?

Exercise 13.12. List the 12 elements of A_4 . Which cycle structures do or do not occur?

Exercise 13.13. Prove that two *transpositions* in S_X commute if and only if their supports are either disjoint or identical.

Exercise 13.14. Let $X = \{1, 2, \dots, n\}$, and let A be a subset of X having k elements. Prove that $\{f \text{ in } S_n : f(A) = A\}$, the *stabilizer* of A , is a subgroup of S_n isomorphic to the direct product $S_k \times S_{n-k}$.

Exercise 13.15. Show that S_5 contains cyclic subgroups of order 6. Does S_5 contain cyclic subgroups with more than six elements?

Exercise 13.16. In the symmetric group S_{10} , find the largest order of a cyclic subgroup, and give an example of an element of that order.

Exercise 13.17. This exercise introduces the 3×3 *permutation matrices*.

If $f \in S_3$ and A_f is the 3×3 matrix with $a_{f(j),j} = 1$ for $j = 1, 2, 3$ and $a_{ij} = 0$ otherwise.

- (a) For each element f of S_3 , find the corresponding matrix A_f .
- (b) If $\mathbf{e}_1 = (1, 0, 0)$, $\mathbf{e}_2 = (0, 1, 0)$, and $\mathbf{e}_3 = (0, 0, 1)$ are the standard basis vectors in \mathbf{R}^3 , prove that $A_f \mathbf{e}_i = \mathbf{e}_{f(i)}$. That is, multiplication by A_f permutes the standard basis vectors.
- (c) Prove that if $f, g \in S_3$, then $A_{fg} = A_f A_g$. Conclude that the set of 3×3 permutation matrices is a group under matrix multiplication, isomorphic to S_3 .

Chapter 14

Examples of Finite Groups

Definition 14.1. Let X be a non-empty set. The *symmetric group* S_X is the set of permutations of X equipped with the operation of mapping composition. If $X = \{1, 2, \dots, n\}$ is the “standard” set of n elements, we write $S_X = S_n$.

14.1 Cayley’s Theorem

By Theorem 14.9 below, a finite group (G, \cdot) of order n is isomorphic to a subgroup of $S_G \simeq S_n$. In this sense, the symmetric groups “contain” all of finite group theory. An example will illustrate the main idea.

Example 14.2. Consider the Cayley table for $(\mathbf{Z}_4, +)$. Associate to each row of the table the corresponding permutation of elements:

$+$	0	1	2	3	
0	0	1	2	3	$\phi_0 = (0)(1)(2)(3)$
1	1	2	3	0	$\phi_1 = (0\ 1\ 2\ 3)$
2	2	3	0	1	$\phi_2 = (0\ 2)(1\ 3)$
3	3	0	1	2	$\phi_3 = (0\ 3\ 2\ 1)$

Each *element* of the group $(\mathbf{Z}_4, +)$ maps to a *permutation* of the set \mathbf{Z}_4 . The resulting set of four permutations is a cyclic group of order 4, namely is isomorphic to the original group $(\mathbf{Z}_4, +)$.

Definition 14.3. Let (G, \cdot) be a group and X a set. An *action* of (G, \cdot) on X is a mapping $\phi : G \rightarrow S_X$ such that

$$(14.1) \quad \phi(a \cdot b) = \phi(a)\phi(b) \quad \text{for all } a \text{ and } b \text{ in } G.$$

A group action is *faithful* if ϕ is one-to-one, namely, if distinct elements of G correspond to different permutations.

Remark 14.4. A faithful group action associates a permutation of X to each element of G in such a way that mapping composition “implements” the group operation.

Since $\phi(a)$ is a *bijection* of X (rather than an element of X), we usually write ϕ_a , allowing us to use $\phi_a(x)$ to denote the image of x ($x \in X$) under the bijection associated to the element a of G . In this notation, the morphism property (14.1) reads $\phi_{a \cdot b} = \phi_a \phi_b$.

Proposition 14.5. *Let $\phi : (G, \cdot) \rightarrow S_X$ be a group action. The image of ϕ is a subgroup of S_X . If ϕ is faithful, the image of ϕ is isomorphic to (G, \cdot) .*

Proof. Let ϕ_a and ϕ_b be arbitrary elements of the image. By (14.1), their composition is $\phi_{a \cdot b}$, which is also an element of the image: The image of ϕ is closed under composition. The identity transformation I_X is equal to ϕ_e by the same argument used to prove Theorem 12.2, and similarly, $\phi_{a^{-1}} = (\phi_a)^{-1}$. The image therefore contains an identity element and the inverse of each of its elements, and so is a group under mapping composition by Theorem 7.32.

If ϕ is injective, then ϕ is a bijection to its image, and therefore an isomorphism. \square

We now turn to *Cayley’s theorem*, a concrete scheme for representing an arbitrary group as a group of permutations *on its own set of elements*.

Lemma 14.6. *Let (G, \cdot) be a group. For each a in G , define $\ell_a : G \rightarrow G$ by $\ell_a(x) = a \cdot x$ for x in G . The mapping ℓ_a is a bijection, i.e., $\ell_a \in S_G$.*

Proof. (Injectivity). If $a \cdot x_1 = \ell_a(x_1) = \ell_a(x_2) = a \cdot x_2$, then $x_1 = x_2$ by the cancellation law.

(Surjectivity). For arbitrary b in G , $\ell_a(a^{-1}b) = b$. \square

Definition 14.7. Let (G, \cdot) be a group. The mapping $\ell : G \rightarrow S_G$ defined by $\ell(a) = \ell_a$ is called the *natural left action* of (G, \cdot) .

Remark 14.8. There is an analogous *natural right action* r , defined by $r_a(x) = xa^{-1}$. The inversion of a is required for the morphism property.

Theorem 14.9 (Cayley's theorem). *The natural left action $\ell : G \rightarrow S_G$ is a faithful action of (G, \cdot) .*

Proof. (ℓ is a group action). Fix a in G , and let $\ell_a(x) = a \cdot x$ for x in G . Equation (14.1) becomes the associative law in (G, \cdot) :

$$\ell_{a \cdot b}(x) = (a \cdot b) \cdot x = a \cdot (b \cdot x) = \ell_a(b \cdot x) = \ell_a \ell_b(x).$$

Since this equation holds for arbitrary x in G , $\ell_{a \cdot b} = \ell_a \ell_b$ as mappings.

(ℓ is faithful). If $\ell_a(x) = \ell_b(x)$ for even a *single* x , then $a = b$ by the cancellation law.

By Proposition 14.5, (G, \cdot) is isomorphic to a subgroup of S_G . \square

14.2 The Dihedral Groups

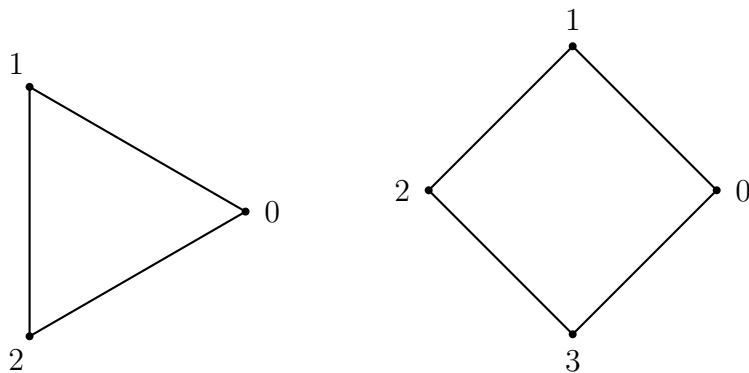
Definition 14.10. Let $n \geq 3$, and let P_n be a regular planar n -gon. The set D_n of symmetries of P_n forms a group under mapping composition, called a *dihedral group*.

Theorem 14.11. *Let P_n be a regular n -gon with $n \geq 3$. The dihedral group D_n is a non-Abelian group of order $2n$. If a is a counterclockwise rotation by $1/n$ of a turn and b is reflection across a diameter through some vertex of P_n , then a has order n , b has order 2, the elements of D_n are*

$$e, a, a^2, \dots, a^{n-1}, \quad b, ab, a^2b, \dots, a^{n-1}b,$$

and $ba = a^{-1}b$.

Before tackling the general case, let's look at triangles and squares. Place one vertex at the rightmost extremity, and label it 0. Number the remaining vertices counterclockwise.



Any symmetry permutes the vertices, and our analysis focuses on determining how the vertices might be permuted.

Example 14.12 (Symmetries of an equilateral triangle). Rotating the triangle one-third of a turn cyclically permutes the vertices. Call this symmetry $a = (0\ 1\ 2)$. Performing a three times returns the triangle to its original configuration: $a^3 = e$, and therefore $a^2 = a^{-1}$.

Another symmetry is obtained by reflecting the triangle across the horizontal axis, fixing vertex 0 and exchanging 1 and 2; call this symmetry $b = (1\ 2)$. Since reflecting twice is the identity map, $b^2 = e$. The symmetries

$$ab = (0\ 1\ 2)(1\ 2) = (0\ 1)$$

(reflect, then rotate; the rightmost factor acts first) and

$$ba = (1\ 2)(0\ 1\ 2) = (0\ 2) = (0\ 2\ 1)(1\ 2) = a^2b$$

(rotate, then reflect) geometrically correspond to reflecting across the axis through vertex 2 and reflecting across the axis through vertex 1.

There are six permutations of three vertices, so the symmetries

$$e, a, a^2, b, ab, a^2b$$

must enumerate D_3 . Since $ba = a^2b \neq ab$, the group of symmetries of an equilateral triangle is non-Abelian.

Because $b = b^{-1}$ and $a^2 = a^{-1}$, the identity $ba = a^2b$ can be written $bab^{-1} = a^{-1}$ or $ba = a^{-1}b$.

Proposition 14.13. *The dihedral group D_3 is isomorphic to the symmetric group S_3 .*

Example 14.14 (Symmetries of a square). Let a be a one-quarter rotation counterclockwise, and let b be reflection across the horizontal axis. In cycle notation, with the vertices numbered as above, $a = (0\ 1\ 2\ 3)$ and $b = (1\ 3)$. The rotation a has order 4, so $a^3 = a^{-1}$ and brief calculations give

$$ba = (1\ 3)(0\ 1\ 2\ 3) = (0\ 3)(1\ 2) = (0\ 3\ 2\ 1)(1\ 3) = a^3b.$$

Again, this identity may be written $bab^{-1} = a^{-1}$ or $ba = a^{-1}b$.

We have found eight symmetries of the square:

$$e, a, a^2, a^3, b, ab, a^2b, a^3b.$$

No two distinct powers of a can be equal since a has order 4. By the cancellation law, no power of a can be equal to a power of a times b , since b is not a power of a . Geometrically, powers of a are rotations of the square, and computation of the remaining four elements shows they are reflections about the four axes of the square.

It's plausible we've listed all symmetries of the square. To prove this, reckon as follows: Every symmetry of the square maps vertex 0 to another vertex, and there are four possibilities. Once the image of 0 is known, there is a two-fold ambiguity, since the square may be reflected across the diagonal axis through the image of 0. Thus, we expect $4 \times 2 = 8$ symmetries of the square; these must be the eight elements listed above.

Remark 14.15. The identity $bab^{-1} = a^{-1}$ has geometric significance. Imagine two “observers”, A^+ viewing the square “from above”, so the vertices are numbered counterclockwise, and A^- viewing the square “from below”, so the vertices are numbered clockwise. The transformation b reflects the square, exchanging the notions of clockwise and counterclockwise.

Now, the transformation seen by A^+ as a counterclockwise quarter turn is seen by A^- as a clockwise quarter turn. In other words, performing bab^{-1} —changing viewpoint, rotating counterclockwise, then changing viewpoint again—is the same as a^{-1} , a clockwise rotation. If this isn't clear, make a square from paper, label the vertices, and physically manipulate the paper.

Remark 14.16. For both the triangle ($n = 3$) and the square ($n = 4$), there were $2n$ symmetries. These were of two types: (i) n rotational symmetries—integer multiples of $1/n$ of a turn, and (ii) n reflections about the diameters passing through a vertex and/or an edge midpoint. Respectively, these symmetries correspond to ending configurations in which the vertices are numbered counterclockwise or clockwise. If a represents a specific rotation by $1/n$ of a turn and b is a specific reflection, the $2n$ symmetries are expressions of the form $a^k b^\ell$, with $0 \leq k < n$, $0 \leq \ell < 2$, and $bab^{-1} = a^{-1}$. Analogous conclusions carry over to a general n -gon.

Proof of Theorem 14.11. Let P_n be a regular n -gon with vertices cyclically numbered $0, 1, \dots, n-1$, and let ϕ be a symmetry. The images

$\phi(0)$ and $\phi(1)$ must be adjacent vertices, since 0 and 1 are adjacent. There are n possibilities for $\phi(0)$, but once $\phi(0)$ is known there are only two choices for $\phi(1)$, namely the two vertices adjacent to $\phi(0)$. Consequently, P_n has at most $2n$ symmetries.

Once both $\phi(0)$ and $\phi(1)$ are known, the symmetry is completely determined: The image of vertex 2 must be the neighbor of $\phi(1)$ not equal to $\phi(0)$, and so on inductively. If $\phi(1)$ lies $1/n$ of a turn counterclockwise from $\phi(0)$, the vertices are numbered counterclockwise.

Let a be counterclockwise rotation through $1/n$ of a turn and b be reflection across the diameter through vertex 0. In cycle notation,

$$a = (0\ 1\ 2\ \dots\ n-1), \quad b = (1\ n-1)(2\ n-2)\dots$$

The $2n$ distinct symmetries $a^k b^\ell$, with $0 \leq k < n$ and $\ell = 0, 1$, enumerate D_n , the set of all symmetries of the n -gon.

Finally, $bab^{-1}(0) = a^{-1}(0)$ and $bab^{-1}(1) = a^{-1}(1)$: The clockwise rotation a^{-1} maps 0 to $n-1$ and 1 to 0, while

$$\begin{aligned} bab^{-1}(0) &= bab(0) = ba(0) = b(1) = n-1, \\ bab^{-1}(1) &= bab(1) = ba(n-1) = b(0) = 0. \end{aligned}$$

Since bab^{-1} and a^{-1} agree on 0 and 1, they are the same symmetry. \square

Remark 14.17. The identity $bab^{-1} = a^{-1}$ in D_n , together with the law $(bab^{-1})^r = ba^r b^{-1}$ from Exercise 7.18, implies $ba^r b^{-1} = a^{-r}$, or

$$ba^r = a^{-r}b = a^{n-r}b \quad \text{for } r \text{ in } \mathbf{Z}.$$

Example 14.18. Label the vertices of a regular hexagon $0, 1, \dots, 5$ with 0 and 3 on a horizontal axis. In D_6 , the symmetry group of a regular hexagon, let a denote $1/6$ of a turn counterclockwise and b reflection over the horizontal axis. In cycle notation,

$$a = (0\ 1\ 2\ 3\ 4\ 5), \quad b = (1\ 5)(2\ 4).$$

The group D_6 has twelve elements, six rotations and six reflections:

$$e, a, a^2, a^3, a^4, a^5, \quad b, ab, a^2b, a^3b, a^4b, a^5b.$$

Products may be computed from the identity $ba^r = a^{-r}b$. For example,

$$(a^4b)(a^2b) = a^4(ba^2)b = a^4(a^{-2}b)b = a^2.$$

Remark 14.19. Since each symmetry of the hexagon permutes the 6 vertices, the dihedral group D_6 is isomorphic to a subgroup of the symmetric group S_6 . This representation is different from the action coming from Cayley's theorem, which implements D_6 in S_{12} .

14.3 The Quaternion Group

The *quaternion group* Q , see Exercise 11.37, is the non-Abelian group of order 8 containing elements $\{\pm 1, \pm i, \pm j, \pm k\}$ that satisfy

$$i^2 = j^2 = k^2 = -1 \quad \text{and} \quad ij = k.$$

Since -1 is a power of each non-identity element, -1 commutes with everything in Q . Further, $i^{-1} = i^3 = -i$, and similarly for j and k .

Starting from $ij = k$, the reverse order law gives

$$-k = k^{-1} = (ij)^{-1} = j^{-1}i^{-1} = (-j)(-i) = ji;$$

i and j do not commute, but “anti-commute”.

The products kj and ik can be computed using $ij = k$:

$$\begin{aligned} -j &= (i^2)j = i(ij) = ik, & (\text{multiplying } ij = k \text{ on the left by } i), \\ -i &= i(j^2) = (ij)j = kj, & (\text{multiplying } ij = k \text{ on the right by } j). \end{aligned}$$

Similarly to the preceding paragraph, the reverse order law implies $jk = i$ and $ki = j$.

\cdot	1	-1	i	$-i$	j	$-j$	k	$-k$
1	1	-1	i	$-i$	j	$-j$	k	$-k$
-1	-1	1	$-i$	i	$-j$	j	$-k$	k
i	i	$-i$	-1	1	k	$-k$	$-j$	j
$-i$	$-i$	i	1	-1	$-k$	k	j	$-j$
j	j	$-j$	$-k$	k	-1	1	i	$-i$
$-j$	$-j$	j	k	$-k$	1	-1	$-i$	i
k	k	$-k$	j	$-j$	$-i$	i	-1	1
$-k$	$-k$	k	$-j$	j	i	$-i$	1	-1

Table 14.1: The Cayley table for the quaternion group Q .

14.4 Subgroups of S_4

Example 14.20. There are 24 permutations on four letters. As in Chapter 13, these can be listed by enumerating all orderings of the digits 1–4, viewing each ordering as a permutation, and converting to cycle notation. For best results, work out these products yourself, then check your answers below.

$$\begin{aligned}
 1\ 2\ 3\ 4 &\leftrightarrow (1)(2)(3)(4), & 1\ 3\ 2\ 4 &\leftrightarrow (1)(2\ 3)(4), & 1\ 4\ 2\ 3 &\leftrightarrow (1)(2\ 4\ 3), \\
 1\ 2\ 4\ 3 &\leftrightarrow (1)(2)(3\ 4), & 1\ 3\ 4\ 2 &\leftrightarrow (1)(2\ 3\ 4), & 1\ 4\ 3\ 2 &\leftrightarrow (1)(2\ 4)(3), \\
 2\ 1\ 3\ 4 &\leftrightarrow (1\ 2)(3)(4), & 2\ 3\ 1\ 4 &\leftrightarrow (1\ 2\ 3)(4), & 2\ 4\ 1\ 3 &\leftrightarrow (1\ 2\ 4\ 3), \\
 2\ 1\ 4\ 3 &\leftrightarrow (1\ 2)(3\ 4), & 2\ 3\ 4\ 1 &\leftrightarrow (1\ 2\ 3\ 4), & 2\ 4\ 3\ 1 &\leftrightarrow (1\ 2\ 4)(3), \\
 3\ 1\ 2\ 4 &\leftrightarrow (1\ 3\ 2)(4), & 3\ 2\ 1\ 4 &\leftrightarrow (1\ 3)(2)(4), & 3\ 4\ 1\ 2 &\leftrightarrow (1\ 3)(2\ 4), \\
 3\ 1\ 4\ 2 &\leftrightarrow (1\ 3\ 4\ 2), & 3\ 2\ 4\ 1 &\leftrightarrow (1\ 3\ 4)(2), & 3\ 4\ 2\ 1 &\leftrightarrow (1\ 3\ 2\ 4), \\
 4\ 1\ 2\ 3 &\leftrightarrow (1\ 4\ 3\ 2), & 4\ 2\ 1\ 3 &\leftrightarrow (1\ 4\ 3)(2), & 4\ 3\ 1\ 2 &\leftrightarrow (1\ 4\ 2\ 3), \\
 4\ 1\ 3\ 2 &\leftrightarrow (1\ 4\ 2)(3), & 4\ 2\ 3\ 1 &\leftrightarrow (1\ 4)(2)(3), & 4\ 3\ 2\ 1 &\leftrightarrow (1\ 4)(2\ 3).
 \end{aligned}$$

Every element of S_4 has order 1, 2, 3, or 4, and is either a cycle (21 elements) or a product of two disjoint transpositions (three elements).

The symmetric group S_4 contains four “copies” of S_3 (the set of permutations fixing 4 is a copy of S_3 , and similarly for the other letters); three “copies” of the dihedral group D_4 (one for each pair $\{\sigma, \sigma^{-1}\}$ of 4-cycles); and the alternating group A_4 .

Example 14.21. The set A_4 of even permutations in S_4 is a non-Abelian group of order 12 containing the identity element, the products of disjoint transpositions (three elements), and the 3-cycles (eight elements, two for each letter 1, 2, 3, and 4):

$$\begin{aligned}
 1\ 2\ 3\ 4 &\leftrightarrow (1)(2)(3)(4), & 1\ 3\ 4\ 2 &\leftrightarrow (1)(2\ 3\ 4), & 1\ 4\ 2\ 3 &\leftrightarrow (1)(2\ 4\ 3), \\
 2\ 1\ 4\ 3 &\leftrightarrow (1\ 2)(3\ 4), & 2\ 3\ 1\ 4 &\leftrightarrow (1\ 2\ 3)(4), & 2\ 4\ 3\ 1 &\leftrightarrow (1\ 2\ 4)(3), \\
 3\ 1\ 2\ 4 &\leftrightarrow (1\ 3\ 2)(4), & 3\ 2\ 4\ 1 &\leftrightarrow (1\ 3\ 4)(2), & 3\ 4\ 1\ 2 &\leftrightarrow (1\ 3)(2\ 4), \\
 4\ 1\ 3\ 2 &\leftrightarrow (1\ 4\ 2)(3), & 4\ 2\ 1\ 3 &\leftrightarrow (1\ 4\ 3)(2), & 4\ 3\ 2\ 1 &\leftrightarrow (1\ 4)(2\ 3).
 \end{aligned}$$

For brevity, we write $\sigma_{12,34} = (1\ 2)(3\ 4)$, $\sigma_{132} = (1\ 3\ 2) = \sigma_{123}^{-1}$, etc.

One way to study a group is to enumerate or otherwise characterize its (proper, non-trivial) subgroups. Such an investigation always begins by understanding the cyclic subgroups, since every subgroup contains the cyclic subgroups generated by each of its elements.

A group of order N obviously has at most N cyclic subgroups, and one of these is always the trivial group $\langle e \rangle$. The “brute force” approach to enumerating the cyclic subgroups of G is therefore to enumerate the elements of G , calculate the cyclic subgroup generated by each, and finally to eliminate “repeats”—subgroups appearing more than once. Said another way, define an equivalence relation on G by declaring elements a and b to be equivalent if they generate the same cyclic subgroup, and describe the resulting partition.

If a and b are elements of G , then $\langle a \rangle = \langle b \rangle$ if and only if $b \in \langle a \rangle$ (b is a power of a) and $a \in \langle b \rangle$ (a is a power of b). In particular, each element of order 2 is in a class by itself. Generally, if a has order n , then by Theorem 12.17, a is equivalent to $b = a^k$ if and only if $a \in \langle a^k \rangle$, if and only if $\gcd(k, n) = 1$.

This process can be carried out feasibly for most of the groups in this chapter. We’ll use A_4 as an example.

Example 14.22. Among the twelve elements of A_4 , there are three elements of order 2 (each a product of disjoint transpositions) and eight elements of order 3 (each a 3-cycle).

As noted above, elements of order 2 generate distinct cyclic subgroups: $a \leftrightarrow \langle a \rangle = \{e, a\}$. The elements of order 3 divide naturally into four pairs of mutual inverses, $(1\ 3\ 2) = (1\ 2\ 3)^{-1}$ and so forth, each pair generating the same cyclic group. In summary, A_4 contains three cyclic subgroups of order 2 and four cyclic subgroups of order 3.

Generator(s)	Cyclic subgroup
e	$\{e\}$
$\sigma_{12,34}$	$\{e, \sigma_{12,34}\}$
$\sigma_{13,24}$	$\{e, \sigma_{13,24}\}$
$\sigma_{14,23}$	$\{e, \sigma_{14,23}\}$
$\sigma_{123}, \sigma_{132}$	$\{e, \sigma_{123}, \sigma_{132}\}$
$\sigma_{124}, \sigma_{142}$	$\{e, \sigma_{124}, \sigma_{142}\}$
$\sigma_{134}, \sigma_{143}$	$\{e, \sigma_{134}, \sigma_{143}\}$
$\sigma_{234}, \sigma_{243}$	$\{e, \sigma_{234}, \sigma_{243}\}$

There is also a non-cyclic (but Abelian) subgroup of order 4 containing the identity and the elements of order 2. These eight subgroups

are the only proper, non-trivial subgroups of A_4 . It will be more convenient to prove this assertion later, once we know more about properties of subgroups.

14.5 Symmetries of Polyhedra

Loosely, a *polyhedron* is a 3-dimensional analog of a polygon, namely a solid region whose boundary consists of a finite union of planar polygons. For this introductory treatment, we consider only the *Platonic solids*, convex solids for which the faces are *regular* polygons and the same number of faces meet at each vertex.

There are exactly five, Figure 14.1: The tetrahedron, octahedron, icosahedron (regular triangular faces with three, four, or five meeting at each vertex); the cube (square faces, three meeting at each vertex); and the dodecahedron (regular pentagonal faces, three meeting at each vertex).

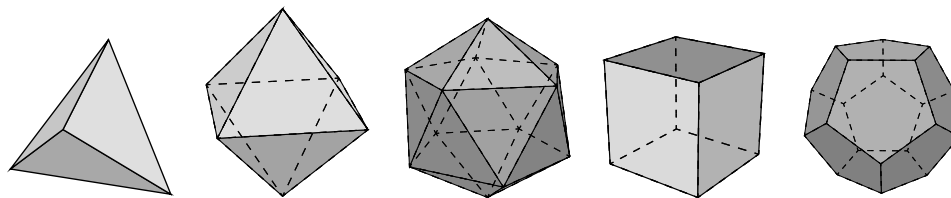


Figure 14.1: The Platonic solids.

As with planar polygons, we are interested in *symmetries*, rigid motions of space mapping a polyhedron to itself. Each polyhedron admits two types of symmetry, rotations and reflections, and the rotational symmetries form a subgroup of the full symmetry group.

Remark 14.23. Each Platonic solid P has a *dual solid* whose vertices are the centers of the faces of P . The tetrahedron is its own dual; the cube and octahedron are dual; the dodecahedron and icosahedron are dual. In this brief study of symmetries, we therefore focus on just three Platonic solids: The tetrahedron, the cube, and the dodecahedron.

The Tetrahedron

Label the vertices of the tetrahedron 0, 1, 2, 3. An arbitrary symmetry of the tetrahedron permutes the vertices. Conversely, a vertex permutation completely determines a symmetry, since the only symmetry

fixing all four vertices is the identity map. Immediately, we learn that a tetrahedron has at most $4! = 24$ symmetries.

Theorem 14.24. *A regular tetrahedron has 24 symmetries, of which 12 are rotations. The full symmetry group is isomorphic to S_4 , and the rotation group is isomorphic to A_4 .*

Proof. Consider a tetrahedron T sitting with one face in the (x, y) plane and with the fourth vertex along the z axis. Each symmetry of the triangular base gives rise to a symmetry of T , and conversely every symmetry of T fixing the fourth vertex induces a symmetry of the base.

Generally, T admits six symmetries fixing any particular vertex v : three rotations (by one-third of a turn about the axis ℓ through v) and three reflections (about any of the three planes containing ℓ and a vertex other than v).

Now fix a labeling of the vertices of T , Figure 14.2, left. Each symmetry f maps the vertex 0 to one of the four vertices of T . Once this vertex is known, the argument in the preceding paragraph shows there are six symmetries (three rotations) fixing $f(0)$. Counting up choices, there are $4 \times 6 = 24$ symmetries in all, of which $4 \times 3 = 12$ are rotations. The symmetry group of T is therefore mapped bijectively to S_4 (with f being sent to the resulting permutation of the vertices), and this map obviously satisfies the morphism condition. It remains to show the rotation group is isomorphic to A_4 . This will be accomplished by explicit calculation.

For each of the four axes through a vertex and the center of the opposite face, there are three rotation symmetries. One of these is the identity; the other two are 3-cycles on the set of vertices. This accounts for $1 + 4 \times 2 = 9$ elements in all: The identity, and eight 3-cycles.

Now, T has six edges, and therefore three axes running through midpoints of a pair of opposite edges. For each axis, the half-turn about that axis maps T to itself; the induced map on vertices is a product of disjoint transpositions. This accounts for 12 rotational symmetries, all of which are even permutations. The rotation group of T is therefore isomorphic to A_4 . \square

The Cube

The easiest way to count symmetries of a cube is not the most obvious. Rather than looking at vertices, consider the four long diagonals, each

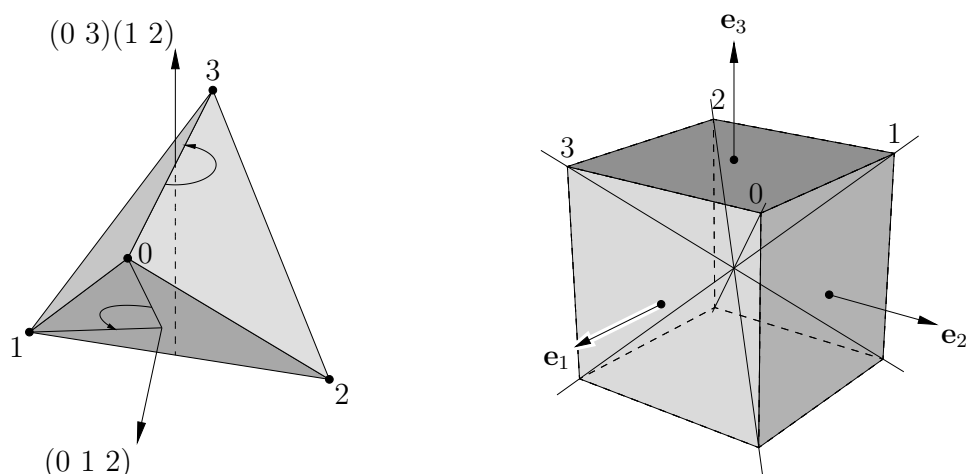


Figure 14.2: Symmetries of the tetrahedron and cube.

passing through a pair of opposite vertices, Figure 14.2, right. Every symmetry of a cube permutes these axes, and a bit of geometric consideration shows every axis permutation is achieved. Unlike the situation for the tetrahedron, however, there exists a non-trivial symmetry fixing all four axes. If the cube is centered at the origin, the symmetry is defined by the *antipodal map* $f(x, y, z) = (-x, -y, -z)$. The cube therefore has at most $2 \cdot 4! = 48$ symmetries.

Theorem 14.25. *A cube has 48 symmetries, of which 24 are rotations. The group of rotational symmetries of a cube is isomorphic to S_4 .*

The symmetry group of a cube can be represented as a set of 3×3 real (or integer) matrices. To see how, consider the standard basis vectors $\mathbf{e}_1 = (1, 0, 0)$, $\mathbf{e}_2 = (0, 1, 0)$, $\mathbf{e}_3 = (0, 0, 1)$ and their negatives. A symmetry of the cube (i) Permutes the pairs $\{\mathbf{e}_1, -\mathbf{e}_1\}$, $\{\mathbf{e}_2, -\mathbf{e}_2\}$, $\{\mathbf{e}_3, -\mathbf{e}_3\}$ (namely, the coordinate axes), and (ii) is otherwise arbitrary. There are six ways of permuting the coordinate axes, and once a permutation is chosen there are eight choices of sign, plus or minus on each axis. In total, this accounts for $6 \times 8 = 48$ symmetries.

The close relationship between symmetries of a cube and a regular tetrahedron is no accident: Inscribed in a cube are *two* regular tetrahedra. If the cube has vertices $(\pm 1, \pm 1, \pm 1)$ (with all eight choices of sign), one tetrahedron has vertices $(1, 1, 1)$, $(1, -1, -1)$, $(-1, 1, -1)$, and $(-1, -1, 1)$, and the other tetrahedron has the negatives of these points as vertices. Any symmetry of a tetrahedron therefore gives rise

to a symmetry of a cube, but only half of the cube's symmetries arise in this way. Indeed, the cube admits symmetries exchanging the two inscribed tetrahedra, accounting for an extra factor of two.

Though the full symmetry group of a tetrahedron is isomorphic to the rotation group of a cube (both are S_4), this fact has no obvious interpretation in 3-dimensional geometry. Rather, the full symmetry group of the cube contains multiple subgroups isomorphic to S_4 .

The Dodecahedron

If the key to studying symmetries of the cube was sneaky (looking at the four long diagonals rather than at the eight vertices), the dodecahedron requires a positively arcane fact: Inscribed in a dodecahedron D are *five* regular tetrahedra, permuted by each symmetry of D , and no non-

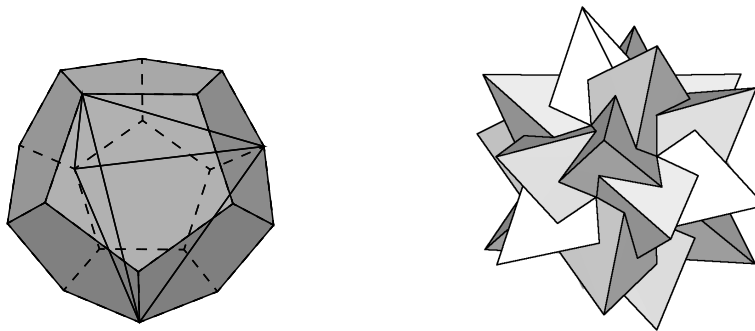


Figure 14.3: Tetrahedra inscribed in a regular dodecahedron.

trivial symmetry preserves all five. A careful analysis is more difficult to carry out only because the geometry of the dodecahedron is less familiar than that of a cube, but the end result is simple to state.

Theorem 14.26. *A regular dodecahedron admits 120 symmetries, of which 60 are rotations. The full symmetry group is isomorphic to S_5 , and the rotation group is isomorphic to A_5 .*

14.6 Rubik's Cube

Mathematicians study groups substantially larger than these. Among the most famous is the group of configurations of *Rubik's cube*, a mechanical puzzle invented by the Hungarian architect Ernő Rubik in the

1970s. Mathematically, Rubik's cube consists of 27 small "cubies" arranged into a $3 \times 3 \times 3$ cube.*

The six faces of the cube are assigned labels, typically up (u) and down (d), left (l) and right (r), front (f) and back (b). Each label corresponds conventionally to a one-quarter turn of that face, counter-clockwise when looking down the axis onto the face. Abstractly, the "Rubik group" R is generated by the six elements u, d, l, r, t, b , each of order 4, subject to relations dictated by the geometry of face rotations. For example, u and d commute, as do rotations about other pairs of parallel faces. However, ur (rotate the right face, then the up face) and ru (up, then right) are different; a similar remark holds for any other pair of rotations about bordering faces.

Each element of R corresponds to a unique configuration of cubies, with the identity element corresponding to a "solved" cube. The group R is generated by six elements of order 4, and every configuration is reachable by at most 29 quarter-turns; the actual minimum number is currently unknown, but believed to be around 20. Despite these seemingly-small numbers, R has order

$$8! \times 3^7 \times 12! \times 2^{10} = 43,252,003,274,489,856,000 \approx 4.3 \times 10^{19}.$$

Starting from a solved cube, you can explore basic group-theoretic concepts by performing restricted sets of "moves".

Example 14.27. The element ru has order 105:

$$(ru)^{105} = \underbrace{ruru \dots ru}_{105 \text{ times}} = e,$$

but no smaller positive power is equal to e . Take care not to get distracted while experimenting with powers of ru unless you are able to solve the scrambled puzzle!

Example 14.28. Rubik's group contains small subgroups generated by "half-turns": u^2, d^2 , etc. For example, $\langle u^2 d^2, l^2 r^2, t^2 b^2 \rangle$, the group generated by pairs of half-turns on parallel faces, has order 8 and contains pleasing checkerboard patterns.

The group $\langle u^2, d^2, l^2, r^2, t^2, b^2 \rangle$ generated by all half-turns is larger, but difficult to get "lost" in. In each such configuration of the cube, colors from opposite faces will be mixed with each other, but nothing worse.

*The actual pieces of a physical cube interlock ingeniously, permitting each of six $3 \times 3 \times 1$ "layers" to pivot about the central axis perpendicular to the layer while holding the assemblage together during the rotation.

Exercises

Exercise 14.1. Prove that the group of symmetries of a non-square rectangle is isomorphic to the Klein 4-group $(\mathbf{Z}_2 \times \mathbf{Z}_2, +)$.

Exercise 14.2. Using the technique of Example 14.22, enumerate the cyclic subgroups of:

- (a) S_3 . (b) D_4 . (c) Q . (d) D_6 .

Exercise 14.3. (a) Enumerate the cyclic subgroups of order 4 in S_4 .

(b) How many cyclic subgroups of order 4 are there in S_5 ? In A_5 ?

(c) How many cyclic subgroups of order 4 are there in S_6 ? In A_6 ?

(d) How many cyclic subgroups of order 5 are there in S_5 ? In A_5 ?

(e) How many cyclic subgroups of order 6 are there in S_5 ?

Exercise 14.4. (a) Using the elements a and b as in the text, list the ten elements of the dihedral group D_5 , and enumerate its cyclic subgroups.

(b) Similarly, list the elements of the dihedral group D_6 , and enumerate its cyclic subgroups.

(c) The dihedral group D_6 contains precisely two distinct subgroups isomorphic to D_3 . Prove this assertion, and list the elements of each subgroup.

Exercise 14.5. Which group has more distinct cyclic subgroups: D_{17} (symmetries of a 17-gon) or D_{18} (symmetries of an 18-gon)? Explain.

Exercise 14.6. Let a, b in D_n be as in Section 14.2.

(a) Prove $a^{r_1}(a^{r_2}b^s) = a^{r_1+r_2}b^s$ and $(a^{r_1}b)(a^{r_2}b^s) = a^{r_1-r_2}b^{s+1}$.

(b) Show that, with exponents on a reduced mod n and exponents on b reduced mod 2, the formulas in part (a) can be combined into

$$(a^{r_1}b^{s_1})(a^{r_2}b^{s_2}) = a^{r_1+(-1)^{s_1}r_2}b^{s_1+s_2}.$$

Exercise 14.7. Prove or disprove:

(a) D_6 is isomorphic to A_4 . (b) D_6 is isomorphic to $S_3 \times (\mathbf{Z}_2, +)$.

Exercise 14.8. We have seen five groups of order 8: D_4 , Q ,

$$G_8 = (\mathbf{Z}_8, +), \quad G_{4,2} = (\mathbf{Z}_4 \times \mathbf{Z}_2, +), \quad G_{2,2,2} = (\mathbf{Z}_2 \times \mathbf{Z}_2 \times \mathbf{Z}_2, +).$$

Find the number of elements of each order in each group, and conclude that no two of these groups are isomorphic. (Every group of order 8 turns out to be isomorphic to one of these five.)

Exercise 14.9. In the group D_8 of symmetries of a regular octagon:

- (a) There is a cyclic subgroup of order 8.
- (b) There exist two subgroups isomorphic to D_4 .
- (c) There exist four subgroups isomorphic to $(\mathbf{Z}_2 \times \mathbf{Z}_2, +)$.
- (d) The groups found in (a)–(c) are the only such subgroups of D_8 .
- (e) There is no subgroup of D_8 isomorphic to $G_{2,2,2}$, to $G_{4,2}$, or to Q .
(See Exercise 14.8 for notation.)

Exercise 14.10. Let $G = GL(2, \mathbf{Z}_2) \subseteq \mathbf{Z}_2^{2 \times 2}$ be the set of invertible 2×2 matrices with entries in $\mathbf{Z}_2 = \{0, 1\}$, viewed as a group under matrix multiplication, see Exercise 11.33.

- (a) Prove algebraically that (G, \cdot) is isomorphic to S_3 : Find a matrix α in G of order 3 and a matrix β in G of order 2, and show that $\beta\alpha = \alpha^2\beta$.
- (b) Prove geometrically that (G, \cdot) is isomorphic to S_3 : Consider the “ \mathbf{Z}_2 -plane” $\mathbf{Z}_2 \times \mathbf{Z}_2$, whose elements are

$$\mathbf{e}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_a = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}_b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{e}_c = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Show that under left multiplication, (G, \cdot) faithfully permutes the set $\{\mathbf{e}_a, \mathbf{e}_b, \mathbf{e}_c\}$. Write the permutation corresponding to each element of G in cycle notation.

Exercise 14.11. In each part, $G = GL(2, \mathbf{R})$ is the multiplicative group of real invertible 2×2 matrices.

- (a) Exhibit a subgroup H_1 of G that is isomorphic to $(\mathbf{Z}_6, +)$.
- (b) Exhibit a subgroup H_2 of G that is isomorphic to S_3 .

Exercise 14.12. Let a be a positive real number, $a \neq 1$, and let P be the rectangular box with eight vertices $(\pm 1, \pm 1, \pm a)$.

- (a) Find a “familiar” group isomorphic to the rotation group of P , make a careful sketch, and describe an isomorphism in detail.
- (b) Describe the full symmetry group of P by determining the number of elements of each order and describing the geometric action of each type of element.

Exercise 14.13. Let $a < b < c$ be positive real numbers, and let P be the rectangular box with eight vertices $(\pm a, \pm b, \pm c)$.

- (a) Find a “familiar” group isomorphic to the rotation group of P , make a careful sketch, and describe an isomorphism in detail.
- (b) Describe the full symmetry group of P by determining the number of elements of each order and describing the geometric action of each type of element.

Exercise 14.14. Let f be a symmetry of the cube that maps each long diagonal to itself. Show f is either the identity map or the map $f(x, y, z) = (-x, -y, -z)$.

Hint: If v_1 and v_2 are endpoints of an edge, so are $f(v_1)$ and $f(v_2)$.

Exercise 14.15. On the left in Figure 14.2, the permutation $(0\ 1\ 2)$ is identified as rotation by $1/3$ of a turn about the face 012 , while $(0\ 3)(1\ 2)$ is a $1/2$ turn about the axis through the midpoints of the edges 03 and 12 . (All rotations are assumed counterclockwise, with the “sense” gauged by looking toward the center of the polyhedron from the edge or face being rotated.)

- (a) Give the vertex permutations corresponding to the indicated symmetries of the tetrahedron:
 - (i) Rotation by $2/3$ of a turn about the face 012 .
 - (ii) A $1/2$ turn about the axis through the midpoints of the edges 01 and 23 .
 - (iii) Reflection about the plane through the center and containing edge 23 .
- (b) Describe the symmetries giving rise to the permutations
 - (i) $(0\ 1)(2\ 3)$
 - (ii) $(0\ 3\ 2)$
 - (iii) $(0\ 1)$.

Exercise 14.16. Referring to the right half of Figure 14.2, and in a similar fashion to the preceding question:

- (a) Describe the axis permutations corresponding to the following symmetries of the cube.
 - (i) Rotation by a quarter turn about the top face.
 - (ii) Rotation by $1/3$ of a turn about axis 0. (Rotate counterclockwise looking down the axis from the label 0 toward the center.)
 - (iii) Rotation by $1/3$ of a turn about axis 1.
 - (iv) A half turn about the axis through the midpoint of edge 12.
 - (v) Reflection in the plane containing axes 0 and 1.
 - (vi) Reflection in the plane containing axes 0 and 2.
- (b) Describe and sketch the *rotation* giving rise to each of the following axis permutations.
 - (i) $(0\ 1)(2\ 3)$
 - (ii) $(0\ 3\ 2)$
 - (iii) $(0\ 1)$.

Exercise 14.17. Referring to the right half of Figure 14.2, let R_1 be a quarter-turn about \mathbf{e}_1 and R_2 a quarter turn about \mathbf{e}_2 .

- (a) Use geometry to find the axis permutations of R_1R_2 and R_2R_1 . (Remember to apply the *right-hand* mapping first.)
- (b) Find the axis permutations of R_1 and R_2 , and use cycle calculations to confirm your answer to part (a).

Chapter 15

Cosets

Definition 15.1. Let (G, \cdot) be a group, and let H be a subgroup. Each element a in G defines a *left coset* of H in G , namely the set

$$aH = \{ah : h \in H\} = \{x \text{ in } G : x = ah \text{ for some } h \text{ in } H\} \subseteq G.$$

Remark 15.2. As we will see presently, any two left cosets are in bijective correspondence, and are either disjoint or identical. The set of left cosets of H consequently defines a partition of G , and if G is finite, then any two left cosets contain the same number of elements.

Example 15.3. Let $(G, \cdot) = (\mathbf{Z}, +)$ be the additive group of integers, and let $H = 3\mathbf{Z}$, the set of integer multiples of 3. Listing elements from a few left cosets “close to” $H = 0 + H$ gives

$$\begin{aligned} -1 + H &= \{\dots, -7, -4, -1, \quad 2, \quad 5, \dots\}, \\ 0 + H &= \{\dots, -6, -3, \quad 0, \quad 3, \quad 6, \dots\}, \\ 1 + H &= \{\dots, -5, -2, \quad 1, \quad 4, \quad 7, \dots\}, \\ 2 + H &= \{\dots, -4, -1, \quad 2, \quad 5, \quad 8, \dots\}, \\ 3 + H &= \{\dots, -3, \quad 0, \quad 3, \quad 6, \quad 9, \dots\}, \\ 4 + H &= \{\dots, -2, \quad 1, \quad 4, \quad 7, \quad 10, \dots\}. \end{aligned}$$

Closer inspection reveals this list is redundant: $-1 + H = 2 + H$, $0 + H = 3 + H$, and $1 + H = 4 + H$. The reason is not difficult to discern. For example, the coset $1 + H$ is the set of integers one larger than a multiple of 3, namely, the set of integers leaving a remainder of 1 upon division by 3. However, every integer 4 greater than a multiple of three is also 1 greater than a multiple of 3, since $4 + 3k = 1 + 3(k + 1)$.

Indeed, the division algorithm allows us to write an arbitrary integer N uniquely in the form $N = 3q + r$ with $0 \leq r < 3$, proving $N \in r + H$. It follows there are exactly three left cosets of $H = 3\mathbf{Z}$ in \mathbf{Z} : $0 + H$, $1 + H$, and $2 + H$.

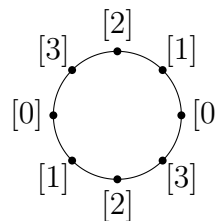
Example 15.4. There is nothing special about the modulus 3 in the preceding example. If $n > 1$ is an integer and we view $H = n\mathbf{Z}$ as a subgroup of $(\mathbf{Z}, +)$, then H has precisely n left cosets

$$r + H = r + n\mathbf{Z} = \{nq + r : q \in \mathbf{Z}\}, \quad 0 \leq r < n,$$

the set of integers leaving a remainder of r on division by n . In the discussion of congruence classes, we wrote $r + n\mathbf{Z} = [r]_n$.

Example 15.5. Let $G = \mathbf{Z}_8$, viewed as a group under addition mod 8, and let $H = \langle [4] \rangle = \{[0], [4]\}$. A bit of thought along the lines of the preceding examples shows H has four distinct left cosets in G :

$$\begin{aligned} [0] + H &= [4] + H = \{[0], [4]\}, \\ [1] + H &= [5] + H = \{[1], [5]\}, \\ [2] + H &= [6] + H = \{[2], [6]\}, \\ [3] + H &= [7] + H = \{[3], [7]\}. \end{aligned}$$



In the figure, the elements of \mathbf{Z}_8 are partitioned into four cosets and labeled accordingly, compare Corollary 15.7.

Cosets turn out to be a convenient tool for studying the structure of groups, particularly finite groups, largely because distinct left cosets of H partition G . To prove this, we first recast the definition of a left coset in terms of an equivalence relation.

Theorem 15.6. Let (G, \cdot) be a group and $H \subseteq G$ a subgroup. Define a relation R on G by aRb if and only if $a^{-1}b \in H$.

- (i) The relation R is an equivalence relation.
- (ii) The equivalence classes of R are precisely the left cosets of H in G .
- (iii) There exists a bijection between arbitrary left cosets aH and bH .

Proof. (i) The relation R is reflexive because $e = a^{-1}a \in H$ for all a in G .

Symmetry of R restates the reverse order law for a group operation: $(a^{-1}b)^{-1} = b^{-1}a$. Since the subgroup H is closed under inversion, aRb if and only if $a^{-1}b \in H$, if and only if $b^{-1}a = (a^{-1}b)^{-1} \in H$, if and only if bRa .

To prove R is transitive, suppose aRb and bRc , namely that $a^{-1}b$ and $b^{-1}c$ are elements of H . Since H is closed under the operation of G , $a^{-1}c = (a^{-1}b)(b^{-1}c) \in H$, which means aRc .

Since R is reflexive, symmetric, and transitive, R is an equivalence.

(ii) Let a and x be arbitrary elements of G . It suffices to prove $x \in aH$ if and only if aRx . But

$x \in aH$ if and only if there exists an h in H such that $x = ah$,
 if and only if there exists an h in H such that $a^{-1}x = h$,
 if and only if aRx .

(iii) If aH and bH are arbitrary left cosets of H , then multiplication on the left by ba^{-1} , a bijection from G to G , maps the left coset aH to $(ba^{-1})aH = b(a^{-1}a)H = bH$. The inverse map, multiplication by ab^{-1} , similarly maps bH to aH . These cosets are therefore in bijective correspondence. \square

Corollary 15.7. *If (G, \cdot) is a group and $H \subseteq G$ is a subgroup, then $aH = bH$ if and only if $b \in aH$.*

Remark 15.8. The decomposition of a finite group G into cosets of H may be reached by a finite computation: Step through the elements of G one by one, “multiplying” each element a by each element of H and collecting the set aH of “products”:

Example 15.9. Let $G = D_3 = \{e, a, a^2, b, ab, a^2b\}$ and $H = \{e, b\}$. To decompose G into left cosets of H , pick an element not in H , say a , and enumerate $aH = \{a, ab\}$. Corollary 15.7 guarantees $abH = aH$ as well. Now pick an element not in $H \cup aH$, say a^2 , and compute $a^2H = \{a^2, a^2b\}$. By Corollary 15.7, $a^2H = a^2bH$. In this example, $G = H \cup aH \cup a^2H$; we have “used up” all the elements of G . Generally, continue until every element of G is contained in some coset.

Lagrange’s Theorem and Applications

Theorem 15.6 has a simple, important consequence for finite groups: The order of a subgroup divides the order of the group.

Theorem 15.10 (Lagrange’s theorem). *If (G, \cdot) is a finite group con-*

taining n elements, $H \subseteq G$ is a subgroup containing m elements, and H has k distinct left cosets in G , then $n = km$. In particular, $m \mid n$.

H	a_1H	a_2H	\cdots	$a_{k-1}H$
-----	--------	--------	----------	------------

Proof. The distinct left cosets of H partition G by part (ii) of Theorem 15.6 (equivalence classes of a relation), and any two left cosets have the same number of elements by part (iii). If there are k distinct left cosets of H in G , then $n = km$. \square

Definition 15.11. Let H be a subgroup of (G, \cdot) . The number of distinct left cosets of H in G is called the *index* of H in G . If G is finite, the index of H is the order of G divided by the order of H .

Corollary 15.12. Let (G, \cdot) be a finite group of order n . If $a \in G$, then the order of a divides n .

Proof. Suppose $a \in G$, and let $H = \langle a \rangle$ be the cyclic subgroup generated by a . The order of a , a.k.a. the order of H , divides n by Lagrange's theorem. \square

Corollary 15.13. Every group (G, \cdot) of prime order p is cyclic, and every non-identity element is a generator.

Proof. Let $a \neq e$ be an element of G , and consider the cyclic subgroup $H = \langle a \rangle$ generated by a . Since $\{e, a\} \subseteq H$, the order of H is at least 2. By Lagrange's theorem, the order of H divides p , the (prime) order of G . The only divisor of p greater than 1 is p itself, so H must have order p , namely, must be equal to G . \square

Remark 15.14. Since two cyclic groups of the same order n are isomorphic, we have shown that for each prime p , there exists exactly one group of order p up to isomorphism.

Remark 15.15. A set containing 101 elements admits a large number of non-commutative binary operations possessing an identity element and inverses. Corollary 15.13 guarantees none of these operations is associative! Indeed, an associative, non-commutative binary operation with an identity element and inverses would define a non-Abelian group of order 101. But 101 is prime, so every group of order 101 is cyclic, hence Abelian.

Remark 15.16. Among groups of order less than 10, a group of order 2, 3, 4, 5, or 7 must be Abelian. This follows immediately for the prime orders 2, 3, 5, and 7. In a group G of order 4, either there exists an element of order 4, so G is cyclic (hence Abelian), or else Lagrange's theorem guarantees every non-identity element of G has order 2. In this event, G is Abelian by Exercise 7.20, and is easily checked to be isomorphic to the Klein 4-group $(\mathbf{Z}_2 \times \mathbf{Z}_2, +)$.

To exhibit a more interesting conclusion using the tools just developed, we'll find all groups of order 9 up to isomorphism. In particular, every group of order 9 is Abelian.

Proposition 15.17. *Let (G, \cdot) be a finite group of order 9. Then G is isomorphic to either \mathbf{Z}_9 or to $\mathbf{Z}_3 \times \mathbf{Z}_3$.*

Proof. The only divisors of 9 are 1, 3, and 9, so every non-identity element of G has order 3 or order 9. If G contains an element of order 9, then G is cyclic, hence isomorphic to \mathbf{Z}_9 .

Otherwise, every non-identity element of G has order 3. Pick $a \neq e$, and write $H = \langle a \rangle = \{e, a, a^2\}$. Next, pick an element b not in H . The left cosets $bH = \{b, ba, ba^2\}$ and $b^2H = \{b^2, b^2a, b^2a^2\}$ are mutually disjoint, for if $bH \cap b^2H \neq \emptyset$, then $b = b^{-1}b^2 \in H$ by Theorem 15.6. It follows that we have listed all nine elements: $G = H \cup bH \cup b^2H$.

We claim $ab = ba$. The element ab of G must appear among

$$e, a, a^2, \quad b, ba, ba^2, \quad b^2, b^2a, b^2a^2.$$

Since $b \notin \langle a \rangle$, the cancellation law implies ab cannot be any of the following: e, a, a^2, b , or b^2 . Moreover, since a, b , and ab have order 3,

$$a(a^2b^2)b = (a^3)(b^3) = e = (ab)^3 = ababab = a(baba)b,$$

so by cancellation $a^2b^2 = baba = b(ab)a$. Consequently, ab cannot be any of b^2a^2, ba^2 , or b^2a . For example, if $ab = ba^2$, then substitution would give

$$a^2b^2 = b(ba^2)a = b^2a^3 = b^2,$$

which would imply $a^2 = e$. The only remaining possibility is $ab = ba$.

The group G is therefore generated by a pair of commuting elements, hence is Abelian. As seen above, the elements of G are precisely the expressions $b^i a^j$ with $i, j = 0, 1, 2$, and the group operation is given by

$$(b^i a^j)(b^k a^\ell) = b^{(i+k) \bmod 3} a^{(j+\ell) \bmod 3}.$$

The bijection $\phi : \mathbf{Z}_3 \times \mathbf{Z}_3 \rightarrow G$ defined by $\phi(i, j) = b^i a^j$ is therefore a group isomorphism. \square

Example 15.18. (Proper subgroups of A_4) The alternating group A_4 contains three elements of order 2 and eight elements of order 3. These generate three cyclic subgroups of order 2 and four cyclic subgroups of order 3 (since σ and σ^{-1} generate the same subgroup).

If a subgroup $H \subseteq A_4$ contains two distinct elements of order 2, it must contain their product, and therefore must have order at least 4. This set of four elements is closed, hence is a non-cyclic subgroup.

We claim the seven cyclic subgroups and the non-cyclic subgroup of order 4 are the only proper, non-trivial subgroups of A_4 . To see why, consider what happens if $H \subseteq A_4$ contains an element τ of order 2 and an element σ of order 3. The cyclic subgroup $K = \langle \sigma \rangle$ contains three elements, the left coset τK contains three more elements, and $\sigma\tau \notin K \cup \tau K$ is a seventh element in H . Since the order of H is both a divisor of twelve *and* strictly larger than 6, H has order 12, namely is all of A_4 .

Similarly, if H contains non-commuting elements σ_1 and σ_2 of order 3, then $K = \langle \sigma_1 \rangle$, $\sigma_2 K$, and $\sigma_2^2 K$ collectively account for at least nine elements of H , and again by Lagrange's theorem $H = A_4$.

15.1 Normal Subgroups

Definition 15.19. If (G, \cdot) is a group, $H \subseteq G$ is a subgroup, and $a \in G$, we define the *right coset* Ha to be the set

$$Ha = \{ha : h \in H\} = \{x \text{ in } G : x = ha \text{ for some } h \text{ in } H\}.$$

Remark 15.20. Obvious modifications of prior arguments show that there is a relation on G whose equivalence classes are precisely the right cosets of H in G , any two right cosets are either disjoint or identical, and arbitrary right cosets are in bijective correspondence, see Exercise 15.10.

Definition 15.21. Let (G, \cdot) be a group and $H \subseteq G$ a subgroup. We say H is a *normal subgroup* of G , written $H \triangleleft G$, if $aH = Ha$ for every a in G . In other words, a subgroup is normal if *every* left coset is also a right coset.

Example 15.22. Let $(G, +)$ be an *Abelian* group and $H \subseteq G$ a subgroup. Every left coset of H is also a right coset. Indeed,

$$aH = \{ah : h \in H\} = \{ha : h \in H\} = Ha$$

because $ah = ha$ for all h in G .

Example 15.23. Let $G = D_3$ be the dihedral group of symmetries of an equilateral triangle, $H = \langle a \rangle$ the cyclic group generated by a one-third rotation, and $K = \langle b \rangle$ the cyclic subgroup generated by reflection across an axis.

Recall the elements of G are $\{e, a, a^2, b, ab, a^2b\}$, and that $ba = a^2b$. There are two distinct left cosets of H in D_3 :

$$\begin{aligned} eH &= aH = a^2H = \{e, a, a^2\}, \\ bH &= baH = ba^2H = \{b, ba, ba^2\} = \{b, a^2b, ab\}. \end{aligned}$$

Similarly, H has two distinct right cosets, and $bH = Hb$.

Note that this equality holds only “at the level of sets”, not “at the level of elements”, since $ab \neq ba$ and $a^2b \neq ba^2$.

Considering the subgroup K , we encounter new behavior:

$$\begin{aligned} eK &= bK = \{e, b\} & Ke &= Kb = \{e, b\}, \\ aK &= abK = \{a, ab\} & Ka &= Kba = \{a, ba\} = \{a, a^2b\}, \\ a^2K &= a^2bK = \{a^2, a^2b\} & Ka^2 &= Kba^2 = \{a^2, ba^2\} = \{a^2, ab\}. \end{aligned}$$

The right cosets of K in D_3 are not the left cosets. Particularly, a left coset and a right coset can intersect without being identical.

In summary, $H \triangleleft D_3$, but K is not a normal subgroup of D_3 .

Products of Sets

Definition 15.24. Let A and B be arbitrary non-empty subsets of a group (G, \cdot) . Their *product* is the set

$$\begin{aligned} AB &= \{ab : a \in A, b \in B\} \\ &= \{x \text{ in } G : x = ab \text{ for some } a \text{ in } A \text{ and } b \text{ in } B\}, \end{aligned}$$

namely, the union of the sets aB as a ranges over A , or the union of the sets Ab as b ranges over B .

Remark 15.25. Left cosets may be viewed as the special cases in which A is a singleton and B is a subgroup: $aH = \{a\}H$.

In this interpretation, normal subgroups are exactly the subgroups commuting with singletons: $aH = Ha$.

Lemma 15.26. Let (G, \cdot) be a group, $H \subseteq G$ a subgroup.

- (i) $HH = H$.

(ii) If A , B , and C are subsets of (G, \cdot) , then $(AB)C = A(BC)$.

(iii) $H \triangleleft G$ if and only if $a^{-1}Ha = H$ for all a in G .

Proof. (i) Clearly $H \subseteq HH$, since $e \in H$ and $h = he$ for all h in H . Conversely, if h_1 and h_2 are elements of H , then $h_1h_2 \in H$ because H is a subgroup of G (hence closed under the operation of G), so $HH \subseteq H$.

(ii) This assertion merely reformulates associativity of the operation of G . In detail, $x \in (AB)C$ if and only if there exist elements a in A , b in B , and c in C such that $x = (ab)c$. By associativity, this condition holds if and only if $x = a(bc)$ for some a , b , and c , if and only if $x \in A(BC)$.

(iii) Fix a in G . Then $aH = Ha$ if and only if $(aH)a^{-1} = (Ha)a^{-1}$. By (ii), $(Ha)a^{-1} = H(aa^{-1}) = He = H$. \square

Theorem 15.27. Let (G, \cdot) be a group, and $H \triangleleft G$ a normal subgroup. For all a and b in G , $(aH)(bH) = (ab)H$.

Proof. This follows almost immediately from the preceding observations:

$$\begin{aligned}
 (aH)(bH) &= a(H(bH)) && \text{Lemma 15.26 (ii)} \\
 &= a(Hb)H && \text{Lemma 15.26 (ii)} \\
 &= a(bH)H && H \triangleleft G \\
 &= a(b(HH)) && \text{Lemma 15.26 (ii)} \\
 &= a(bH) && \text{Lemma 15.26 (i)} \\
 &= (ab)H && \text{Lemma 15.26 (ii)}
 \end{aligned}$$

for all a and b . \square

Remark 15.28. This technical result is fundamental: It asserts that left cosets of a *normal* subgroup can be “multiplied” as if they were themselves elements of a group. This group, the “quotient group” G/H , is studied in the next chapter.

If H is a non-normal subgroup, coset multiplication is not well-defined, in that the coset of the product depends on the choice of representative elements.

Exercises

Exercise 15.1. Let $G = (\mathbf{Z}_{15}, +)$ and $H = \langle 5 \rangle \subseteq \mathbf{Z}_{15}$. Partition \mathbf{Z}_{15} into left cosets of H .

Exercise 15.2. Let $G = (\mathbf{Z}_4 \times \mathbf{Z}_3, +)$ and $H = \langle (2, 1) \rangle$. List the elements of $\mathbf{Z}_4 \times \mathbf{Z}_3$ in a 4×3 array, list the elements of H , and partition G into left cosets of H .

Exercise 15.3. For each group G and subgroup H , (i) partition G into left cosets of H , (ii) partition G into right cosets of H , and (iii) determine whether $H \triangleleft G$.

- (a) $G = D_3$ the dihedral group, $H = \langle ab \rangle = \{e, ab\}$.
- (b) $G = D_6$, $H = \langle a^2 \rangle = \{e, a^2, a^4\}$.
- (c) $G = D_6$, $H = \{e, a^3, b, a^3b\}$.
- (d) $G = Q$ the quaternion group, $H = \langle -1 \rangle$.
- (e) $G = Q$, $H = \langle i \rangle$.
- (f) $G = A_4$ the alternating group on four letters, $H = \langle (1\ 2\ 3) \rangle$.
- (g) $G = A_4$, H the non-cyclic subgroup of order 4.

Exercise 15.4. Consider $G = (\mathbf{Z}_{10}, \cdot)$, the set of residue classes mod 10 under multiplication, and let $H = \{[0], [2], [4], [6], [8]\} \subseteq \mathbf{Z}_{10}$.

- (a) Show H is closed under \cdot , and that \cdot has an identity element in H . Which elements of H have inverses?
Suggestion: Write out the Cayley table for (H, \cdot) .
- (b) Explain why existence of an identity element in (H, \cdot) does not contradict Proposition 7.31 (i). (“The identity element in H is the same as the identity element in G .”)
- (c) Show H^\times , the set of invertible elements of H , is a group under multiplication mod 10.
- (d) Explain why part (c) does not contradict Lagrange’s theorem.

Exercise 15.5. Let (G, \cdot) be an arbitrary group, and let $H = \{e\}$ be the trivial subgroup. What are the left cosets of H in G ? The right cosets? Is H normal in (G, \cdot) ?

Exercise 15.6. Let (G, \cdot) be a group, and $H \subseteq G$ a subgroup of index 2. Prove $H \triangleleft G$.

Hint: What are the left cosets of H ? The right cosets?

Exercise 15.7. Use the result of the preceding exercise to prove:

- (a) $A_n \triangleleft S_n$ for all $n > 2$.
- (b) The dihedral group D_n contains a normal subgroup isomorphic to \mathbf{Z}_n .
- (c) The dihedral group D_6 contains a normal subgroup isomorphic to S_3 .
Hint: Consider $\langle a^2, b \rangle$.

Exercise 15.8. (a) Let (G, \cdot) be a finite group, and suppose G contains a *unique* subgroup H of some order. Prove $H \triangleleft G$.

Hint: Show that for each a in G , $a^{-1}Ha \subseteq G$ is a subgroup of G having the same order as H .

- (b) Use part (a) to give alternative proofs of parts (b), (d), and (g) of Exercise 15.3.

Exercise 15.9. Let (G, \cdot) be the multiplicative group $GL(2, \mathbf{R})$ of invertible 2×2 real matrices.

- (a) Prove that $H = \{cI_2 : c \in \mathbf{R}\}$, the group of scalar matrices, is normal in (G, \cdot) .
- (b) Determine whether the group of invertible diagonal matrices is normal in (G, \cdot) .
- (c) Prove that $SL(2, \mathbf{R})$, the group of matrices of determinant 1, is normal in (G, \cdot) .
Hints: Use the criterion in Lemma 15.26 (iii). By Exercise 11.10, $\det(AB) = \det A \det B$ for all A, B in $\mathbf{R}^{2 \times 2}$.

Exercise 15.10. Let (G, \cdot) be a group, H a subgroup. Define a relation R by aRb if and only if $ba^{-1} \in H$.

- (a) Prove R is an equivalence relation.
- (b) Prove the equivalence classes of R are precisely the right cosets of H in G .
- (c) Prove that any two right cosets are in bijective correspondence.
- (d) Show that all cosets (left or right) are in bijective correspondence.

Chapter 16

Homomorphisms

16.1 Definition and Properties

Definition 16.1. Let (G_1, \cdot) and $(G_2, *)$ be arbitrary groups. A mapping $\phi : G_1 \rightarrow G_2$ is a *homomorphism* if

$$\phi(a \cdot b) = \phi(a) * \phi(b) \quad \text{for all } a \text{ and } b \text{ in } G_1.$$

If there exists a *surjective* homomorphism $\phi : G_1 \rightarrow G_2$, we say G_2 is a *homomorphic image* of G_1 .

That is, a homomorphism is “just like an isomorphism, except possibly not bijective”. The following properties mirror parts of Theorem 12.2, and are proven in exactly the same way.

Proposition 16.2. Let $\phi : G_1 \rightarrow G_2$ be a group homomorphism, and let e_i in G_i be the respective identity elements.

- (i) $\phi(e_1) = e_2$.
- (ii) For all a in G_1 , $\phi(a^{-1}) = \phi(a)^{-1}$.
- (iii) For all a in G_1 and k in \mathbf{Z} , $\phi(a^k) = \phi(a)^k$.

The following sets “measure” the extent to which a homomorphism ϕ fails to be injective and/or surjective.

Definition 16.3. Let $\phi : G_1 \rightarrow G_2$ be a homomorphism. The *kernel* of ϕ is the preimage of the identity element e_2 of G_2 :

$$\ker \phi = \phi^{-1}(\{e_2\}) = \{a \text{ in } G_1 : \phi(a) = e_2\} \subseteq G_1.$$

The *image* of ϕ is the image of the mapping in the usual sense:

$$\text{img } \phi = \phi(G_1) = \{b \text{ in } G_2 : b = \phi(a) \text{ for some } a \text{ in } G_1\} \subseteq G_2.$$

Example 16.4. Let (G, \cdot) be a group with identity element e , and $H = \{\bullet\}$ a trivial (i.e., one-element) group. The mappings $\phi : G \rightarrow H$ and $\psi : H \rightarrow G$ defined by $\phi(a) = \bullet$ for all a in G , $\psi(\bullet) = e$ are “trivial” homomorphisms.

Example 16.5. Let (G, \cdot) be an arbitrary group. If $a \in G$, the mapping $\phi_a : (\mathbf{Z}, +) \rightarrow (G, \cdot)$ defined by $\phi(k) = a^k$ is a homomorphism by the law of exponents:

$$\phi(k + \ell) = a^{k+\ell} = a^k \cdot a^\ell = \phi(k) \cdot \phi(\ell) \quad \text{for all } k, \ell \text{ in } \mathbf{Z}.$$

If a has order m , $\ker \phi_a = m\mathbf{Z}$; if a has infinite order, $\ker \phi_a = \{0\}$. In either case, $\text{img } \phi_a = \langle a \rangle$, the cyclic subgroup generated by a .

Example 16.6. Let (G, \cdot) be an *Abelian* group. For each integer k , there is a homomorphism $\phi_k : G \rightarrow G$ defined by $\phi_k(x) = x^k$ for all x in G . Our ability to rearrange operands (Exercise 7.21) guarantees the morphism condition:

$$\phi_k(a \cdot b) = (a \cdot b)^k = a^k \cdot b^k = \phi_k(a) \cdot \phi_k(b).$$

In general, $(a \cdot b)^k \neq a^k \cdot b^k$ if a and b are non-commuting elements, so commutativity of (G, \cdot) is crucial.

Example 16.7. Let $(G, \cdot) = (\mathbf{Z}_n, +)$ in the preceding example. The mapping $\phi_2 : (\mathbf{Z}_8, +) \rightarrow (\mathbf{Z}_8, +)$ defined by $\phi_2(x) = x + x = 2x$ is a homomorphism. Tabulating values,

$$\begin{array}{cccccccc} x = & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2x = & 0 & 2 & 4 & 6 & 0 & 2 & 4 & 6 \end{array}$$

from which we read off $\ker \phi_2 = \{0, 4\}$ and $\text{img } \phi_2 = \{0, 2, 4, 6\}$.

For each $k = 0, 1, 2, \dots, 7$, there is a homomorphism $\phi_k : \mathbf{Z}_8 \rightarrow \mathbf{Z}_8$ satisfying $\phi_k(x) = kx$. The image is the cyclic subgroup generated by $d = \gcd(k, 8)$, and the kernel is the cyclic subgroup generated by $8/d$. In particular, ϕ_k is an isomorphism if and only if $\gcd(k, 8) = 1$.

Lemma 16.8. Let $\phi : (G_1, \cdot) \rightarrow (G_2, *)$ be a homomorphism. For all a and b in G_1 , $\phi(a) = \phi(b)$ if and only if $a^{-1} \cdot b \in \ker \phi$.

Proof. Let a and b be arbitrary elements of G_1 . By Proposition 16.2 and the morphism condition,

$$\phi(a)^{-1} * \phi(b) = \phi(a^{-1}) * \phi(b) = \phi(a^{-1} \cdot b).$$

Thus $a^{-1} \cdot b \in \ker \phi$ if and only if $\phi(a)^{-1} * \phi(b) = \phi(a^{-1} \cdot b) = e_2$, if and only if $\phi(a) = \phi(b)$. \square

Proposition 16.9. *Let $\phi : (G_1, \cdot) \rightarrow (G_2, *)$ be a homomorphism.*

- (i) $\ker \phi$ is a normal subgroup of G_1 .
- (ii) ϕ is injective if and only if $\ker \phi = \{e_1\}$.
- (iii) $\text{img } \phi$ is a subgroup of G_2 .

Proof. (i) Since $e_2 = \phi(e_1)$, both $\ker \phi$ and $\text{img } \phi$ are non-empty.

If a and b are arbitrary elements of $\ker \phi$, then $\phi(a) = e_2 = \phi(b)$. By Lemma 16.8, $a^{-1} \cdot b \in \ker \phi$. By Theorem 7.32, $\ker \phi$ is a subgroup of G_1 .

By Lemma 15.26 (iii), to prove the kernel is a *normal* subgroup it suffices to show $g^{-1}ag \in \ker \phi$ for all a in $\ker \phi$ and all g in G_1 . But

$$\phi(g^{-1}ag) = \phi(g^{-1})\phi(a)\phi(g) = \phi(g^{-1})e_2\phi(g) = \phi(g)^{-1}\phi(g) = e_2,$$

which means $g^{-1}ag$ is in $\ker \phi$.

(ii) If a is an arbitrary element of $\ker \phi$, then $\phi(a) = e_2 = \phi(e_1)$. If ϕ is injective, we have $a = e_1$; that is, $\ker \phi = \{e_1\}$.

Conversely, suppose $\ker \phi = \{e_1\}$. If $\phi(a) = \phi(b)$, then $a^{-1}b \in \ker \phi$ by Lemma 16.8. This implies $a^{-1}b = e_1$, i.e., $a = b$, so ϕ is injective.

(iii) If a_2 and b_2 are elements of $\text{img } \phi$, there exist elements a_1 and b_1 in G_1 such that $\phi(a_1) = a_2$, i.e., $\phi(a_1^{-1}) = a_2^{-1}$, and $\phi(b_1) = b_2$. By the morphism condition,

$$a_2^{-1} * b_2 = \phi(a_1^{-1}) * \phi(b_1) = \phi(a_1^{-1} \cdot b_1),$$

so $a_2^{-1} * b_2 \in \text{img } \phi$. By Theorem 7.32, $\text{Im } \phi$ is a subgroup of G_2 . \square

Remark 16.10. The image of a homomorphism need not be a normal subgroup of G_2 . Let G_2 be an arbitrary group having a non-normal subgroup G_1 . The inclusion map $\phi : G_1 \rightarrow G_2$ is obviously a homomorphism, but by construction the image is not normal in G_2 .

Example 16.11. The parity mapping $\text{sgn} : S_n \rightarrow \{\pm 1\}$ from the group of permutations on n letters to the multiplicative group $\{\pm 1\}$ is a homomorphism: $\text{sgn}(fg) = \text{sgn}(f) \text{sgn}(g)$ for all f and g in S_n .

The kernel is the set $\text{sgn}^{-1}(1)$ of permutations with sgn equal to 1, namely the alternating group A_n of even permutations. By Proposition 16.9, the alternating group A_n is a normal subgroup of the symmetric group S_n .

Example 16.12. Let $(G, \cdot) = GL(2, \mathbf{R})$ be the group of invertible 2×2 real matrices under matrix multiplication. The determinant mapping $\det : G \rightarrow (\mathbf{R}^\times, \cdot)$ is a homomorphism by Exercise 11.10.

The kernel $\det^{-1}(1)$ is $SL(2, \mathbf{R})$, the group of 2×2 real matrices of determinant 1. Consequently, $SL(2, \mathbf{R}) \triangleleft GL(2, \mathbf{R})$.

Lemma 16.13, the foundational technical result of this chapter, says a homomorphism ϕ is constant on left cosets of $K = \ker \phi$, but not on any larger sets. In Figure 16.1, left cosets of K are vertical lines in G_1 , and the mapping ϕ is vertical projection.

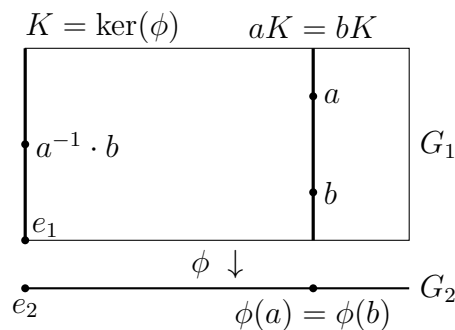


Figure 16.1: Homomorphisms, level sets, and cosets.

Lemma 16.13. Let $\phi : (G_1, \cdot) \rightarrow (G_2, *)$ be a homomorphism, and write $K = \ker \phi$. The following are equivalent for all a and b in G_1 .

- (a) $\phi(a) = \phi(b)$.
- (b) $a^{-1} \cdot b \in \ker \phi$.
- (c) $aK = bK$ as left cosets in G_1 .

Proof. (i) Statements (a) and (b) are equivalent by Lemma 16.8. Statements (b) and (c) are equivalent by Theorem 15.6 (ii). \square

16.2 Homomorphisms and Cyclic Groups

Let $\phi : (G, \cdot) \rightarrow (G', *)$ be a homomorphism. If $a \in G$, and we set $\phi(a) = a'$, then $\phi(a^k) = (a')^k$ by Proposition 16.2 (iii).

Lemma 16.14. *Let $\phi : (G, \cdot) \rightarrow (G', *)$ be a homomorphism, a in G , and $a' = \phi(a)$ in G' . If a has finite order n , then a' has finite order, and the order of a' divides n .*

Proof. Since $a^n = e$, Proposition 16.2 (iii) implies $(a')^n = e'$. Thus a' has finite order, say m .

The division algorithm allows us to write $n = mq + r$ uniquely with $0 \leq r < m$. Since $r = n - qm$, the law of exponents gives

$$(a')^r = (a')^n ((a')^m)^{-q} = e'.$$

But m is the smallest positive exponent k such that $(a')^k = e'$, so $r = 0$. That is, $m \mid n$. \square

Proposition 16.15. *Let (G, \cdot) be a cyclic group and $\phi : (G, \cdot) \rightarrow (G', *)$ a homomorphism. The image $G' = \phi(G)$ is a cyclic subgroup of $(G', *)$, and if G is finite, then the order of G' divides the order of G .*

Proof. Let a be a generator of G , and let $a' = \phi(a)$. If y in G' is arbitrary, there exists an x in G such that $\phi(x) = y$, so there exists an integer k such that $x = a^k$ (G is cyclic and a is a generator). Proposition 16.2 (iii) implies

$$y = \phi(x) = \phi(a^k) = \phi(a)^k = (a')^k.$$

That is, an arbitrary element of G' is a power of a' , i.e., $(G', *)$ is cyclic, and in fact generated by $a' = \phi(a)$. The assertion about orders follows immediately from Lemma 16.14. \square

Proposition 16.16. *Let $(G', *)$ be an arbitrary group, a' in G' of finite order m , and $n \geq 2$ an integer. If $m \mid n$, then there exists a homomorphism $\phi : (\mathbf{Z}_n, +) \rightarrow (G', *)$ with $\phi([1]) = a'$. Moreover, ϕ is injective if and only if $m = n$.*

Proof. Suppose $m \mid n$. Consider the homomorphism $\Phi : \mathbf{Z} \rightarrow G'$ defined by $\Phi(k) = (a')^k$. If $k_1 \equiv k_2 \pmod{n}$, then $k_1 \equiv k_2 \pmod{m}$ because $m \mid n$, so $\Phi(k_1) = \Phi(k_2)$. This means Φ is constant on residue classes mod n , so there is an induced homomorphism $\phi : (\mathbf{Z}_n, +) \rightarrow (G', *)$ satisfying $\phi([k]_n) = \Phi(k) = (a')^k$.

In particular, $\phi([1]) = a'$. Since $\ker \phi = \langle [m] \rangle \subseteq (\mathbf{Z}_n, +)$, ϕ is injective if and only if $\langle [m] \rangle = \langle [0] \rangle$, if and only if $m = n$. \square

The preceding two propositions may be summarized in the following useful form.

Theorem 16.17. *Let (G, \cdot) be a cyclic group of order $n \geq 2$, a in G a generator, $(G', *)$ an arbitrary group, and a' in G' of finite order m . There exists a homomorphism $\phi : (G, \cdot) \rightarrow (G', *)$ with $\phi(a) = a'$ if and only if $m \mid n$, and ϕ is injective if and only if $m = n$.*

16.3 Quotient Groups

In Theorem 15.27, we saw that left cosets of a normal subgroup $H \triangleleft G$ can be multiplied in a well-defined way.

Definition 16.18. Let (G, \cdot) be a group, $H \triangleleft G$. Denote by G/H the set of left cosets of H in G , and define an associative binary operation $*$ on G/H by

$$(aH) * (bH) = (ab)H.$$

The left coset $H = eH$ is the identity element, and the inverse of an arbitrary left coset aH is the left coset $(a^{-1})H$. The group $(G/H, *)$ is the *quotient group* $G \bmod H$.

Example 16.19. Let $G = (\mathbf{Z}_6, +)$ and $H = \langle 3 \rangle = \{0, 3\}$. Since G is Abelian, every subgroup is normal, so there is a quotient group G/H having three elements, the distinct left cosets of H in G . The Cayley table for G may be written so as to highlight the structure of the quotient by grouping elements of G lying in the same cosets of H , Table 16.1.

Well-definedness of the binary operation on cosets manifests itself visually: Each “large box” (comprising four “small boxes”) contains elements from a single coset. If we label the elements of G/H with the “nicknames”, $[0] = 0 + H = 3 + H$, $[1] = 1 + H = 4 + H$, and $[2] = 2 + H = 5 + H$, Table 16.1 may be written

$*$	$[0]$	$[1]$	$[2]$
$[0]$	$[0]$	$[1]$	$[2]$
$[1]$	$[1]$	$[2]$	$[0]$
$[2]$	$[2]$	$[0]$	$[1]$

In other words, the quotient group $\mathbf{Z}_6 / \langle 3 \rangle$ is isomorphic to \mathbf{Z}_3 , the cyclic group of order 3. The isomorphism maps the left coset of a generator, $1 + H$, to a generator $[1]$.

\cdot	$0 + H$	$3 + H$	$1 + H$	$4 + H$	$2 + H$	$5 + H$
$0 + H$	$0 + H$	$3 + H$	$1 + H$	$4 + H$	$2 + H$	$5 + H$
$3 + H$	$3 + H$	$0 + H$	$4 + H$	$1 + H$	$5 + H$	$2 + H$
$1 + H$	$1 + H$	$4 + H$	$2 + H$	$5 + H$	$3 + H$	$0 + H$
$4 + H$	$4 + H$	$1 + H$	$5 + H$	$2 + H$	$0 + H$	$3 + H$
$2 + H$	$2 + H$	$5 + H$	$3 + H$	$0 + H$	$4 + H$	$1 + H$
$5 + H$	$5 + H$	$2 + H$	$0 + H$	$3 + H$	$1 + H$	$4 + H$

Table 16.1: The Cayley table for $\mathbf{Z}_6/\langle 3 \rangle$.

A surjective homomorphism $\phi : \mathbf{Z}_6 \rightarrow \mathbf{Z}_3$ defined by $\phi(a) = a + H$ lurks in this example. Since H (viewed as a left coset) is the identity element of G/H , the kernel of ϕ is H (viewed as a subgroup of G).

A surprising amount of information about this example could have been deduced on general grounds. The subgroup $H = \langle 3 \rangle$ has two elements; since G has six elements, there are $6/2 = 3$ distinct left cosets. The group \mathbf{Z}_6 is cyclic, so the homomorphic image $\mathbf{Z}_6/\langle 3 \rangle$ is also cyclic by Theorem 16.17.

Example 16.20. Let $Q = \{\pm 1, \pm i, \pm j, \pm k\}$ be the quaternion group, see Exercise 11.37, and $H = \{\pm 1\}$. While G itself is non-Abelian, every element of H commutes with every element of G , so $H \triangleleft G$.

The distinct left cosets of H in Q are H itself, $iH = -iH = \{\pm i\}$, $jH = -jH = \{\pm j\}$, and $kH = -kH = \{\pm k\}$. As in the preceding example, the Cayley table of Q can be organized to highlight the structure of the quotient group $Q/\{\pm 1\}$, Table 16.2.

Again, each “large box” contains only elements from a single coset. If we call the elements of $Q/\{\pm 1\}$ by the “nicknames” $[1] = H = -H$, $[i] = iH = -iH$, $[j] = jH = -jH$, and $[k] = kH = -kH$, Table 16.2

\cdot	$1H$	$-1H$	iH	$-iH$	jH	$-jH$	kH	$-kH$
$1H$	$1H$	$-1H$	iH	$-iH$	jH	$-jH$	kH	$-kH$
$-1H$	$-1H$	$1H$	$-iH$	iH	$-jH$	jH	$-kH$	kH
iH	iH	$-iH$	$-1H$	$1H$	kH	$-kH$	$-jH$	jH
$-iH$	$-iH$	iH	$1H$	$-1H$	$-kH$	kH	jH	$-jH$
jH	jH	$-jH$	$-kH$	kH	$-1H$	$1H$	iH	$-iH$
$-jH$	$-jH$	jH	kH	$-kH$	$1H$	$-1H$	$-iH$	iH
kH	kH	$-kH$	jH	$-jH$	$-iH$	iH	$-1H$	$1H$
$-kH$	$-kH$	kH	$-jH$	jH	iH	$-iH$	$1H$	$-1H$

Table 16.2: The Cayley table for $Q/\{\pm 1\}$.

may be written

$*$	$[1]$	$[i]$	$[j]$	$[k]$
$[1]$	$[1]$	$[i]$	$[j]$	$[k]$
$[i]$	$[i]$	$[1]$	$[k]$	$[j]$
$[j]$	$[j]$	$[k]$	$[1]$	$[i]$
$[k]$	$[k]$	$[j]$	$[i]$	$[1]$

This is the table for the Klein 4-group $(\mathbf{Z}_2 \times \mathbf{Z}_2, +)$.

Note carefully that while Q is non-Abelian, both $\{\pm 1\}$ and $Q/\{\pm 1\}$ are Abelian.

Example 16.21. Let $(G, \cdot) = S_n$, the symmetric group on n letters, and $H = A_n$ the alternating group. There are two left cosets, H and

its complement H^c . The quotient group S_n/A_n is isomorphic to \mathbf{Z}_2 , the “standard” cyclic group of two elements.

16.4 The Homomorphism Theorem

Let (G, \cdot) be a group, $H \triangleleft G$ a normal subgroup, and $G' = G/H$ the quotient group. There is a quotient map $\phi : G \rightarrow G'$ sending each element a of G to its left coset aH in G' . This map is a homomorphism (by the definition of coset multiplication) and obviously surjective. Briefly, a normal subgroup, or equivalently a quotient group, gives rise to a surjective group homomorphism.

This chain of argument can be run backward as well. The precise formulation is the *Homomorphism Theorem*:

Theorem 16.22. *If $\phi : G \rightarrow G'$ is a group homomorphism with kernel K , the image $\phi(G) \subseteq G'$ is isomorphic to the quotient G/K .*

Proof. There is no loss of generality in assuming ϕ is surjective, so $G' = \phi(G)$.

By Lemma 16.13, ϕ is constant on left cosets of $K = \ker \phi$. Consequently, ϕ induces a mapping $\bar{\phi} : G/K \rightarrow G'$ by

$$\bar{\phi}(aK) = \phi(a) \quad \text{for all } a \text{ in } G.$$

It suffices to prove $\bar{\phi}$ is an isomorphism of groups.

(Morphism condition). For all a and b in G ,

$$\begin{aligned} \bar{\phi}((aK)(bK)) &= \bar{\phi}((ab)K) && \text{Definition of coset product} \\ &= \phi(ab) && \text{Definition of } \bar{\phi} \\ &= \phi(a)\phi(b) && \phi \text{ is a homomorphism} \\ &= \bar{\phi}(aK)\bar{\phi}(bK) && \text{Definition of } \bar{\phi}. \end{aligned}$$

(Injectivity). By Lemma 16.13, $aK \neq bK$ implies $\phi(a) \neq \phi(b)$, which means $\bar{\phi}(aK) \neq \bar{\phi}(bK)$.

(Surjectivity). If $a' \in G'$, there exists an a in G with $\phi(a) = a'$ because ϕ is surjective, so $\bar{\phi}(aK) = a'$. \square

Remark 16.23. As a practical matter, normal subgroups and surjective homomorphisms are in natural bijective correspondence. To find all homomorphic images of a group (G, \cdot) up to isomorphism, we need only enumerate the normal subgroups $H \triangleleft G$ and compute the quotient groups G/H .

Exercises

Exercise 16.1. Describe all homomorphisms $\phi : G_1 \rightarrow G_2$ between the indicated groups, and give the kernel and image of each.

- (a) $G_1 = (\mathbf{Z}_8, +)$, $G_2 = (\mathbf{Z}_4, +)$. (b) $G_1 = (\mathbf{Z}_4, +)$, $G_2 = (\mathbf{Z}_8, +)$.

Exercise 16.2. Find all homomorphisms $\phi : \mathbf{Z}_6 \rightarrow \mathbf{Z}_8$, and for each, find the kernel and image. Are any of these homomorphisms injective?

Exercise 16.3. In each part, let $G = (\mathbf{Z}_4 \times \mathbf{Z}_2, +)$. For each subgroup H , partition G into left cosets, write out the Cayley table for G/H , and determine the isomorphism class of G/H .

- (a) $H = \langle (1, 1) \rangle$. (b) $H = \langle (2, 0) \rangle$. (c) $H = \langle (2, 1) \rangle$.

Exercise 16.4. Let (G, \cdot) be a *finite* group. Use the Homomorphism Theorem to prove that if $\phi : G \rightarrow G'$ is a *surjective* homomorphism, then the order of G is the order of $\ker \phi$ multiplied by the order of G' .

Exercise 16.5. Recall that if $z = a + bi \in \mathbf{C}$, the norm of z is defined to be $|z| = \sqrt{z\bar{z}} = \sqrt{a^2 + b^2}$. Prove $|\cdot| : (\mathbf{C}^\times, \cdot) \rightarrow (\mathbf{R}^+, \cdot)$ is a surjective homomorphism, and find the kernel. Suggestion: Review Section 2.2.

Exercise 16.6. Let $(\mathbf{R}, +)$ be the additive group of real numbers and $(U(1), \cdot)$ the multiplicative group of complex numbers of length 1.

- (a) Show $\phi(t) = e^{it}$ is a surjective homomorphism from $(\mathbf{R}, +)$ to $U(1)$.
 (b) Prove $\mathbf{R}/(2\pi\mathbf{Z})$ is isomorphic to $U(1)$.

Exercise 16.7. Let $G = (\mathbf{R}^2, +)$ be the additive group of vectors in the plane. Fix real numbers a and b , not both zero, and let

$$H = \{(x, y) \text{ in } \mathbf{R}^2 : ax + by = 0\} \subseteq \mathbf{R}^2.$$

- (a) Show H is a subgroup of G , and describe H geometrically.
 (b) Show every coset of H in G has the form $\{(x, y) : ax + by = c\}$ for some real c . Describe these sets geometrically.
 (c) Since G is Abelian, $H \triangleleft G$. Show G/H is isomorphic to the additive group of real numbers.
 Hint: Map the coset in part (b) to the real number c .

Exercise 16.8. Find all homomorphic images of the following groups:

- (a) S_3 . (b) D_4 . (c) Q . (d) A_4 .

Exercise 16.9. Let a and b denote the usual rotation and reflection in D_6 . For each normal subgroup H given, partition D_6 into left cosets of H and determine the isomorphism class of the quotient D_6/H .

- (a) $H = \langle a^2 \rangle = \{e, a^2, a^4\}$. (b) $H = \langle a^3 \rangle = \{e, a^3\}$.

Exercise 16.10. Let (G, \cdot) be a group. The *center* of G is the set

$$Z(G) = \{g \text{ in } G : ag = ga \text{ for all } a \text{ in } G\}.$$

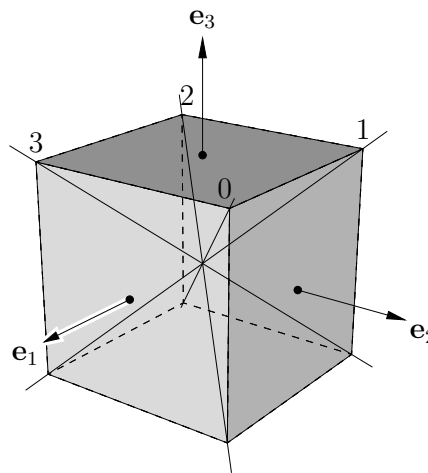
- (a) Prove $Z(G)$ is an Abelian subgroup of G .
 (b) If $H \subseteq Z(G)$ is a subgroup, prove $H \triangleleft G$.

Exercise 16.11. Let $\phi : G \rightarrow G'$ be a surjective homomorphism. If G is Abelian, prove G' is Abelian.

Exercise 16.12. Let $H_1 \triangleleft G_1$ and $H_2 \triangleleft G_2$.

- (a) Prove $H_1 \times H_2 \triangleleft G_1 \times G_2$.
 (b) Prove $(G_1 \times G_2)/(H_1 \times H_2)$ is isomorphic to $(G_1/H_1) \times (G_2/H_2)$.

The group of symmetries of the cube is a rich source of examples of normal subgroups, homomorphisms, and quotient groups. In the following exercises, G denotes the group of all symmetries of the cube, and G^+ denotes the group of rotation symmetries of the cube. As we saw in Chapter 14, G contains 48 elements, and G^+ is isomorphic to S_4 , the symmetric group on four letters. We are now in a better position to prove such statements rigorously.



Exercise 16.13. To each rotation symmetry of a cube, associate the resulting permutation of the long diagonals. Prove this association is an injective homomorphism (and hence an isomorphism) $G^+ \rightarrow S_4$.

Exercise 16.14. Let $-I$ be the antipodal map.

- (a) Prove $\langle -I \rangle \triangleleft G$.

Hint: Use Exercise 16.10.

- (b) Prove $G/\langle -I \rangle \approx S_4$.

Hint: Show the “axis permutation map” is well-defined modulo $-I$.

Exercise 16.15. Each symmetry of the cube permutes the three coordinate axes, defining a homomorphism $\phi : G \rightarrow S_3$.

- (a) List the elements of $G^+ \cap \ker \phi$.

Hint: Which rotations of the cube preserve *all three* coordinate axes?

- (b) List all the elements of $\ker \phi$.

Suggestion: How does part (a) help?

Exercise 16.16. Let $G_0 \subseteq G$ be the group of rotations of the cube preserving the axis 0.

- (a) Prove $G_0 \approx S_3$.

- (b) Is $G_0 \triangleleft G$? If not, why not? If so, what is G/G_0 ?

Exercise 16.17. Let $H \subseteq G$ be the group generated by all half-turn symmetries of the cube.

- (a) List the elements of H .

- (b) Is $H \triangleleft G$? If not, why not? If so, what is G/H ?

Chapter 17

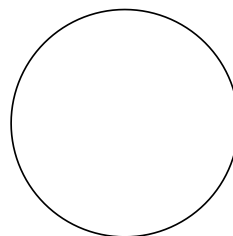
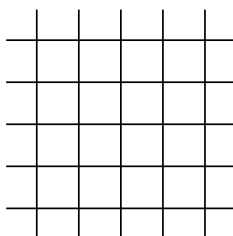
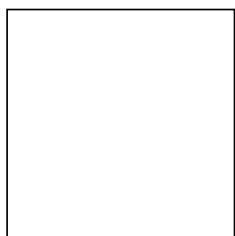
Continuous Symmetries

Group theory arises naturally when we consider “symmetries” of an object or structure, namely, invertible transformations preserving some property. Many of the groups we have encountered so far act on finite sets, such as the set of vertices of a plane polygon. A group of symmetries acting on a finite set \mathcal{U} is, up to isomorphism, contained in the finite group $S_{\mathcal{U}}$ of permutations of \mathcal{U} .

This chapter introduces two settings in which infinite groups arise: the geometry of the Euclidean plane, whose group of rigid motions is infinite, and time-independent differential equations, in which the “time translation” group $(\mathbf{R}, +)$ acts on the set of “instants” or “states of the universe”, providing a simple model of “the passage of time”.

Types of Symmetry

Consider three familiar shapes: a square, an infinite grid of squares, and a circle.



As noted above, a square has only finitely many symmetries: Label the vertices 0, 1, 2, 3, and note that (i) vertex 0 can map only to finitely many locations (namely to another vertex), and (ii) once the image of 0

is known, a rigid motion is completely determined up to reflection over the diagonal containing the image of 0. Conceptually, a square possesses finitely many “distinguished” points, and any symmetry must preserve the set of distinguished points.

The square grid has infinitely many symmetries, including arbitrary horizontal and vertical translations by an integer multiple of the mesh size. As with corners of the square, grid intersection points are “distinguished”, but now there are infinitely many such points.

The group of symmetries of the grid is “discrete”; we cannot “continuously deform” the identity map to another symmetry without temporarily leaving the set of symmetries. For example, if we slide the grid one unit to the right, the intermediate translations, such as sliding by one-half of a unit, are not symmetries of the grid, since they do not map the grid to itself.

Contrast with the symmetries of the circle. A circle has infinitely many symmetries, including rotation about the center through an arbitrary real angle. Rotational symmetries of the circle are “continuous”, in that rotating through angle θ may be accomplished by rotating through angle $t\theta$ as the parameter t “runs” from 0 to 1, and every intermediate transformation is also a symmetry of the circle. Conceptually, *no* point of the circle is distinguished; every point is just like every other from the perspective of Euclidean geometry.

17.1 Euclidean Planar Motions

In this section we describe all transformations of the plane that preserve distances between arbitrary pairs of points. Ideas we’ve studied can be used to break this problem into more manageable pieces. You may want to review the geometry of linear transformations in Chapter 11.

Distance and Isometries

In coordinate geometry, you learned the formula for the distance between points with Cartesian coordinates $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2}.$$

This is nothing but the Pythagorean theorem, the segment joining the points being the hypotenuse of a right triangle with sides parallel to the coordinate axes, Figure 17.1.

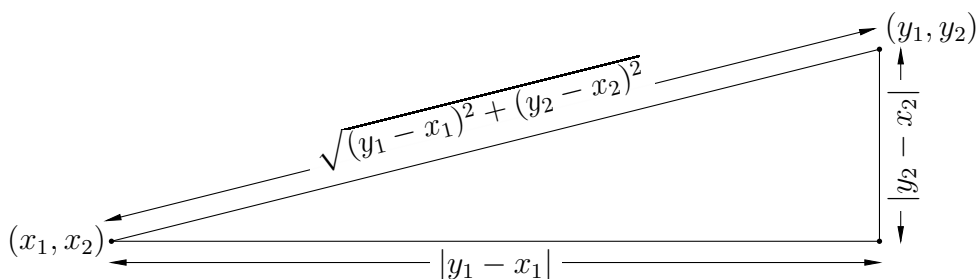


Figure 17.1: The Euclidean distance formula.

Definition 17.1. If $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$ are plane vectors, their *dot product* is

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 = \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \mathbf{u}^T \mathbf{v}.$$

The quantity $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}} = \sqrt{u_1^2 + u_2^2}$ is the *length* of \mathbf{u} . In terms of length, the distance between points \mathbf{x} and \mathbf{y} is

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{y} - \mathbf{x}\|.$$

Definition 17.2. A mapping $\phi : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ is a *rigid motion* or an *isometry* if $d(\phi(\mathbf{x}), \phi(\mathbf{y})) = d(\mathbf{x}, \mathbf{y})$ for all \mathbf{x} and \mathbf{y} in \mathbf{R}^2 .

Example 17.3. The identity map $I : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ is an isometry.

Example 17.4. A composition of isometries is an isometry: If ϕ and ψ are isometries of \mathbf{R}^2 , \mathbf{x}_1 and \mathbf{x}_2 are arbitrary elements of \mathbf{R}^2 , and we set $\mathbf{y}_i = \phi(\mathbf{x}_i)$ and $\mathbf{z}_i = \psi(\mathbf{y}_i) = (\psi\phi)(\mathbf{x}_i)$ for $i = 1, 2$, then

$$d((\psi\phi)(\mathbf{x}_1), (\psi\phi)(\mathbf{x}_2)) = d(\mathbf{z}_1, \mathbf{z}_2) = d(\mathbf{y}_1, \mathbf{y}_2) = d(\mathbf{x}_1, \mathbf{x}_2),$$

so $\psi\phi$ is an isometry.

If we knew that every isometry $\phi : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ were a bijection of \mathbf{R}^2 , it would follow easily that the inverse map ϕ^{-1} is an isometry. This, in turn, would imply the set of isometries of the plane is a group under mapping composition.

Now, every isometry of the plane *is* in fact a bijection. The group of isometries of \mathbf{R}^2 under mapping composition is called the *isometry group* of the plane, or the group of *Euclidean motions*, denoted $\mathcal{E}(2)$.

Why is every isometry of the plane a bijection? Injectivity is obvious: If $\phi(\mathbf{x}_1) = \phi(\mathbf{x}_2)$, then $0 = d(\phi(\mathbf{x}_1), \phi(\mathbf{x}_2)) = d(\mathbf{x}_1, \mathbf{x}_2)$, which implies $\mathbf{x}_1 = \mathbf{x}_2$. The proof of surjectivity, however, relies implicitly on deeper properties of Euclidean geometry, not merely the definition of an isometry! The strategy, outlined in the exercises, is to prove that every Euclidean plane isometry has the form $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, with A a 2×2 real orthogonal matrix and \mathbf{b} in \mathbf{R}^2 . For such mappings, surjectivity is clear.

We first turn our attention to “affine” isometries, those expressible in the form $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$. This allows us to become acquainted with some familiar mappings in a formal setting and to investigate important subgroups of $\mathcal{E}(2)$, all with the foreknowledge that we’re really “meeting” all isometries, not merely a select group.

The Translation Group

Example 17.5. If $\mathbf{b} = (b_1, b_2) \in \mathbf{R}^2$, then the mapping $\phi = T_{\mathbf{b}}$ defined by

$$T_{\mathbf{b}}(\mathbf{x}) = \mathbf{x} + \mathbf{b}, \quad \mathbf{x} \in \mathbf{R}^2,$$

called *translation* by \mathbf{b} , is an isometry. Indeed,

$$\begin{aligned} d(T_{\mathbf{b}}(\mathbf{x}), T_{\mathbf{b}}(\mathbf{y})) &= d(\mathbf{x} + \mathbf{b}, \mathbf{y} + \mathbf{b}) \\ &= \|(\mathbf{y} + \mathbf{b}) - (\mathbf{x} + \mathbf{b})\| \\ &= \|\mathbf{y} - \mathbf{x}\| = d(\mathbf{x}, \mathbf{y}) \quad \text{for all } \mathbf{x} \text{ and } \mathbf{y} \text{ in } \mathbf{R}^2. \end{aligned}$$

Example 17.6. The set \mathcal{T} of *all* translations of the plane is a subgroup of $\mathcal{E}(2)$ isomorphic to the additive group $(\mathbf{R}^2, +)$. This claim amounts to little more than the associative and commutative laws for vector addition:

$$T_{\mathbf{b}_1} T_{\mathbf{b}_2}(\mathbf{x}) = (\mathbf{x} + \mathbf{b}_2) + \mathbf{b}_1 = \mathbf{x} + (\mathbf{b}_2 + \mathbf{b}_1) = T_{\mathbf{b}_1 + \mathbf{b}_2}(\mathbf{x}).$$

In words, the net result of successive translations is translation by the sum of the displacements.

Example 17.7. (Lattices) Let \mathbf{v}_1 and \mathbf{v}_2 be non-proportional elements of \mathbf{R}^2 . The set $\Lambda = \{n_1\mathbf{v}_1 + n_2\mathbf{v}_2 : n_1, n_2 \text{ integers}\}$ is called the (integer) *lattice* generated by \mathbf{v}_1 and \mathbf{v}_2 . An integer lattice is an additive subgroup of \mathbf{R}^2 . The set of translations by elements of Λ is a subgroup of the translation group \mathcal{T} .

The parallelogram $D = \{t_1\mathbf{v}_1 + t_2\mathbf{v}_2 : 0 \leq t_i \leq 1\}$ is called a *fundamental domain* for the group of translations by elements of Λ , see Figure 17.2. Conceptually, a fundamental domain is a bathroom tile and \mathbf{R}^2 is an infinite wall covered with translated copies of the fundamental domain.

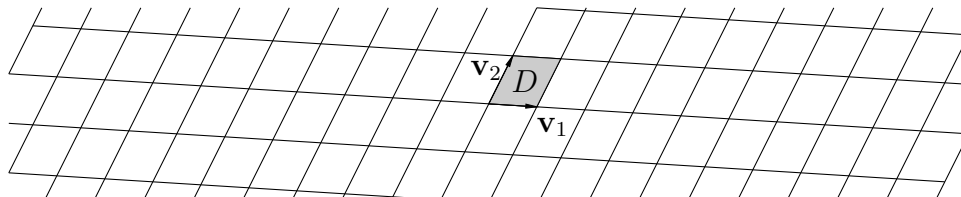


Figure 17.2: The fundamental domain of a lattice. Lattice elements are grid intersections, and grid parallelograms are “copies” of D translated by elements of Λ .

Algebraically, every element of the plane is a real linear combination $c_1\mathbf{v}_1 + c_2\mathbf{v}_2$, and the real number c_i can be written uniquely as an integer n_i plus a real number $0 \leq t_i < 1$. Thus,

$$\begin{aligned} c_1\mathbf{v}_1 + c_2\mathbf{v}_2 &= (n_1 + t_1)\mathbf{v}_1 + (n_2 + t_2)\mathbf{v}_2 \\ &= \underbrace{n_1\mathbf{v}_1 + n_2\mathbf{v}_2}_{\in \Lambda} + \underbrace{t_1\mathbf{v}_1 + t_2\mathbf{v}_2}_{\in D}. \end{aligned}$$

In words, every element of the plane lies in some “copy” of the fundamental domain translated by an element of Λ , and this representation is (almost) unique.

Example 17.8. Let $\Lambda \subseteq \mathbf{R}^2$ be the lattice generated by \mathbf{v}_1 and \mathbf{v}_2 . Since the additive group \mathbf{R}^2 is Abelian, the subgroup Λ is a normal subgroup. The quotient group $T = \mathbf{R}^2/\Lambda$ is called the *torus* associated to the lattice Λ .

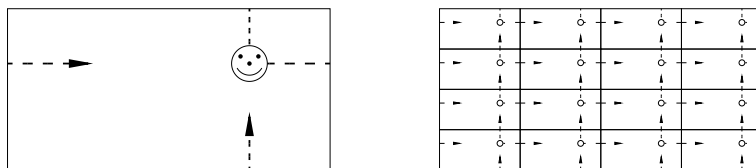


Figure 17.3: A torus from “inside”.

From “inside”, a torus looks like the universe of a video game such as Asteroids or Pac Man. The fundamental domain is the screen on which the game is played. Traveling off the right edge of the screen causes the player’s avatar to return instantly to the corresponding point on the left-hand side, and *vice versa*. Similarly, traveling off the top of the screen returns one at the bottom and *vice versa*.

From “outside”, a torus can be visualized (at least if D is a rectangle) by “rolling up” D into a cylinder, then stretching and bending the ends of the cylinder together to form the surface of a donut, Figure 17.4.

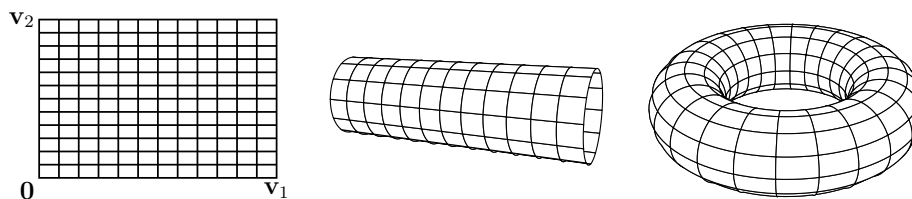


Figure 17.4: Rolling up a fundamental domain into a torus.

Let $D \subseteq \mathbf{R}^2$ be a fixed choice of fundamental domain for Λ . Elements of T are left cosets of Λ , namely, translates $(x, y) + \Lambda$. Two points (x_1, y_1) and (x_2, y_2) represent the same point of T if and only if their difference is an element of the lattice: $(x_2 - x_1, y_2 - y_1) \in \Lambda$, or

$$(x_2, y_2) = (x_1, y_1) + n_1 \mathbf{v}_1 + n_2 \mathbf{v}_2.$$

Consequently, every point of the torus corresponds to an element of D , and distinct elements of D represent distinct elements of T with the following exceptions: Each point on the left edge of D has a matching element on the right edge, each point on the bottom edge of D has a matching element on the top edge, and all four corners represent the same point.

The identity element of T is the left coset Λ . While every non-zero element of \mathbf{R}^2 has infinite order (generates an infinite cyclic group), T has elements of finite order. Indeed, an element \mathbf{x} in \mathbf{R}^2 projects to an element of finite order in T if and only if $n\mathbf{x} \in \Lambda$ for some integer n , if and only if there exist *rational* numbers r_i such that $\mathbf{x} = r_1 \mathbf{v}_1 + r_2 \mathbf{v}_2$.

Linear Isometries

Let A be a 2×2 matrix with real entries. Under what conditions does the linear mapping $T(\mathbf{x}) = A\mathbf{x}$ define an isometry? The following gives

convenient necessary and sufficient conditions.

Theorem 17.9. *Let $T : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ be a linear transformation. The following are equivalent:*

- (i) $T(\mathbf{u}) \cdot T(\mathbf{v}) = \mathbf{u} \cdot \mathbf{v}$ for all \mathbf{u} and \mathbf{v} in \mathbf{R}^2 ,
- (ii) $d(T(\mathbf{0}), T(\mathbf{x})) = d(\mathbf{0}, \mathbf{x})$ for all \mathbf{x} in \mathbf{R}^2 ,
- (iii) T is an isometry.

Proof. ((i) implies (ii)). Assume $\mathbf{x} \in \mathbf{R}^2$. Condition (i) with $\mathbf{u} = \mathbf{v} = \mathbf{x}$ implies

$$\|T(\mathbf{x})\|^2 = T(\mathbf{x}) \cdot T(\mathbf{x}) = \mathbf{x} \cdot \mathbf{x} = \|\mathbf{x}\|^2,$$

or $\|T(\mathbf{x})\| = \|\mathbf{x}\|$. Since T is linear, $T(\mathbf{0}) = \mathbf{0}$, so

$$d(T(\mathbf{0}), T(\mathbf{x})) = \|T(\mathbf{0}) - T(\mathbf{x})\| = \|T(\mathbf{x})\| = \|\mathbf{x}\| = \|\mathbf{0} - \mathbf{x}\| = d(\mathbf{0}, \mathbf{x})$$

for all \mathbf{x} in \mathbf{R}^2 .

((ii) implies (iii)). Let \mathbf{x} and \mathbf{y} be arbitrary elements of \mathbf{R}^2 . Then

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= d(\mathbf{0}, \mathbf{y} - \mathbf{x}) && \text{translation is an isometry} \\ &= d(T(\mathbf{0}), T(\mathbf{y} - \mathbf{x})) && \text{condition (ii)} \\ &= d(T(\mathbf{0}), T(\mathbf{y}) - T(\mathbf{x})) && T \text{ is linear} \\ &= d(T(\mathbf{x}), T(\mathbf{y})). && \text{translation is an isometry} \end{aligned}$$

Since \mathbf{x} and \mathbf{y} were arbitrary, T is an isometry.

((iii) implies (i)). Let \mathbf{u} and \mathbf{v} be arbitrary elements of \mathbf{R}^2 . The dot product can be expressed in terms of the distance function using the *polarization identity*

$$\begin{aligned} \frac{1}{4}(\|\mathbf{u} + \mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2) &= \frac{1}{4}((\mathbf{u} + \mathbf{v}) \cdot (\mathbf{u} + \mathbf{v}) - (\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v})) \\ &= \frac{1}{4}(\|\mathbf{u}\|^2 + 2\mathbf{u} \cdot \mathbf{v} + \|\mathbf{v}\|^2 - (\|\mathbf{u}\|^2 - 2\mathbf{u} \cdot \mathbf{v} + \|\mathbf{v}\|^2)) \\ &= \mathbf{u} \cdot \mathbf{v}. \end{aligned}$$

If T is an isometry, $\|T(\mathbf{x})\| = \|\mathbf{x}\|$ for all \mathbf{x} , since each side can be interpreted as the distance from some vector to the zero vector. Thus

$$\begin{aligned} T(\mathbf{u}) \cdot T(\mathbf{v}) &= \frac{1}{4}(\|T(\mathbf{u}) + T(\mathbf{v})\|^2 - \|T(\mathbf{u}) - T(\mathbf{v})\|^2) && \text{polarization} \\ &= \frac{1}{4}(\|T(\mathbf{u} + \mathbf{v})\|^2 - \|T(\mathbf{u} - \mathbf{v})\|^2) && T \text{ is linear} \\ &= \frac{1}{4}(\|\mathbf{u} + \mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2) && T \text{ is an isometry} \\ &= \mathbf{u} \cdot \mathbf{v} && \text{polarization} \end{aligned}$$

for all \mathbf{u} and \mathbf{v} in \mathbf{R}^2 . □

Theorem 17.9 characterizes linear transformations that are isometries. We'd like an easy-to-use criterion in terms of the 2×2 matrix A defining the linear transformation.

Definition 17.10 (See also Exercise 11.17). A square matrix A is *orthogonal* if $A^\top = A^{-1}$, namely if $A^\top A = AA^\top = I_n$. The set of all 2×2 real orthogonal matrices is denoted $O(2)$.

Proposition 17.11. *A 2×2 matrix A is orthogonal if and only if its columns are mutually orthogonal unit vectors.*

Proof. The (i, j) th entry of $A^\top A$ is $(A_i^\top)(A_j)$, the dot product of the i th column and the j th column of A . To say this entry is δ_{ij} means each column dotted with itself is 1 (each column is a unit vector) and the first column dotted with the second is 0 (the columns are perpendicular). □

Theorem 17.12. *Multiplication by the 2×2 matrix A defines an isometry of \mathbf{R}^2 if and only if $A \in O(2)$, if and only if $A^\top A = I_2$.*

Proof. Recall the dot product $\mathbf{u} \cdot \mathbf{v}$ may be viewed as a matrix product $\mathbf{u}^\top \mathbf{v}$. By Theorem 17.9 (i), multiplication by a 2×2 matrix A defines an isometry if and only if

$$\mathbf{u}^\top (A^\top A) \mathbf{v} = (A\mathbf{u})^\top (A\mathbf{v}) = (A\mathbf{u}) \cdot (A\mathbf{v}) = \mathbf{u} \cdot \mathbf{v} = \mathbf{u}^\top \mathbf{v}$$

for all \mathbf{u} and \mathbf{v} in \mathbf{R}^2 . Writing $A^\top A = [a_{ij}]$ and taking $\mathbf{u} = \mathbf{e}_i$ and $\mathbf{v} = \mathbf{e}_j$ to be standard basis vectors, the preceding condition becomes $a_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}$ (the Kronecker delta symbol), or $A^\top A = I_2$. □

The Orthogonal Group

Proposition 17.13. *The set $O(2)$ of 2×2 orthogonal matrices is a group under matrix multiplication. The subset $SO(2)$ comprising matrices of determinant 1 is a normal subgroup of $O(2)$.*

Definition 17.14. The group $O(2)$ is called the (2-dimensional) *orthogonal group*. The subgroup $SO(2)$ is the *special orthogonal group*. A 2×2 orthogonal matrix of determinant 1 is a *rotation*. A 2×2 orthogonal matrix of determinant -1 is a *reflection*.

Proof of proposition. First, $O(2)$ is closed under matrix multiplication: If A and B are orthogonal, then

$$(AB)^T = B^T A^T = B^{-1} A^{-1} = (AB)^{-1},$$

so AB is also orthogonal.

The identity matrix is obviously orthogonal, either by direct calculation or because its columns are the standard basis vectors, which are mutually orthogonal unit vectors.

To prove the inverse of an orthogonal matrix is orthogonal, first recall the formula for the inverse of a 2×2 matrix:

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

It follows that $(A^{-1})^T = (A^T)^{-1}$ for every invertible 2×2 matrix.

In particular, if A is orthogonal and $B = A^{-1} = A^T$, then

$$B^T = (A^{-1})^T = (A^T)^{-1} = (A^{-1})^{-1} = A = B^{-1};$$

the inverse of an orthogonal matrix is orthogonal.

To see $SO(2) \triangleleft O(2)$, note that $\det : O(2) \rightarrow \mathbf{R}^\times$ is a homomorphism with kernel $SO(2) = \det^{-1}(1)$. By Proposition 16.9, the kernel of a homomorphism is a normal subgroup of the domain. \square

Every unit vector in the plane can be written $\mathbf{e}_1(\theta) = (\cos \theta, \sin \theta)$ for some real θ (well-defined modulo 2π). There are precisely two unit vectors perpendicular to this: $\mathbf{e}_2(\theta) = (-\sin \theta, \cos \theta)$ and its negative, $(\sin \theta, -\cos \theta)$. These two choices give rise to two orthogonal matrices, respectively a rotation ($\det = 1$) and a reflection ($\det = -1$):

$$\text{Rot}_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}, \quad \text{Ref}_{\theta/2} = \begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{bmatrix}.$$

The corresponding linear transformations act on the standard F as shown in Figure 17.5.

The Rotation Group

As shown abstractly above, the set $SO(2)$ of 2×2 orthogonal matrices of determinant 1 is a group under matrix multiplication. This can also be shown explicitly, since a composition of rotations is a rotation (see

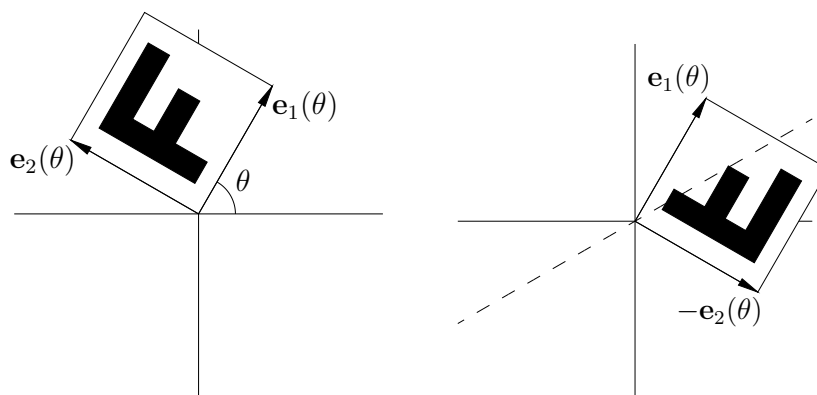


Figure 17.5: The actions of Rot_θ and $\text{Ref}_{\theta/2}$ on the standard F.

below), the identity map is a rotation (with $\theta = 0$), and the inverse of a rotation is a rotation.

Algebraically, the sum formulas for the trig functions give

$$\begin{aligned} \text{Rot}_\theta \text{Rot}_\phi &= \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{bmatrix} \\ &= \begin{bmatrix} \cos \theta \cos \phi - \sin \theta \sin \phi & -\cos \theta \sin \phi - \sin \theta \cos \phi \\ \sin \theta \cos \phi + \cos \theta \sin \phi & -\sin \theta \sin \phi + \cos \theta \cos \phi \end{bmatrix} \\ &= \begin{bmatrix} \cos(\theta + \phi) & -\sin(\theta + \phi) \\ \sin(\theta + \phi) & \cos(\theta + \phi) \end{bmatrix} = \text{Rot}_{\theta+\phi}. \end{aligned}$$

As expected, “angles add” under successive rotations.

Remark 17.15. The preceding equation is the morphism condition for the mapping $\phi : \mathbf{R} \rightarrow SO(2)$ defined by $\phi(\theta) = \text{Rot}_\theta$. The kernel of ϕ is the set of real θ for which $\phi(\theta) = I_2$, namely $2\pi\mathbf{Z}$, the set of integer multiples of 2π . By Theorem 16.22, $SO(2) \simeq \mathbf{R}/2\pi\mathbf{Z}$. Conceptually, a rotation about the origin can be made through an arbitrary real angle, but the result is only well-defined up to integer multiples of a full turn. (Compare Examples 4.57 and 4.62.)

Composition of Orthogonal Transformations

The orthogonal group contains elements of two types, see Figure 17.5, characterized by their determinants:

$$\text{Rot}_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}, \quad \text{Ref}_{\theta/2} = \begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{bmatrix}.$$

Geometrically, the linear transformation Rot_θ rotates the plane counterclockwise about the origin through an angle of θ radians. The transformation $\text{Ref}_{\theta/2}$ reflects the plane about the line through the origin and making an angle of $\theta/2$ with the positive horizontal axis.

Remark 17.16. The set of 2×2 reflection matrices is the left coset $O(2) \setminus SO(2)$, which is not a subgroup of $O(2)$. Every reflection matrix is its own inverse, since $A^T = A$.

Proposition 17.17. *For real θ , $\text{Rot}_\theta \text{Ref}_0 \text{Rot}_{-\theta} = \text{Ref}_\theta$ is reflection about the line making angle θ with the positive horizontal axis.*

Proof. It suffices to check

$$\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix},$$

a straightforward calculation. \square

Remark 17.18. Geometrically, this result can be interpreted in terms of conjugation. Rotating clockwise by θ , reflecting across the horizontal axis, and rotating counterclockwise by θ is the same as reflecting across the line at angle θ , which is obtained from the horizontal axis by a counterclockwise rotation by θ .

Example 17.19. Though there are infinitely many 2×2 orthogonal matrices, a useful “Cayley table” of $O(2)$ can be formed as follows:

\cdot	Rot_ϕ	$\text{Ref}_{\phi/2}$
Rot_θ	$\text{Rot}_{(\theta+\phi)}$	$\text{Ref}_{\frac{1}{2}(\theta+\phi)}$
$\text{Ref}_{\theta/2}$	$\text{Ref}_{\frac{1}{2}(\theta-\phi)}$	$\text{Rot}_{(\theta-\phi)}$

In particular, the orthogonal group $O(2)$ is non-Abelian.

The easiest ways to justify the entries of this table are to calculate appropriate matrix products, or to find the aggregate effect of a composition on the standard basis vector \mathbf{e}_1 , and to remember that composing a rotation with a reflection (in either order) gives a reflection, while composing two rotations or two reflections gives a rotation.

Remark 17.20. As is customary with Cayley tables, the “product” ST corresponds to the entry S on the left side of the table and the entry T along the top of the table. Since these are mappings, however, T is applied *first*, then S .

Affine Isometries

Composition of translations and orthogonal transformations gives additional isometries of the plane: If $A \in \mathbf{R}^{2 \times 2}$ is orthogonal and $\mathbf{b} \in \mathbf{R}^2$, the *affine map* $T_{A,\mathbf{b}}(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ is an isometry.

Since $T_{A,\mathbf{b}}(\mathbf{0}) = \mathbf{b}$ is not generally $\mathbf{0}$, the map $T_{A,\mathbf{b}} = T_{\mathbf{b}} T_A$ cannot generally be represented using only multiplication of 2×2 matrices. However, $T_{A,\mathbf{b}}$ can be represented using multiplication of 3×3 matrices. To accomplish this, “suspend” each element of \mathbf{R}^2 by lifting one unit along a third coordinate axis. Algebraically, identify $\mathbf{x} = (x_1, x_2)$ in \mathbf{R}^2 with $\widehat{\mathbf{x}} = (x_1, x_2, 1)$ in \mathbf{R}^3 . Now form the “block matrix”

$$\widehat{A, \mathbf{b}} = \begin{bmatrix} A & \mathbf{b} \\ \mathbf{0}^\top & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ 0 & 0 & 1 \end{bmatrix}.$$

A block matrix calculation shows

$$\widehat{A, \mathbf{b}} \cdot \widehat{\mathbf{x}} = \begin{bmatrix} A & \mathbf{b} \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} = \begin{bmatrix} A\mathbf{x} + \mathbf{b} \\ 1 \end{bmatrix} = \begin{bmatrix} T_{A,\mathbf{b}}(\mathbf{x}) \\ 1 \end{bmatrix} = \widehat{T_{A,\mathbf{b}}(\mathbf{x})}.$$

To be sure you understand this calculation, expand it out in terms of 3×3 matrices and elements of \mathbf{R}^3 .

The transformation $T_{A,\mathbf{b}}$ is essentially an ordered pair (A, \mathbf{b}) consisting of an orthogonal 2×2 matrix and an element of \mathbf{R}^2 . The set of all such pairs can be made into a group, with the group operation coming from composition of affine transformations. Since

$$\begin{aligned} (T_{A_2, \mathbf{b}_2} T_{A_1, \mathbf{b}_1})(\mathbf{x}) &= A_2(A_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2 \\ &= (A_2A_1)\mathbf{x} + (A_2\mathbf{b}_1 + \mathbf{b}_2), \end{aligned}$$

the composition rule is

$$(17.1) \quad (A_2, \mathbf{b}_2) \cdot (A_1, \mathbf{b}_1) = (A_2A_1, A_2\mathbf{b}_1 + \mathbf{b}_2).$$

It turns out *every* isometry of the plane is an affine transformation, see Exercises 17.1–17.6.

Remark 17.21. Though the translation group \mathcal{T} and the rotation group $SO(2)$ are Abelian, by (17.1) they generate a non-Abelian subgroup of $\mathcal{E}(2)$, the group $\mathcal{E}^+(2)$ of *orientation preserving* isometries.

Example 17.22. The translation group \mathcal{T} is a *normal* subgroup of the isometry group. Let $T_{A,\mathbf{b}}$ be an arbitrary isometry and assume $\mathbf{b}' \in \mathbf{R}^2$. Since $A\mathbf{x} + \mathbf{b} = \mathbf{y}$ if and only if $\mathbf{x} = A^{-1}(\mathbf{y} - \mathbf{b}) = A^{-1}\mathbf{y} - A^{-1}\mathbf{b}$, we have

$$\begin{aligned} T_{A,\mathbf{b}} T_{\mathbf{b}'} (T_{A,\mathbf{b}})^{-1}(\mathbf{y}) &= T_{A,\mathbf{b}} T_{\mathbf{b}'}(A^{-1}\mathbf{y} - A^{-1}\mathbf{b}) \\ &= T_{A,\mathbf{b}}(A^{-1}\mathbf{y} - A^{-1}\mathbf{b} + \mathbf{b}') \\ &= A(A^{-1}\mathbf{y} - A^{-1}\mathbf{b} + \mathbf{b}') + \mathbf{b} \\ &= \mathbf{y} - \mathbf{b} + A\mathbf{b}' + \mathbf{b} = \mathbf{y} + A\mathbf{b}', \end{aligned}$$

which is translation by $A\mathbf{b}'$, hence an element of \mathcal{T} .

Alternatively, the inverse of $T_{A,\mathbf{b}}$ is $T_{A^{-1}, -A^{-1}\mathbf{b}}$, so

$$\begin{aligned} T_{A,\mathbf{b}} T_{\mathbf{b}'} (T_{A,\mathbf{b}})^{-1} &= T_{A,\mathbf{b}} T_{\mathbf{b}'} T_{A^{-1}, -A^{-1}\mathbf{b}} \\ &= T_{A,\mathbf{b}} T_{A^{-1}, -A^{-1}\mathbf{b} + \mathbf{b}'} \\ &= T_{I, A\mathbf{b}'} = T_{A\mathbf{b}'} \end{aligned}$$

by the composition rule (17.1).

Remark 17.23. Much of the formalism developed for $\mathcal{E}(2)$, including the composition rule (17.1), generalizes to Euclidean spaces of dimension three or more. However, rotations of \mathbf{R}^n , $n \geq 3$, are more complicated to describe than plane rotations, and the rotation group is non-Abelian in dimension $n \geq 3$.

17.2 Time-Invariant Planar Flows

Definition 17.24. Let $D \subseteq \mathbf{R}^2$ be a region. A function $f : D \rightarrow \mathbf{R}$ is *smooth* if f has continuous partial derivatives of all orders. A mapping $\mathbf{f} : D \rightarrow \mathbf{R}^2$ is smooth if each component function of \mathbf{f} is smooth.

A bijection $\phi : D \rightarrow D$ is a *diffeomorphism* of D if both ϕ and the inverse mapping ϕ^{-1} are smooth.

Example 17.25. Let $D = \mathbf{R}^2$ be the entire plane. The following maps are diffeomorphisms:

- (i) Translation by a fixed vector \mathbf{b} in \mathbf{R}^2 ,

$$T_{\mathbf{b}}(x_1, x_2) = (x_1 + b_1, x_2 + b_2).$$

(ii) Rotation about the origin by an angle θ ,

$$\text{Rot}_\theta(x_1, x_2) = (x_1 \cos \theta - x_2 \sin \theta, x_1 \sin \theta + x_2 \cos \theta).$$

(iii) Scaling about the origin, $\mathbf{f}(x_1, x_2) = (cx_1, cx_2)$ with $c \neq 0$.

(iv) The “non-linear shear”

$$\mathbf{f}(x_1, x_2) = (x_1, x_2 + f(x_1)),$$

with $f : \mathbf{R} \rightarrow \mathbf{R}$ an infinitely-differentiable function.

A composition of smooth mappings is smooth by the chain rule and induction on the number of derivatives. Consequently, a composition of diffeomorphisms is a diffeomorphism. The identity map is obviously a diffeomorphism of an arbitrary region D , and the inverse of a diffeomorphism is a diffeomorphism. The set $\text{Diff}(D)$ of all diffeomorphisms of D therefore forms a group under mapping composition, naturally called the *diffeomorphism group* of D .

Think of the additive group $(\mathbf{R}, +)$ as representing time, with $t = 0$ meaning now, $t < 0$ representing the past, and $t > 0$ corresponding to the future.

Definition 17.26. A homomorphism $\phi : (\mathbf{R}, +) \rightarrow \text{Diff}(D)$ is called a (time-independent) *flow* on D .

Remark 17.27. A flow ϕ associates, to each time t , a diffeomorphism $\phi_t : D \rightarrow D$ representing “elapsed time t ”. The map ϕ_0 is the identity map of D .

Physically, the morphism condition $\phi_{s+t} = \phi_s \phi_t$ for real s and t means that allowing time to “run” for $s+t$ from some initial condition gives the same result as letting time run for t from the same initial condition, then letting time run for s from the state at time t .

Example 17.28. Let $D = \mathbf{R}^2$ be the plane. The following are flows:

(i) Translation in the direction of a fixed vector \mathbf{b} in \mathbf{R}^2 ,

$$\phi_{\mathbf{b}}(t, x_1, x_2) = (x_1 + tb_1, x_2 + tb_2).$$

(ii) Rotation about the origin,

$$\phi(t, x_1, x_2) = (x_1 \cos t - x_2 \sin t, x_1 \sin t + x_2 \cos t).$$

(iii) Scaling about the origin, $\phi(t, x_1, x_2) = (e^t x_1, e^t x_2)$.

(iv) Any mapping of the form

$$\phi(t, x_1, x_2) = (x_1, x_2 + tf(x_1)),$$

with $f : \mathbf{R} \rightarrow \mathbf{R}$ an infinitely-differentiable function.

Flows arise naturally throughout mathematics and the physical sciences as “solutions” of systems of differential equations. A flow may be viewed as a smooth map $\phi : \mathbf{R} \times D \rightarrow D$, with $\phi(t, x_1, x_2)$ representing the location at time t of the particle having initial position (x_1, x_2) . The velocity of the particle at time t is $\partial\phi/\partial t(t, x_1, x_2) = (\dot{x}_1, \dot{x}_2)$. By the morphism condition, the velocity of a particle at time t depends only on the particle’s location at time t , not on t itself. Consequently, there exists a *vector field* \mathbf{F} on D defined by

$$\mathbf{F}(x_1, x_2) = \left. \frac{\partial\phi}{\partial t} \right|_{t=0} (t, x_1, x_2).$$

Each particle’s motion is governed by the system of differential equations

$$(\dot{x}_1, \dot{x}_2) = \mathbf{F}(x_1, x_2).$$

Example 17.29. The flows in Example 17.28 correspond to the indicated vector fields.

(i) Translation in the direction of a fixed vector \mathbf{b} in \mathbf{R}^2 ,

$$\phi_{\mathbf{b}}(t, x_1, x_2) = (x_1 + tb_1, x_2 + tb_2) \longleftrightarrow \mathbf{F}(x_1, x_2) = \mathbf{b} = (b_1, b_2).$$

(ii) Rotation about the origin,

$$\begin{aligned} \phi(t, x_1, x_2) &= (x_1 \cos t - x_2 \sin t, x_1 \sin t + x_2 \cos t) \\ &\longleftrightarrow \mathbf{F}(x_1, x_2) = (-x_2, x_1). \end{aligned}$$

(iii) Scaling about the origin,

$$\phi(t, x_1, x_2) = (e^t x_1, e^t x_2) \longleftrightarrow \mathbf{F}(x_1, x_2) = (x_1, x_2).$$

(iv) The non-linear shear

$$\phi(t, x_1, x_2) = (x_1, x_2 + tf(x_1)) \longleftrightarrow \mathbf{F}(x_1, x_2) = (0, f(x_1)),$$

with $f : \mathbf{R} \rightarrow \mathbf{R}$ an infinitely-differentiable function.

Exercises

The first several exercises ask you to prove every rigid motion of the plane is an affine transformation. The idea is to post-compose an arbitrary isometry $\phi : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ with a “conditioning” isometry so that the composition ψ satisfies tighter and tighter constraints. Without using more than translations and orthogonal transformations to condition ϕ , we can arrange that ψ is the identity, which means ϕ is affine.

Exercise 17.1. Let ϕ be an isometry, and define $\phi_0(\mathbf{x}) = \phi(\mathbf{x}) - \phi(\mathbf{0})$. Prove ϕ_0 is an isometry of the plane, and $\phi_0(\mathbf{0}) = \mathbf{0}$.

Exercise 17.2. Let ϕ be an isometry of the plane satisfying $\phi(\mathbf{0}) = \mathbf{0}$. Prove ϕ is a linear transformation.

Outline: Set $\mathbf{u}_1 = \phi(\mathbf{e}_1)$ and $\mathbf{u}_2 = \phi(\mathbf{e}_2)$. The aim is to show

$$\phi(c_1\mathbf{e}_1 + c_2\mathbf{e}_2) = c_1\mathbf{u}_1 + c_2\mathbf{u}_2 \quad \text{for all real } c_1, c_2.$$

First use geometry and the isometry condition to prove $\phi(c_1\mathbf{e}_1) = c_1\mathbf{u}_1$. Similarly prove $\phi(c_2\mathbf{e}_2) = c_2\mathbf{u}_2$. Finally, prove ϕ is linear.

Hints: ϕ maps triangles to congruent triangles, and therefore maps triples of collinear points to collinear points. Let \mathbf{x} and \mathbf{y} be arbitrary, distinct points of \mathbf{R}^2 . The general point on the line through \mathbf{x} and \mathbf{y} is $\mathbf{x} + t(\mathbf{y} - \mathbf{x})$, with t real. Use the fact that ϕ preserves distances to prove $\phi(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) = \phi(\mathbf{x}) + t(\phi(\mathbf{y}) - \phi(\mathbf{x}))$.

Exercise 17.3. Let ϕ be an isometry satisfying $\phi(\mathbf{0}) = \mathbf{0}$. Use Theorem 17.9 to prove the vectors $\phi(\mathbf{e}_1)$ and $\phi(\mathbf{e}_2)$ are an orthonormal set, namely, each is a unit vector, and their dot product is equal to 0.

Exercise 17.4. Let ϕ be an isometry satisfying $\phi(\mathbf{0}) = \mathbf{0}$. Prove there exists an orthogonal transformation A such that $A\phi(\mathbf{e}_1) = \mathbf{e}_1$ and $A\phi(\mathbf{e}_2) = \mathbf{e}_2$.

Exercise 17.5. Let ϕ be an isometry fixing the origin and the standard basis vectors. Prove ϕ is the identity map.

Hint: First use the isometry condition to prove ϕ fixes every point on the horizontal axis. Similarly, prove ϕ fixes every point on an arbitrary vertical line.

Exercise 17.6. Let ϕ be an isometry. Use the results of the preceding exercises to prove there exists an orthogonal matrix A and a vector \mathbf{b} in \mathbf{R}^2 such that $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ for all \mathbf{x} in \mathbf{R}^2 .

Exercise 17.7. Explain *geometrically* why the translation group is a normal subgroup of the isometry group of \mathbf{R}^2 .

Hint: You may assume every isometry of \mathbf{R}^2 is affine. If T is translation by \mathbf{b} and A is an orthogonal transformation, express the action of ATA^{-1} in terms of A and the displacement \mathbf{b} . It may help to consider separately the cases where A is a rotation or a reflection.

Exercise 17.8. An isometry ϕ of \mathbf{R}^2 is *orientation preserving* if the angle from $\phi(\mathbf{e}_1)$ to $\phi(\mathbf{e}_2)$ is counterclockwise, and is *orientation reversing* if this angle is clockwise.

- (a) Prove ϕ is orientation preserving if and only if there exist a rotation matrix A and a vector \mathbf{b} such that $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ for all \mathbf{x} in \mathbf{R}^2 .
- (b) Prove that the set $\mathcal{E}^+(2)$ of orientation-preserving isometries is a non-Abelian subgroup of the isometry group $\mathcal{E}(2)$. (That is, verify the claim made in Remark 17.21.)
- (c) Let $\text{Ref}_0 : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ be the reflection $\text{Ref}_0(x, y) = (x, -y)$. Prove that $\mathcal{E}(2) = \mathcal{E}^+(2) \cup \text{Ref}_0 \mathcal{E}^+(2)$. Conclude that this is the decomposition of the isometry group into left cosets of $\mathcal{E}^+(2)$, and that $\mathcal{E}^+(2) \triangleleft \mathcal{E}(2)$.

Exercise 17.9. Let ϕ be an orientation-preserving isometry that is *not* a translation. That is, $\phi(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ for some A in $SO(2)$ and \mathbf{b} in \mathbf{R}^2 , and $A \neq I$. Prove then ϕ has a fixed point: There exists \mathbf{x}_0 in \mathbf{R}^2 such that $\phi(\mathbf{x}_0) = \mathbf{x}_0$.

Hint: Use Proposition 11.31 to show $A - I$ is invertible.

Exercise 17.10. Let $\text{Ref}_0 : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ be the reflection defined by $\text{Ref}_0(x, y) = (x, -y)$, and assume $\mathbf{b} \in \mathbf{R}^2$.

- (a) Describe the transformation $T_{\mathbf{b}} \text{Ref}_0(T_{\mathbf{b}})^{-1}$ geometrically. In particular, which points of \mathbf{R}^2 , if any, are fixed by this transformation?
- (b) Confirm your answer to part (a) with an algebraic calculation.

Exercise 17.11. Let a and b be real numbers.

- (a) Show the vector field $\mathbf{F}(x, y) = (ax, by)$ on the plane \mathbf{R}^2 corresponds to the flow $\phi(t, x, y) = (xe^{at}, ye^{bt})$.

(b) For each of the following pairs (a, b) of parameters, sketch at least two solution curves in each quadrant, and one curve on each half of each coordinate axis.

- (i) $(1, 2)$. (ii) $(1, -1)$. (iii) $(-1, -2)$. (iv) $(3, 2)$. (v) $(1, 0)$.

Exercise 17.12. This exercise introduces a flow, closely related to rotation about the origin, arising in physics (special relativity) and pure mathematics (hyperbolic geometry).

Define the *boost transformation* ϕ_t by

$$\phi_t(x, y) = \begin{bmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \cosh t + y \sinh t \\ x \sinh t + y \cosh t \end{bmatrix}.$$

- (a) Show ϕ is a flow, and find the associated vector field \mathbf{F} .
- (b) Show the image of the horizontal axis under the time- t flow is the line of slope $\tanh t$, the hyperbolic tangent of t , and illustrate with a sketch. What is the analogous assertion for ordinary rotation about the origin?
- (c) Show the hyperbolas $x^2 - y^2 = c$ are preserved by a boost. What is the analogous assertion for ordinary rotation about the origin?
- (d) Sketch the image of the standard F under a boost with $t > 0$, and with $t < 0$. What happens as $t \rightarrow \infty$?
- (e) Let α and β be real numbers, and let $\mathbf{e}_1 = (1, 0)$ be the horizontal standard basis vector. Prove the vector $\phi_{\alpha+\beta}(\mathbf{e}_1)$ lies along the line of slope

$$\tanh(\alpha + \beta) = \frac{\tanh \alpha + \tanh \beta}{1 + (\tanh \alpha)(\tanh \beta)} = \frac{a + b}{1 + ab},$$

where we have written $a = \tanh \alpha$, $b = \tanh \beta$. (This formula describes addition of velocities in special relativity, with a and b interpreted as fractions of the speed of light. The bijective mapping $\tanh : \mathbf{R} \rightarrow (-1, 1)$ is an isomorphism from the additive group of reals to the group studied in Exercise 7.15. Physically, closure of $(-1, 1)$ means added velocities never exceed the speed of light.)

Euler's Formula

In calculus, you may have encountered the Maclaurin series for the exponential function and the circular trig functions:

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \frac{x^7}{7!} + \dots, \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots, \\ \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \end{aligned}$$

If $i^2 = -1$, these series are related by *Euler's formula*

$$e^{i\theta} = \cos \theta + i \sin \theta.$$

Indeed, the powers of i are $i^0 = 1$, $i^1 = i$, $i^2 = -1$, $i^3 = -i$, and the cyclic pattern continues *ad infinitum*. If we set $x = i\theta$ in the exponential series, then formally

$$\begin{aligned} e^{i\theta} &= 1 + i\theta + \frac{(i\theta)^2}{2!} + \frac{(i\theta)^3}{3!} + \frac{(i\theta)^4}{4!} + \frac{(i\theta)^5}{5!} + \frac{(i\theta)^6}{6!} + \frac{(i\theta)^7}{7!} + \dots \\ &= 1 + i\theta - \frac{\theta^2}{2!} - i\frac{\theta^3}{3!} + \frac{\theta^4}{4!} + i\frac{\theta^5}{5!} - \frac{\theta^6}{6!} - i\frac{\theta^7}{7!} + \dots \\ &= \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \dots\right) + i\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \dots\right) \\ &= \cos \theta + i \sin \theta. \end{aligned}$$

By elementary trigonometry, we have

$$\begin{aligned} e^{\pm 2\pi i} &= \cos(2\pi) \pm i \sin(2\pi) = 1 = e^0, \\ e^{\pm \pi i} &= \cos(\pi) \pm i \sin(\pi) = -1, \\ e^{\pm \pi i/2} &= \cos(\pi/2) \pm i \sin(\pi/2) = \pm i. \end{aligned}$$

Similar formulas may be given for $e^{\pm\pi i/3}$, $e^{\pm\pi i/4}$, and $e^{\pm\pi i/6}$.

In Chapter 2 we saw how Euler's formula together with the addition formulas for cosine and sine implies the law of exponents for the complex exponential function. Conversely, the law of exponents (which most people find easy to remember) can be used to give an easy derivation of the addition formulas for cosine and sine (which most people find difficult to remember).

If a and b are real numbers, then

$$\begin{aligned}\cos(a+b) + i\sin(a+b) &= e^{i(a+b)} = e^{ia} e^{ib} \\ &= (\cos a + i\sin a)(\cos b + i\sin b) \\ &= (\cos a \cos b - \sin a \sin b) + i(\sin a \cos b + \cos a \sin b).\end{aligned}$$

Two complex numbers $A + Bi$ and $C + Di$ are equal if and only if their real parts are equal ($A = C$) and their imaginary parts are equal ($B = D$). That is,

$$\begin{aligned}\cos(a+b) &= \cos a \cos b - \sin a \sin b, \\ \sin(a+b) &= \sin a \cos b + \cos a \sin b.\end{aligned}$$

Index

- All-you-can-eat buffet, *see* Buffet
- Alternating group, 204, 238
 - cyclic subgroups of A_4 , 215
 - list of elements of A_4 , 214
 - subgroups of A_4 , 230
- Axioms for the integers, 34
- Binary operation, 93
 - associativity of, 96, 101
 - Cayley table of, 94
 - commutativity of, 100
 - identity element for, 98
 - inverse elements under, 99
 - uniqueness of, 102
- Binomial coefficient
 - definition of, 86
- Binomial theorem, 86
- Boost transformation, 189, 264
 - group law for, 123
- Buffet
 - all-you-can-eat, 84
- Cancellation law in a group, 108
- Cartesian product, 46
- Cayley table, 94
 - and isomorphism, 182
 - of the orthogonal group, 257
- Closed under
 - a binary operation, 95, 113
 - addition, 16
 - inversion, 113
 - multiplication, 21
- Complex conjugate, 15, 177, 187
- Complex number
 - argument of, 20
 - imaginary part of, 15
 - magnitude of, 20
 - non-real, 15
 - real part of, 15
 - unit, 21
- Complex numbers
 - groups of, 118, 188
 - lattice subgroup of, 119, 250
 - product of, 18
- Congruence mod n , 126
- Coprime integers, 136
- cosh, *see* Hyperbolic trig functions
- Cube
 - symmetry group of, 217, 245
- Cyclic group, 111, 228
 - classification of, 184
 - homomorphic image of, 239
 - as symmetries of an n -gon, 209
- De Morgan's laws, 4, 29
- Dihedral group, 209, 221
- Division algorithm, 41

-
- Equivalence class
 - definition of, 60
 - mod n , 126
 - Euclid's algorithm, 132
 - Euler's formula, 20, 68, 92, 120, 265
 - Factorial, definition of, 77
 - Fibonacci numbers, 77
 - Fundamental theorem of arithmetic, 138
 - Gaussian integers, 17
 - as subgroup, 119
 - Greatest common divisor, 133, 141
 - definition of, 130
 - Euclid's algorithm, 132
 - Group
 - action, 207
 - cancellation law, 108
 - cyclic, 111, 184, 228
 - finitely generated, 116
 - of units, 146
 - of order 9, 229
 - Group element
 - integer power of, 110
 - order of, 116
 - Groups
 - direct product of, 109, 134
 - of order 8, 222
 - of prime order, 228
 - Heisenberg group, 176
 - Hexagon
 - symmetry group of, 212, 245
 - Homomorphic image, 235, 240
 - Homomorphisms
 - definition of, 235
 - existence of, 239
 - kernel of, 235
 - and left cosets, 238
 - Hyperbolic trig functions, 189, 264
 - Image of a set, 47
 - Induced mapping, 63
 - Induction, mathematical, 71
 - Integers
 - axioms for, 34
 - coprime, 136
 - even, 15
 - odd, 15
 - positive, 14
 - prime, 135
 - residue classes of, 126
 - unique factorization of, 138
 - Isometry
 - definition of, 249
 - Isometry group, 249
 - structure of, 262
 - Isomorphic groups
 - properties of, 180
 - Isomorphism
 - as a commutative diagram, 95
 - and law of exponents, 181
 - natural logarithm as, 181
 - of quotient group, 243
 - Joke
 - black sheep, 61
 - negative numbers, 32
 - Kernel of a homomorphism, 235
 - Klein 4-group, 121, 172, 205, 221, 229, 242
 - Kronecker δ symbol, 167, 254

-
- Lagrange's theorem, 227
 - Law of exponents, 110
 - integer power, 80
 - as morphism condition, 181
 - Least common multiple, 133, 141
 - Left cosets
 - definition of, 225
 - and homomorphisms, 238
 - number of, 227
 - product of, 231
 - Librarian's nightmare, 201
 - Logarithmic spiral, 119
 - Mapping
 - definition of, 47
 - image of a set under, 47
 - induced, 63
 - level set of, 69
 - preimage of a set under, 48, 65
 - Mathematical induction, 71
 - principle of, 72
 - Matrix
 - complex, 176
 - conjugate, 177
 - determinant of, 163
 - diagonal, 175
 - invertibility, 163
 - invertibility of, 255
 - multiplication, 166
 - orthogonal, 172, 175, 254
 - scalar, 174
 - transpose of, 166
 - unitary, 177
 - Morphism condition, 179, 235
 - commutative diagram for, 95
 - Natural numbers, 36, 71
 - Normal subgroup, 230
 - n th roots of unity, 21, 120, 181
 - Order of a group element, 116, 185
 - Orthogonal group
 - Cayley table for, 257
 - composition rules, 256
 - Orthogonal matrix, 172, 175, 254
 - Partition of a set, 23
 - defined by an equivalence relation, 60
 - Pascal's triangle, 88
 - Pauli matrices, 176
 - Permutations
 - algorithm for multiplying, 196
 - commuting, 194
 - conjugacy, 198
 - disjoint cycle structure of, 192
 - order of, 195
 - parity of, 202
 - versus order of, 204
 - of vertices of an n -gon, 211
 - Pigeonhole Principle, 65
 - Playing cards, 84
 - Power of a group element, 110
 - Power set, 23
 - Preimage of a set, 48
 - Principle of mathematical induction, 72
 - Quaternion group, 213
 - as matrices, 178
 - quotients of, 241
 - Quaternions, 176
 - Quotient
 - group, 240
 - torus as, 250
 - mapping, 63
 - of a set by an equivalence relation, 62

- Residue class
 - definition of, 126
 - group of units, 146
 - automorphisms of, 183
 - generators of, 186
 - zero divisor, 149
- Reverse order law, 123
- Roots of unity, 21, 120, 181
- Russell's paradox, 14
- Set
 - complement of, 22
 - empty, 14
 - partition of, 23
 - by an equivalence relation, 60
 - subsets of, *see* Power set
- Sets
 - Cartesian product of, 46
 - difference of, 30
 - disjoint, 22
 - equality of, 14
 - intersection of, 22
 - subsets of, 14
 - union of, 22
- Sheep, 31
 - joke regarding, 61
- \sinh , *see* Hyperbolic trig functions
- Special relativity, 123, 189, 264
- Spiral
 - logarithmic, 119
- Subgroup
 - generated by a set, 115
 - index of, 228
 - normal, 230
- Symmetric group
 - Cayley's theorem, 207
 - definition of, 191
 - list of elements of S_4 , 214
- Symmetries
 - of Platonic solids, 216
 - of a regular n -gon, 209
 - of Rubik's cube, 219
- \tanh , *see* Hyperbolic trig functions
- Tower of Hanoi, 74
- Unitary matrix, 177