Abstract Algebra: An Inquiry-Based Approach Second Edition Supplemental Material

Jonathan K. Hodge
St. Edward's University

Steven Schlicker
Grand Valley State University

Ted Sundstrom
Grand Valley State University

Contents

Preface

Abstract Algebra: An Inquiry-Based Approach, Second Edition is a textbook that covers material in ring theory and group theory. To keep the keep the size (and cost) of the text manageable, some additional and reference material is offered in this document.

Supplemental Investigations

- **Investigation 34: RSA Encryption.** This investigation describes the RSA algorithm and assumes familiarity with modular congruence and prime numbers from Investigation 1. This investigation is referred to in Exercise 20 of Investigation 23 concerning Fermat's Little Theorem.
- **Investigation 35: Check Digits.** This investigation introduces the idea of check digits in several contexts and assumes familiarity with modular congruence (Investigation 1) and the dihedral groups (Investigation 21).
- **Investigation 36: Games: NIM and the 15 Puzzle.** This investigation applies group theory to develop a winning strategy in the game of NIM and to determine which 15 Puzzles are solvable. It assumes knowledge of groups (Investigation 17) and subgroups (Investigation 19), along with the symmetric groups (Investigation 22).
- **Investigation 37: Groups of Order 8 and 12: Semidirect Products of Groups.** In this investigation, we classify all groups of order 8, introduce semidirect products of groups, and then classify all groups of order 12. We assume familiarity with the earlier classification of groups of various orders (Investigation 26) and with products of groups (Investigation 25).

Appendices

- **Appendix A: Functions.** This appendix appeared in the first edition and provides background information for students on properties of functions. This material is a helpful reference for the study of homomorphisms and isomorphisms in Investigations 7, 13, 26, symmetries in Investigation 17, as well as permutations in Investigation 22 in the text.
- **Appendix B: Mathematical Induction and the Well-Ordering Principle.** Mathematical induction and the Well-Ordering Principle are used throughout *Abstract Algebra: An Inquiry-Based Approach, Second Edition*. This appendix can be used as a review of these important items, and also provides proofs of the equivalencies of the Well-Ordering Principle and the different flavors of mathematical induction for those who are interested.
- **Appendix C: Methods of Proof.** This is a new appendix that provides review material on different methods of proof other than induction, including direct proofs, proofs using logical equivalencies, proof by contradiction, and proofs using cases.
- **Appendix D: Proof that $R[x]$ is a Ring.** The formal proof that $R[x]$ is a ring when R is a ring is long and notationally complex. The details in the general case are omitted in Investigation 8 in the text, but are included in this appendix for those who want to see a complete proof.

- **Appendix E: The Cubic Formula.** This material was in the first edition as a supplement to the investigation on irreducible polynomials. A complete derivation of the cubic formula is presented here. This formula is useful for Exercise 11 in Investigation 11 and is also referenced in Investigation 33 related to solvability by radicals in the text.
- **Appendix F: The Fundamental Theorem of Algebra.** The Fundamental Theorem of Algebra is an important result regarding irreducible polynomials in Investigation 11. Since proofs of this theorem are not algebraic in nature, they don't usually appear in modern algebra texts. In this appendix we present what we believe is an accessible proof for the interested reader.
- **Appendix G: Complex Roots of Unity.** Complex roots of unity appear throughout Investigations 32 and 33 related to field extensions and Galois theory. Many students may already have a firm background in this topic. In this appendix we present a review of complex roots of unity for those who may benefit from one.

Please address any questions or comments to the authors.

Copyright

Investigation 34

RSA Encryption

Focus Questions

By the end of this investigation, you should be able to give precise and thorough answers to the questions listed below. You may want to keep these questions in mind to focus your thoughts as you complete the investigation.

- What is public key encryption, and what are some of its applications?
- How does RSA encryption work, and how is the security of RSA encryption related to prime factorization of integers?
- What mathematical results are necessary to establish the validity of RSA encryption, and how do these results follow from previously established properties of the integers?

Congruence and Modular Arithmetic

In this investigation, we will use "mod" notation in two distinct but related ways. As we have done before, we will write

$$a \equiv b \pmod{n}$$

when we want to specify a *relationship* between the integers a and b—namely, that n divides $(a-b)$, or equivalently, a and b have the same remainder when divided by n. This is the standard usage that we are familiar with from Investigation 2, except that we have omitted the parentheses that typically surround the "mod n" portion of the notation. This omission is common, and it will often make our notation easier to read, especially when we are working with expressions that already contain several sets of parentheses.

We will also sometimes write $a = b \pmod{n}$ (note the use of $=$ instead of $\equiv$) for the purposes of *defining* a to be the unique remainder guaranteed by the Division Algorithm when b is divided by n. This means that

$$a \equiv b \pmod{n} \text{ and } 0 \leq a < n.$$

For example, if $x = 56 \pmod{17}$, then since $56 = 17 \cdot 3 + 5$, we see that $x = 56 \pmod{17} = 5$.

Preview Activity 34.1. Determine the value of x for each of the following:

(a) $x = 29 \pmod{17}$.

(b) $x = 138 \pmod{17}$.

(c) $x = 200 \pmod{17}$.

We can also define a function using this mod operator. For a natural number n, we let $R_n = \{0, 1, 2, \ldots, n-1\}$ and define $f : \mathbb{Z} \to R_n$ so that for each integer x,

$$f(x) = x \pmod{n}.$$

This is sometimes referred to as the **mod *n*** function.

(d) For $n = 17$, compute $f(29)$, $f(138)$, and $f(200)$, and $f(546)$.

Introduction

Throughout history, secret codes have been used to send private messages from one person to another since in many situations, there is a desire for security against unauthorized interpretation of coded data. That is, there is a desire for secrecy. **Cryptography** is the science and art of concealing the content of communications between parties where the communications channel between them can be accessed by an unfriendly third party. Following are some standard terms used in modern cryptography,

- Plaintext. This is the original message that is in readable form.
- Encryption. The process of transforming the plaintext message into a disguised message that is then transmitted to another party.
- Encrypted message. This is the disguised message that is transmitted to another party.
- Decryption. The process of converting the encrypted message back to the original plaintext message.
- Decrypted message. The message that is converted from the encrypted message by the decryption process. The decrypted message should be identical with the original plaintext message.
- Cryptosystem. This term refers to the system that contains both the encryption process and the decryption process.

One of the first known cryptosystems is the so-called Caesar cipher, which was used by Julius Caesar to send messages to his troops. To encrypt a message, Caesar simply shifted each letter in the alphabet three places to the right. Using the English alphabet, this means that the letter A is encrypted as D, the letter B is encrypted as E, and so on. When we get to the end of the alphabet, we "wrap around" to the beginning of the alphabet. So X, Y, and Z are encrypted as A, B, and C, respectively. To decrypt a message, we simply shift each letter in the encrypted message 3 letters to the left.

Although it may not be necessary for this cryptosystem, we will illustrate the modern practice of converting text into numerical information. We will encode each letter with the numbers shown in the following table.

A	B	C	D	E	F	G	H	I	J	K	L	M
00	01	02	03	04	05	06	07	08	09	10	11	12

N	O	P	Q	R	S	T	U	V	W	X	Y	Z
13	14	15	16	17	18	19	20	21	22	23	24	25

So a plaintext message such as SAY would be encoded as 180024, where it is understood that each block of two numbers corresponds to a letter. We can then use the mod 26 function f to encrypt this message. So

$$f(x) = x + 3 \pmod{26},$$

where x is a two-digit representation of a letter.

Activity 34.2.

(a) Use the encryption function f to encrypt the message 180024.

(b) Suppose you receive an encrypted message 16070022060301 that you know was encrypted using the Caesar cipher. Decrypt this message and obtain the plaintext message. Describe the decryption function that you used to do this.

(c) If somebody knows how to encrypt a message using the Caesar cipher, do they then know how to decrypt a message using the Caesar cipher? Explain how this is a weakness in the cryptosystem.

The Caesar cipher used a shift of 3 letters to the right. This can be generalized to the concept of a **shift cipher**, which will be explored in Exercise (1).

RSA Encryption

Preview Activity 34.3. In Investigation 1, we learned about divisibility, greatest common divisors, and prime factorization in the integers. In this investigation, we will use these ideas together to study an interesting and important application: public-key encryption. The system we will study is one that is commonly used to transmit sensitive data over the internet. Its security rests on an important observation that should become apparent as you attempt to complete the following tasks.

(a) For each of the parts below, find a number whose prime factors are exactly the numbers listed:

(i) 4861 and 2621

(ii) 7907 and 619

(iii) 1753 and 1759

(b) Each of the numbers below has exactly two prime factors. Using any mathematically correct method, find these two factors.

(i) 13494211

(ii) 3902233

(iii) 1776977

(c) Which was easier: part (a) or part (b)? Explain why.

From the primitive Caesar cipher, which simply shifts each letter in the original message by a fixed amount, to the fascinating Enigma machines used by the Germans during World War II, numerous encryption schemes and devices have been invented in an attempt to keep sensitive data out of the hands of unauthorized (and potentially malicious!) third parties. In this investigation, we will study one of the most commonly used modern encryption schemes, named *RSA encryption* for Ron Rivest, Adi Shamir, and Leonard Adleman, who publicly described the system in 1977. An equivalent system was secretly developed in 1973 by Clifford Cocks at the Government Communications Headquarters (GCHQ), an intelligence and security organization responsible for providing signals intelligence and information assurance to the government and armed forces of the United Kingdom. This work was declassified in 1997.

Since the late 1970s, RSA encryption has become one of the most important systems for what is now known as *public-key encryption*. The basic idea behind a public-key scheme is that anyone should be able to send a message, but only the intended recipient should be able to read it. Thus, the key to *encrypt* a message is made public, but the key to *decrypt* the message is kept secret and distributed only to those who are authorized to view the encrypted text. In order for such a system to work, it has to be very difficult or even impossible for a potential attacker to determine the *private* decryption key just by knowing the *public* encryption key. Compare this to the Caesar cipher (or any shift cipher) where knowledge of the encryption key makes it easy to determine the decryption key.

Public-key encryption is particularly useful for tasks such as sending data over the Internet. For instance, a banking web site might want to allow any user to send information (such as a user name and password) over a secure connection. However, for the transaction to be truly secure, only the bank should be able to decode the information sent. RSA encryption achieves this design feature by using the properties of prime factorizations suggested in Preview Activity 34.3. In particular, RSA encryption exploits the fact that it is relatively easy to multiply two prime numbers together (even if they are large primes), but nearly impossible to efficiently factor a large number into its prime factors. As we will see shortly, the RSA scheme translates this theoretical fact into a very practical and secure encryption method.

The Basics of RSA Encryption

The first step in any encryption scheme is to decide what *alphabet* will be used and how the elements of the alphabet will be assigned numerical representations. We will use the same method of assignment used in the Caesar and shift ciphers. So we will use the standard A through Z alphabet, with A represented by 0, B represented by 1, and so on. As we will see, however, there are many security advantages to using a larger alphabet that consists not only of letters, but entire words. So, for instance, we might use a two-digit numerical representation of each letter (00 to 25) and encode our messages using blocks of letters. If we were to do so, the word "SECRET" would be encoded as 180402170419. Keep in mind that at this point, we haven't actually encrypted anything. We have simply developed a way of translating letters or blocks of letters into the numerical representations that will be used by our encrypting function. Note that we could have also included numerical codes for spaces, numbers, punctuation, and the like.

Once we have established our alphabet, it takes only a few simple steps to encrypt a message:

(1) First, we generate two different prime numbers, p and q, and calculate the quantity $m = pq$ (called the *modulus*). For the resulting system to be secure, p and q need to be extremely large, perhaps having hundreds or even thousands of digits. Because of the size of the primes required, the two prime numbers are generated by a computer. * The value of m becomes public—it can be shared with anyone—but p and q must be kept secret. In practice, modern computers can easily generate prime numbers with several hundred digits. However, in that case, the value of m must be factored to determine p and q and this can take several hundred years to do so using current technology.

(2) Next, we obtain a positive integer e such that

$$\gcd(e, (p-1)(q-1)) = 1.$$

This number e is called the *encryption key*, and the value of e is made public. The quantity $(p-1)(q-1)$ is often called the *totient*, denoted t. The totient must be kept secret, as it plays an essential role in the decrypting process.

(3) Finally, to encrypt a message, we form blocks of letters of a specified size and input the numerical representation of each block of letters into the encoding function f defined by

$$f(x) = x^e \pmod{m}.$$

Once we have encrypted a message, the decryption process is similar and can be described as follows:

(1) First, we find a positive integer d (the *decryption key*) such that

$$ed \equiv 1 \pmod{(p-1)(q-1)}.$$

(2) To decrypt an encoded message, we input each block of data into the decoding function g defined by

$$g(x) = x^d \pmod{m}.$$

It should be noted that not all of the steps described above are straightforward or easy to complete. For instance, the decryption key is defined to be an integer that satisfies a particular congruence relation. It is not immediately obvious that such an integer will always exist. We will have to use what we have learned about the integers to prove not only that a suitable decryption key exists, but also that it can be found in a relatively straightforward manner, and that the corresponding decoding function actually returns encrypted messages to their original, unencrypted state.

Two Examples

Before we go any further into investigating the details of RSA encryption and why it works, we will examine two different examples. For the first example, most of the computations can be done using

*As we will see later on, in order for RSA encryption to work, p and q each need to be larger than the number of elements in the given alphabet. This is usually not a problem since the security of RSA encryption relies on choosing primes that are huge, and certainly larger than the size of any reasonable alphabet.

a calculator but it would be easier to use a computer algebra system. A computer algebra system was used to do the computations in the second example. It is only necessary to do one of the examples.

Example 34.4. For the first example, we will use very small prime numbers so that some of the calculations can be done using a calculator. Namely, the calculations for the encryption process can be done with a calculator, but something like a computer algebra system is required to do the computations for the decryption process.

So we will use $p = 29$ and $q = 41$. So the modulus m and totient t are

$$\begin{aligned} m &= pq = 1189 \\ t &= (p-1)(q-1) = 1120 \end{aligned}$$

(This example is only for illustration and would not be used since the values of p and q could be easily determined from $m = 1189$.)

Our encryption key e must be chosen so that $\gcd(e, t) = 1$. We will use $e = 3$. So the modulus $m = 1189$ and encryption key $e = 3$ are made public so that anyone can send a message to this receiver. Suppose someone wants to sent the message "DOG". The numerical code for this is 031406. We now divide this into blocks of digits. Because of the size of m, we will use only two-digit blocks since a 4 digit block could be greater that m. (Note: Three digit blocks could also be used.) To encrypt the message, we use the encryption function f where

$$f(x) = x^e \pmod{m} = x^3 \pmod{1189}.$$

For the message 031406, we have

$$\begin{aligned} f(03) &= 03^3 \pmod{1189} = 27 \\ f(14) &= 14^3 \pmod{1189} = 366 \\ f(06) &= 06^3 \pmod{1189} = 216 \end{aligned}$$

To decrypt this message, the receiver must now determine d so that $ed \equiv 1 \pmod{(p-1)(q-1)}$ or $3d \equiv 1 \pmod{1120}$. To do this, we use the Euclidean Algorithm to determine the coefficients for $1 = 3r + 1120s$ according to Bezout's Identity (Theorem 1.17 on page 13.) We will omit the details, but it can be verified that

$$3(-373) + 1120 = 1$$

and so $3(-373) \equiv 1 \pmod{1120}$. By adding $3 \cdot 1120$ to both sides of this congruence, we obtain

$$\begin{aligned} 3(-373) + 3(1120) &\equiv 1 + 3(1120) \pmod{1120} \\ 3(747) &\equiv 1 \pmod{1120} \end{aligned}$$

and so $d = 747$. So the user must now decrypt each of these blocks using the decryption function g where

$$g(x) = x^d \pmod{m} = x^{747} \pmod{1189}.$$

For our three blocks, we use a computer algebra system to obtain

$$\begin{aligned} g(27) &= 27^{747} \pmod{1189} = 3 \\ g(366) &= 366^{747} \pmod{1189} = 14 \\ g(216) &= 216^{747} \pmod{1189} = 6 \end{aligned}$$

So the numerical decrypted message is 031406, which is the plaintext message DOG.

Example 34.5. All computations in this example were done using a computer algebra system such as Maple, Mathematica, or Sage. Suppose we want to use RSA encryption to encode the highly sensitive, top-secret message, "JOHNNY LOVES SALLY". We will begin by choosing p and q. In practice, p and q are often hundreds of digits long, but we will choose two smaller primes,

$$p = 400043344212007458013 \text{ and } q = 500030066366269001203.$$

With these choices of p and q, our modulus m and totient t are

$$m = pq = 200033699955714283345172521584008468989639$$

and

$$t = (p-1)(q-1) = 200033699955714283344272448173430192530424.$$

Recall that m will be made public, but p, q, and t will be kept secret. This is significant, since in order to calculate t from m, we would have to first factor m, a task that a computer could fairly easily complete for this example, but not for examples involving larger primes. In fact, one website on RSA cryptography notes that "if p and q are each 1024 bits long, the sun will burn out before the most powerful computers presently in existence can factor the modulus into p and q." †

The next step in encoding our message is to choose an encryption key. We need to choose a number e such that $\gcd(e, (p-1)(q-1)) = 1$. Note that any prime number that does not divide $t = (p-1)(q-1)$ will suffice here; a common choice is $e = 2^{16} + 1 = 65537$, and this is what we will use.

To perform the actual encryption, we will first break our message into three 6-character blocks. We will use the standard 00 – 25 encoding for the letters A – Z, and we will also use the code 99 to denote a space. Thus, the numerical representation of our message is:

$$091407131324 \;\; 991114210418 \;\; 991800111124$$

We then apply our encoding function to each of these 12-digit numbers:

$$\begin{aligned} f(091407131324) &= (091407131324)^e \pmod{m} \\ &= 009505729493564929202343371764084584555016 \\ f(991114210418) &= (991114210418)^e \pmod{m} \\ &= 012196119237767316793050190360104919489384 \\ f(991800111124) &= (991800111124)^e \pmod{m} \\ &= 124080637343749317837866219863773135637684 \end{aligned}$$

Note that we have appended zeros to the beginning of each encoded block so that each has exactly 42 digits. This would allow the entire message to be sent as a single string and then unambiguously decomposed into its three distinct parts prior to decryption.

To decrypt the message, we would first need to find the decryption key, d. It is at this step that some of the theory we have been studying in previous investigations is particularly useful (and in fact necessary). For now, however, we will skip over the "how" and simply assume that we have been able to find an integer, say

$$d = 921894177863251936178098635065731653140 81,$$

such that

$$ed \equiv 1 \pmod{(p-1)(q-1)}.$$

† `http://fringe.davesource.com/Fringe/Crypt/RSA/Algorithm.html`

A computer algebra system can then readily verify that

$$\begin{aligned} &g(009505729493564929202343371764084584555016) \\ &\qquad = (009505729493564929202343371764084584555016)^d \pmod m \\ &\qquad = 091407131324, \end{aligned}$$

$$\begin{aligned} &g(012196119237767316793050190360104919489384) \\ &\qquad = (012196119237767316793050190360104919489384)^d \pmod m \\ &\qquad = 991114210418, \end{aligned}$$

and

$$\begin{aligned} &g(124080637343749317837866219863773135637684) \\ &\qquad = (124080637343749317837866219863773135637684)^d \pmod m \\ &\qquad = 991800111124. \end{aligned}$$

In other words, the decoding function g returns each block of the encrypted message to its original, unencrypted state, as desired. It is interesting to note that, even in this example, the computations in the decryption process involve raising one 40-digit number to another 40-digit exponent. Fortunately, there are efficient algorithms for performing such exponentiations, even when the numbers involved are much larger (as would be the case if larger, and hence more realistic, values of p and q were chosen.)

Why RSA Decryption Works

Now that we've seen an example, we are ready to get to work and show that RSA encryption actually works the way it is intended. In particular, we must show three important facts:

- First, we must show that no matter what primes are used for p and q, it will always be possible to find an encryption key, e, that satisfies
$$\gcd(e, (p-1)(q-1)) = 1.$$
- Next, we must show that it is always possible to find a decryption key, d, such that
$$ed \equiv 1 \pmod{(p-1)(q-1)}.$$
- Finally, we must show that the decoding function $g(x) = x^d \pmod m$ is the inverse of the encoding function $f(x) = x^e \pmod m$. In other words, we must show that for all x in our alphabet,
$$(x^e)^d \pmod m = x^{ed} \pmod m = x.$$

The goal for the rest of this investigation is to establish these three facts, and the activities below suggest a series of steps that will accomplish exactly that goal.

Task 1: Finding the Encryption Key

Activity 34.6. Let p and q be any prime numbers.

(a) Explain why it is always possible to find a prime number e that does not divide $(p-1)(q-1)$.

(b) Explain why the number e found in part (a) would always satisfy

$$\gcd(e, (p-1)(q-1)) = 1.$$

Task 2: Finding the Decryption Key

Activity 34.7. Consider the fact that the encryption key e is chosen specifically so that

$$\gcd(e, (p-1)(q-1)) = 1.$$

(a) Use Bezout's Identity (Theorem 1.17 on page 13) to write down a linear combination corresponding to $\gcd(e, (p-1)(q-1))$.

(b) Use your answer to part (a) to explain why there must exist an integer d such that

$$ed \equiv 1 \pmod{(p-1)(q-1)}.$$

(c) What process or algorithm would allow you to actually determine the value of d that is guaranteed to exist by part (b)? (Hint: We have studied this algorithm in a previous investigation.)

(d) Suppose that we were able to find an integer d' such that

$$ed' \equiv 1 \pmod{(p-1)(q-1)},$$

but $d' < 0$. Explain how we could use d' to find a positive integer d that also satisfies

$$ed \equiv 1 \pmod{(p-1)(q-1)}.$$

(Hint: Add a convenient quantity to d'.)

Task 3: Proving an Inverse Relationship Between f and g

In order to prove that the decoding function g actually undoes the work of the encoding function f, we must show that for all x in our alphabet,

$$x^{ed} \pmod{m} = x.$$

Doing so will require the following three intermediate results:

Theorem 34.8 (Binomial Theorem)**.** *Let n be a positive integer, and let a and b be any real numbers. Then*

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

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Theorem 34.9 (Freshman's Dream)**.** *Let p be a prime number. Then for all integers a and b,*

$$(a+b)^p \equiv (a^p + b^p) \bmod p.$$

Theorem 34.10 (Fermat's Little Theorem)**.** *Let p be a prime number. Then for every positive integer a,*

$$a^p \equiv a \bmod p.$$

Proofs of the Binomial Theorem and the Freshman's Dream are outlined in Exercises 10 and 11 of Investigation 5. (See page 81.) The proof of Fermat's Little Theorem follows from the Freshman's Dream. (If you have completed Investigation 23, an alternative proof using group theory is suggested in Exercise (20) on page 330.)

Activity 34.11. Use induction on a, along with the Freshman's Dream, to prove Fermat's Little Theorem.

Activity 34.12. Explain why the conclusion of Fermat's Little Theorem (namely, that $a^p \equiv a \pmod{p}$) is equivalent to

$$a^{p-1} \equiv 1 \pmod{p}$$

as long as $p \nmid a$.

We can now use the Freshman's Dream and Fermat's Little Theorem to establish an inverse relationship between the RSA encoding and decoding functions.

Activity 34.13. Let p, q, m, d, and e be as stated previously. Note that, by definition,

$$ed \equiv 1 \pmod{(p-1)(q-1)}.$$

(a) Use the definition of congruence to explain why $ed \equiv 1 \pmod{(p-1)}$ and $ed \equiv 1 \pmod{(q-1)}$.

(b) Use part (a), Activity 34.12, and the fact that p and q were chosen to be very large (and, in particular, much larger than the number of letters in the alphabet, $\mathbb{A}$) in order to prove that

$$x^{ed} \equiv x \pmod{p} \text{ and } x^{ed} \equiv x \pmod{q}$$

for all $x \in \mathbb{A}$.

(c) Use part (b) to explain why for all $x \in \mathbb{A}$, $x^{ed} \equiv x \pmod{m}$.

(d) Explain how your answer to part (c) actually implies that $x^{ed} \pmod{m} = x$. (Hint: It again matters that p and q, and thus m, are larger than the number of elements in $\mathbb{A}$. Remember that if two numbers are both less than m and yet congruent modulo m, then they must be equal.)

(e) Deduce from your answer to part (d) that the decoding function used in RSA encryption always returns an encrypted message to its original, unencrypted state.

Concluding Thoughts and Notes

Before we conclude our investigations of RSA encryption, a few additional observations are worth mentioning.

- Since p and q are chosen to be very large, it is important that e be large enough so that x^e is typically greater than the modulus, $m = pq$. If x^e is not greater than m, then messages can be easily decrypted by simply taking the e^{th} root of the encrypted data, since the encryption function in this case does not involve a reduction modulo m.

- RSA encryption schemes are deterministic, meaning that they have no random component to them. Because of this, potential attackers could use the public encryption key to develop a dictionary of likely words and their encryptions. This dictionary could then be used to try to decipher encrypted messages by comparing the encrypted words to the entries in the dictionary. Encrypting larger blocks of data (instead of individual letters) reduces this security vulnerability.

- One aspect of public key encryption that we have not considered is that of *signing* or *authentication*. Since RSA schemes enable anyone to encrypt a message, it is important to be able to verify that encrypted messages are actually from who they claim to be from. Several methods are available for this purpose, many of which involve introducing an additional private key that identifies the sender. See Exercise (7).

- Most experts agree that with large enough primes, RSA encryption seems to be secure for the near future. That is, even though the values of e and m are made public, the value of the decryption key d cannot really be determined from e and m. However, advances in computer science, and especially in the field of quantum computing, have the potential to solve computational problems such as integer factorization much faster than present-day computers. This would then render RSA encryption obsolete since the decryption key could be determined from e and m.

- Interest in RSA was primarily confined to mathematicians, computer scientists, and cryptography hobbyists until the invention of the World Wide Web. Beginning in the 1990s, there has been an explosion of online commerce. Some form of encryption was needed in order to send credit card numbers and other sensitive information over the Internet. Now, if you log into any secure web server, there is a good chance your computer has the server's public key and used it to secure the information you have sent. However, this information is often not encrypted directly using RSA. This is due to the fact that the RSA system is a so-called asymmetric cryptosystem. This simply means that there are two keys, one for encryption and one for decryption. A symmetric cryptosystem uses the same key for encryption and decryption and is usually much faster and uses fewer resources than an asymmetric system.

  Quite often, a file will be encrypted with a symmetric-key algorithm, and the symmetric key will be encrypted with RSA encryption. Under this process, only an entity that has access to the RSA private key will be able to decrypt the symmetric key. Without being able to access the symmetric key, the original file cannot be decrypted. This method can be used to keep messages and files secure, without taking too long or consuming too many computational resources.

  One widely used symmetric system is the Advanced Encryption Standard or AES. This was developed in the late 1990s by the National Institute of Standards and Technology (NIST) as the new government cipher standard. It is expected that at some time, AES will need to be replaced. The original expectation was that the standard should last 30 years, and NIST is supposed to review it every five years. So far, no problems have arisen with AES, but cryptographers are working on what might succeed AES to be ready just in case problems arise.

Exercises

(1) Caesar's's cipher is an example of a **shift cipher**, in that it encrypts messages by simply shifting each letter in the message by a fixed amount. (For example A $\rightarrow$ E, B $\rightarrow$ F, etc.) The encoding functions associated with shift ciphers always have the form

$$f(x) = x + a \pmod{n},$$

where n is the number of letters in the alphabet (typically 26).

(a) Assuming that the message below was encrypted using a shift cipher, decrypt the message. Make sure you explain how you determined which shift cipher was used.

```
XLMW QIWWEKI AEW IRGVCTXIH YWMRK E WLMJX GMTLIV.
```

(b) Was the following message encrypted using a shift cipher? Why or why not?

```
AJMZ YBZZGLB EGZ QCA BQPSOUABD
KZMQL G ZJMHA PMUJBS.
```

(2) In contrast to a shift cipher (see Exercise (1)), a **stretch cipher** uses multiplication instead of addition to encode messages. That is, instead of using an encoding function of the form

$$f(x) = x + a \pmod{26},$$

it uses one of the form

$$f(x) = ax \pmod{26},$$

where a is some integer.

(a) Use a stretch cipher of your choosing to encode this message: ABSTRACT ALGEBRA MAKES ME SMILE.

(b) Does the stretch cipher you used in part (a) have a corresponding decoding function? In other words, is there a rule that can be used to decode any message encoded by the cipher?no

(c) Is it always possible to decode a message that has been encoded using a stretch cipher? If so, explain how. Otherwise, determine the values of a (assuming a 26 letter alphabet) for which the corresponding stretch cipher is decode-able.

(3) The frequency with which each letter in the alphabet occurs in ordinary English is given in Table 34.1. [‡] Explain in a precise way how this table could be used to break shift and stretch ciphers.

(4) In Exercise (3) , we saw how analyzing the frequency of letters in a message encrypted with a shift or stretch cipher could help someone break such a cipher. So people began devising encryption methods in which a given letter that appears more than once in a message would be encrypted differently depending on its location in the message. For example, there are four A's in the plaintext message is ALABAMA, and we know that the Caesar cipher would encrypt each of these A's with the letter D.

[‡] This table originally appeared in *Applications of Abstract Algebra with Maple and MATLAB* (2nd ed.) by Klima, Stitzinger, and Sigmon, CRC Press, 2006.

Letter	Frequency (%)	Letter	Frequency (%)
A	8.167	N	6.749
B	1.492	O	7.507
C	2.782	P	1.929
D	4.253	Q	0.095
E	12.702	R	5.987
F	2.228	S	6.327
G	2.015	T	9.056
H	6.094	U	2.758
I	6.966	V	0.978
J	0.153	W	2.360
K	0.772	X	0.150
L	4.025	Y	1.974
M	2.406	Z	0.074

Table 34.1
Frequency of each letter in the English language

The Vigenére cipher is a method of encryption that uses a series of interwoven shift ciphers based on a keyword. For example, if the keyword is GOLF, then we determine the keyword number sequence, which is 6 – 14 – 11 – 5. This means that the first letter in the plaintext message will be shifted by 6, then second letter will be shifted by 14, the third letter will be shifted by 11, and the fourth letter will be shifted by 5. When we have used all of the letters in the keyword, we start over at the beginning of the keyword. So the fifth letter in the plaintext message will be shifted by 6, the sixth letter will be shifted by 14, the seventh letter will be shifted by 11, and the eighth letter will be shifted by 5.

We still follow the basic rule that if a plaintext letter corresponds to the number c, and we shift s places, then the encrypted letter will correspond to the number

$$f(c) = (c + s) \pmod{26}.$$

The numerical code for the plaintext message ALABAMA is 00110001001200.

(a) Use the keyword GOLF with a Vigenére cipher to encrypt the message ALABAMA, first as a numerical sequence and then using letters.

(b) The plaintext message had 4 A's. Were these four A's encrypted with same letter in the encrypted message?

(c) Explain how the receiver would decrypt this message.

(5) **Hill ciphers** use matrices to encode and decode messages. For instance, using the 2×2 matrix

$$A = \begin{bmatrix} 2 & 5 \\ 1 & 4 \end{bmatrix},$$

the message "ATTACK" would be encrypted by multiplying A by the vector representation of each pair of consecutive letters. Doing so, we obtain

$$A\begin{bmatrix}0\\19\end{bmatrix}=\begin{bmatrix}95\\76\end{bmatrix},\quad A\begin{bmatrix}19\\0\end{bmatrix}=\begin{bmatrix}38\\19\end{bmatrix},\text{ and }A\begin{bmatrix}2\\10\end{bmatrix}=\begin{bmatrix}54\\42\end{bmatrix}.$$

(a) Reduce these vectors modulo 26 in order to finish encrypting the message.

(b) Are the two A's in the original message encrypted to the same letter? What about the two T's?

(c) Are Hill ciphers more or less susceptible to frequency analysis (that is, the analysis suggested in Exercise (3)) than shift and stretch ciphers? Clearly explain your answer.

(d) What conditions must be placed on the encrypting matrix in order to guarantee that the resulting Hill cipher will be decode-able? Give a convincing argument to justify your answer.

(e) Use a Hill cipher with a matrix different than A to encrypt the message, "CRYPTOGRAPHY IS FUN". Then find the corresponding decrypting matrix, and verify that it does in fact return the encrypted message to its original form.

(6) Are any of the systems mentioned in Exercises 1 – 5 public-key schemes? That is, is it ever possible to encode messages using one of these schemes without also knowing how to decode messages?

(7) Suppose that person B wants to send a message to person A in such a way that A knows it is from B. To do this, person B needs a digital signature. So for this, we will assume that A has made public their encryption key consisting of n_A and e_A. Person B also has a public encryption key consisting of n_B and e_B. Of course, both of them have their own decryption keys: d_A for person A and d_B for person B.

Explain how B can use their decryption key d_B to "sign" the message to A so that A knows that the message is actually from person B.

Connections

Pure mathematics can be described as the study of mathematical concepts independently of any application outside of mathematics. Mathematicians have had differing opinions as to what constitutes pure versus applied mathematics. One of the more famous (at least among mathematicians) examples of this debate can be found in the essay from 1940 called "A Mathematician's Apology"' by the well-known British mathematician G.H. Hardy. Hardy preferred pure mathematics, which he often compared to painting and poetry, but he argued that the distinction was that applied mathematics sought to express physical truth in a mathematical framework, whereas pure mathematics expressed truths that were independent of the physical world.

In this investigation, we considered one of the applications of number theory and abstract algebra that is very important in our digital world. We saw how congruence and modular arithmetic from Investigation 2, along with Bezout's Identity from Investigation 1 and Fermat's Little Theorem (see Exercises (20) and (24) in Investigation 23) can be used to encrypt data in such a way that unauthorized people cannot decrypt the data. This allows for secure electronic transmission of information.

Investigation 35

Check Digits

Focus Questions

By the end of this investigation, you should be able to give precise and thorough answers to the questions listed below. You may want to keep these questions in mind to focus your thoughts as you complete the investigation.

- What are check digits, and how are they used?
- What are some common check digit schemes for credit card and ISBN numbers?
- What is Verhoeff's check digit scheme, and how is it related to the dihedral group of order 10?

Preview Activity 35.1. This investigation will involve both congruence of integers and the dihedral group D_5. In this activity, we will review a few of the basics.

(a) Determine the value of x that satisfies the congruence equation

$$2(1) + 3(2) + 4(2) + 5(6) + 6x \equiv 0 \pmod{7}.$$

(b) Consider the permutation $\pi = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 5 & 7 & 6 & 2 & 8 & 3 & 0 & 9 & 4 \end{pmatrix}$ in S_{10}. (Note that we are permuting the digits 0 through 9 instead of 1 through 10.)

(i) Explain why π is an even permutation.

(ii) What is the order of π?

(iii) Find $\pi^4(5)$, where π^4 denotes the composition of 4 copies of π (that is, $\pi \circ \pi \circ \pi \circ \pi$).

(iv) If x is any integer between 0 and 9, what is $\pi^8(x)$? Explain.

Introduction

In this investigation, we will discuss check digits and explore various check digit schemes, including one that uses the dihedral group D_5. Check digits are important because nowadays most information is transmitted electronically. When we use an ATM or pay with a credit card, there is always

the possibility that an error in transmission can occur. For instance, noise can be introduced in a message, information can be lost, and data can be altered during transmission. Another important source of error is human error. Data can be confused when humans enter numbers into machines or communicate to others. Some mistakes are more prevalent than others. For example, according to Richard Hamming,* the two most common human errors when dealing with information are interchanging digits (e.g., typing 12 instead of 21) and changing one of a string of three digits when two adjacent digits are the same (e.g., using 112 instead of 122). Clearly, problems can arise when data is not encoded or transmitted properly. Fortunately, there are ways to compensate for these errors. The first step is to determine when they occur, and that is where check digits come into play.

Check Digits

A check digit is a digit appended to a string, usually at the end, to make the sum of the digits in that string congruent to a specific number modulo a given integer. For example, the 10 digit identification number 2361068754 might have an extra digit d appended to the end so that the digit sum is congruent to 0 modulo 9. In this case, the check digit would be 3 and the ID number would be 23610687543. As their name suggests, check digits perform a check to ensure that the number received is a valid number. However, just because the number appears to be valid, that does not necessarily make it legitimate. For instance, a credit card company may choose to only use a small subset of the set of valid credit card numbers. We should also note that even when check digits are used to detect errors, they do not necessarily provide a way to correct the errors they find. (There are other methods for doing that.)

Check digits are used in UPC codes, credit card numbers, ISBN numbers, and most other identification systems. We will now consider a few of the more common and/or mathematically interesting check digit schemes.

Credit Card Check Digits

Different credit cards have account numbers, or ID codes, of different lengths and with different prefixes. Each code consists of a string of numbers, with each digit between 0 and 9. The prefix of a card is the one or two digit block at the beginning (the leftmost digits) of the ID code. In particular:

- MasterCard codes have 16 digits and use prefixes of 51, 52, 53, 54, and 55.
- VISA codes have either 13 or 16 digits and use a prefix of 4.
- American Express codes have 15 digits and use prefixes 34 or 37.
- Discover codes have 16 digits and use a prefix of 6011.

The prefix is the Major Industry Identifier (MII) and indicates the type of industry that issues the card. For example, VISA and MasterCard are issued by the banking and financial sector (with prefix

**Coding and Information Theory* (2nd ed.), Prentice-Hall, 1986, p. 27.

numbers 4 or 5) while American Express is in the travel and entertainment category (prefix number 3). All credit cards compute a check digit modulo 10.[†] To find the check digit, we can use a process known as the *Luhn algorithm*:[‡]

(1) Beginning with the second digit from the right (in other words, don't include the check digit) and moving from right to left, double every other digit. Add the individual *digits* of these numbers (e.g., if the doubled number is 16, add 1 and 6).

(2) Sum the digits (but not the check digit) not considered in step (1).

(3) Add the results of steps (1) and (2). Call this result s.

(4) The check digit d is the solution to $s + d \equiv 0 \pmod{10}$.

Activity 35.2.

(a) Find the correct check digit d for the sample VISA card with number 4417 1234 5678 911d.

(b) Create your own valid American Express card number.

ISBN Check Digits

The acronym ISBN is an abbreviation for the International Standard Book Number, which is used to identify books. An ISBN-10 has the form

$$X_1X_2X_3X_4X_5X_6X_7X_8X_9X_{10},$$

where each X_i is a digit between 0 and 9. In an ISBN, the first digit, X_1, represents the language of the book (0 is English), the next block (2 or 3 digits) identifies the publisher, the third block (5 or 6 digits) is a publisher's number for the book, and the last digit is a check digit.

The check digit in an ISBN is determined by first attaching a weight to each digit, with the leftmost digit (X_1) having a weight of 1, the next digit (X_2) having a weight of 2, and so on. (In general, the weight of the digit X_k is k.) Next, we multiply each digit (except the check digit) by its weight and compute the weighted sum. The check digit is congruent to the weighted sum modulo 11, with X representing a check digit of 10.

A quick way to implement this scheme is through the use of weight vectors and dot products. Recall that the dot product of vectors $[v_1, v_2, \ldots, v_n]$ and $[w_1, w_2, \ldots, w_n]$ is the scalar

$$[w_1, w_2, \ldots, w_n] \cdot [v_1, v_2, \ldots, v_n] = \sum_{i=1}^{n} w_i v_i.$$

By taking the dot product of the first 9 digits of an ISBN with the weight vector $[1, 2, 3, 4, 5, 6, 7, 8, 9]$, we can easily determine what the check digit should be.

Activity 35.3. The ISBN-10 for a very interesting book is 0-471-33193-?. Find the check digit.

[†]Some books and web sites state that MasterCard and VISA use the prefix digits when determining the check digit, while American Express and Discover do not. To the best of our knowledge, all companies use the prefix digits in their calculations, and so we will do the same.

[‡]The Luhn Algorithm was created by Hans Peter Luhn, who worked for IBM. It is patented in U.S. Patent No. 2.950.048.

This ISBN-10 check digit scheme will find all single digit errors, but will also catch errors obtained by interchanging digits (for example, typing 12 instead of 21). However, the ISBN-10 scheme is restricted to ID numbers with 10 digits, and we have to introduce the extra symbol X to represent the digit 10 as a possible check digit. In the next section, we will examine a check digit scheme that works for ID numbers with any number of digits.

Verhoeff's Dihedral Group D_5 Check

In the late 1960s, Dutch mathematician Jacobus Verhoeff proposed a check digit scheme based on the dihedral group D_5.[§] This scheme is an improvement on others in that it works for any length number, and it detects all single digit errors and all transposition errors involving two adjacent digits. However, the Verhoeff scheme is a little more complicated to implement.

We begin with the operation table for D_5 given in Table 35.1. (Note that the elements in this table are listed in a different order than usual; this is done to match Verhoeff's labeling.)

	I	R	R^2	R^3	R^4	rR^4	rR^3	rR^2	rR	r
I	I	R	R^2	R^3	R^4	rR^4	rR^3	rR^2	rR	r
R	R	R^2	R^3	R^4	I	rR^3	rR^2	rR	r	rR^4
R^2	R^2	R^3	R^4	I	R	rR^2	rR	r	rR^4	rR^3
R^3	R^3	R^4	I	R	R^2	rR	r	rR^4	rR^3	rR^2
R^4	R^4	I	R	R^2	R^3	r	rR^4	rR^3	rR^2	rR
rR^4	rR^4	r	rR	rR^2	rR^3	I	R^4	R^3	R^2	R
rR^3	rR^3	rR^4	r	rR	rR^2	R	I	R^4	R^3	R^2
rR^2	rR^2	rR^3	rR^4	r	rR	R^2	R	I	R^4	R^3
rR	rR	rR^2	rR^3	rR^4	r	R^3	R^2	R	I	R^4
r	r	rR	rR^2	rR^3	rR^4	R^4	R^3	R^2	R	I

Table 35.1
Operation table for D_5.

We then replace the elements in the D_5 table with the digits 0 to 9 (keeping the elements in the same order as in Table 35.1) to obtain Table 35.2:

Verhoeff's check digit scheme requires an ID number of the form $a_{n-1}a_{n-2}\cdots a_1a_0$ (note that the digits are indexed from right to left, starting with an index of 0) to satisfy the equation

$$\pi^0(a_0)\cdot\pi^1(a_1)\cdot\pi^2(a_2)\cdots\pi^{n-1}(a_{n-1}) = 0,$$

where

$$\pi = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 5 & 7 & 6 & 2 & 8 & 3 & 0 & 9 & 4 \end{pmatrix}$$

[§] J. Verhoeff, "Error detecting decimal codes," *Mathematical Centre* Tract 29, The Mathematical Centre, Amsterdam, 1969.

$\cdot$	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
1	1	2	3	4	0	6	7	8	9	5
2	2	3	4	0	1	7	8	9	5	6
3	3	4	0	1	2	8	9	5	6	7
4	4	0	1	2	3	9	5	6	7	8
5	5	9	8	7	6	0	4	3	2	1
6	6	5	9	8	7	1	0	4	3	2
7	7	6	5	9	8	2	1	0	4	3
8	8	7	6	5	9	3	2	1	0	4
9	9	8	7	6	5	4	3	2	1	0

Table 35.2
Operation table for the Verhoeff check digit scheme.

is a permutation, $\pi^i = \pi \circ \pi \circ \cdots \circ \pi$ is the composition of i copies of π, and the $\cdot$ operation is that which arises from D_5 as indicated in Table 35.2. Note that the permutation applied to each digit depends on the position of the digit in the ID number. For example, if 4 is the digit in the third position from the right, then we apply π^2 to 4 to obtain 7. The result of applying the powers of π to any position can be described in the *permutation table* shown in Table 35.3, where the i^{th} row shows the result of applying the permutation π^i to each possible digit. Note that the powers of π are periodic, so the rows repeat after row 8. In other words, $\pi^{i+8}(k) = \pi^i(k)$ for all k.

	0	1	2	3	4	5	6	7	8	9
π^0	0	1	2	3	4	5	6	7	8	9
π^1	1	5	7	6	2	8	3	0	9	4
π^2	5	8	0	3	7	9	6	1	4	2
π^3	8	9	1	6	0	4	3	5	2	7
π^4	9	4	5	3	1	2	6	8	7	0
π^5	4	2	8	6	5	7	3	9	0	1
π^6	2	7	9	3	8	0	6	4	1	5
π^7	7	0	4	6	9	1	3	2	5	8

Table 35.3
Permutation table for the Verhoeff check digit scheme.

Now that we understand some of the mechanics involved, the Verhoeff scheme can be implemented as follows:

(1) Start with the n digit number

$$a_{n-1}a_{n-2}a_{n-3}\cdots a_1a_0,$$

with the digits labeled from right to left, starting with a_0.

(2) Let c denote the *checksum*, and set $c = 0$ initially.

(3) Step through the n-digit number digit by digit, each time replacing c with $c \cdot \pi^i(a_i)$ (where the operation $\cdot$ is indicated in Table 35.2).

The original number has a valid check digit if and only if, at the end of the process, the checksum c is equal to 0.

As an example, Table 35.4 shows how the steps described above can be followed to validate the checksum for the ID number 4134705. Since the final value of c is 0, we have a valid ID number.

i	a_i	$\pi^i(a_i)$	Old c	New c: Old $c \cdot \pi^i(a_i)$
0	5	5	0	5
1	0	1	5	9
2	7	1	9	8
3	4	0	8	8
4	3	3	8	5
5	1	2	5	8
6	4	8	8	0

Table 35.4
A Verhoeff check digit example.

The Verhoeff algorithm detects all single digit errors and all transposition errors made by interchanging two adjacent digits. Also, the Verhoeff algorithm detects over 95% of twin errors (where aa is changed to bb), over 94% of jump transpositions (abc replaced with cba) and jump twin errors (aca replaced with bcb), and most phonetic errors ($a0$ replaced with $1a$—for example, 40 replaced with 14; notice how these two numbers sound similar when read aloud). Gallian[¶] describes an interesting application in which the German government used a slight modification of the Verhoeff scheme to append check digits to serial numbers on their banknotes.

Activity 35.4. How do we find the check digit for an ID number using the Verhoeff scheme? Consider the ID number $1023857d$, where d is the check digit.

(a) Determine the value of the checksum c for the related ID number 10238570.

(b) Let $d = c^{-1}$. Show that $1023857d$ is a valid ID number. (As it turns out, this technique will always yield a correct check digit. Exercise (5) asks you to prove this result.)

Concluding Activities

Activity 35.5. Since 2007, books have been identified with 13 digit ISBNs (the ISBN-13). The first 12 digits of the ISBN-13 contain the identifying code and the 13th digit is the check digit. The

[¶] *Contemporary Abstract Algebra* (5th ed.), Houghton Mifflin Company, 2002.

check digit scheme is similar to the one used in ISBN-10 IDs. Research the ISBN-13 check digit scheme and explain how it works. Be sure to cite all of your sources in your explanation. Then find the check digit d for the ISBN-13 978-082183798d.

Exercises

(1) Determine if each of the following is a valid ID number. Assume that the last digit of each number is the check digit. If the check digit is not correct, fix it.

(a) 10513485, using the Luhn algorithm

(b) The ISBN-10 ID number 0-131-87718-4

(c) 2401346782, using the Verhoeff scheme.

(2) **Airline ticket ID numbers.** Airlines tickets have a 15-digit identification number. The first digit (reading left to right) is the coupon number and identifies the leg of the trip. (Coupon number 1 indicates the first flight in the trip, 2 the second, and so on, while coupon number 0 is the customer's receipt.) The next three digits identify the airline, and the next 10 digits comprise the document number, while the last digit is the check digit. Airlines use a simple mod 7 system to determine the check digit. If the airline ticket has digits

$$d_1d_2d_3d_4d_5d_6d_7d_8d_9d_{10}d_{11}d_{12}d_{13}d_{14}d_{15},$$

then the check digit d_{15} satisfies

$$d_{15} \equiv d_1d_2d_3d_4d_5d_6d_7d_8d_9d_{10}d_{11}d_{12}d_{13}d_{14} \pmod 7.$$

(a) Verify that 1-101-2134601379-2 is a valid airline ticket. Use a computer algebra system if necessary.

(b) Calculating the remainder when dividing a 14-digit number by 7 is easy for a computer algebra system but a bit time-consuming for humans. Here we will investigate one method that can be done by hand: the method of *casting out sevens*. This method works by determining the remainders when dividing powers of 10 by 7.

(i) Determine a general formula for calculating $10^n \pmod 7$ for nonnegative integers n.

(ii) Expand 11012134601379 in powers of 10, and then use your answer to part (a) to calculate the remainder when 11012134601379 is divided by 7.

(c) Another algorithm that can be used to determine if a large number is divisible by 7 is the following:

- Remove the last (rightmost) digit from the number.
- Subtract twice the value of the removed digit from the remaining number.
- Repeat until you can tell if the number obtained is divisible by 7.

(i) Apply this algorithm to show that 11012134601377 is divisible by 7. Then explain why the algorithm works.

(ii) Can a method similar to this one, where we remove the last digit and subtract some single-digit multiple of that digit from the new number, be used to test for divisibility by any other single digit integer? If so, find and explain all of the cases in which such a method can be used.

(3) In this exercise, we will examine the types of errors that are detected (or not detected) by the Luhn algorithm.

(a) Does the Luhn algorithm find all single digit errors? That is, if

$$a_{n-1} \cdots a_{i+1} a_i a_{i-1} \cdots a_0 d$$

and

$$a_{n-1} \cdots a_{i+1} b_i a_{i-1} \cdots a_0 d'$$

are valid IDs with $a_i \neq b_i$ and check digits d and d', must d and d' be different? If the answer is yes, prove it. If no, provide a counterexample.

(b) Will the Luhn algorithm detect all errors obtained by interchanging digits (e.g., typing 12 instead of 21)? If the answer is yes, prove it. If no, provide a counterexample.

(4) In this exercise, we will examine the types of errors that are detected (or not detected) by the ISBN-10 scheme.

(a) Show that the ISBN-10 scheme detects all single digit errors. (See part (a) of Exercise (3).)

(b) Show that the ISBN-10 scheme detects all errors obtained by interchanging digits (for example, typing 12 instead of 21).

(c) Another common family of errors are twin errors. One example of a twin error is changing aa to bb. Does the ISBN-10 scheme detect all twin errors? If the answer is yes, prove it. If no, provide a counterexample.

(d) Two other types of common errors are jump transposition errors (for example, abc replaced with cba) and jump twin errors (for example, aca replaced with bcb). Does the ISBN-10 scheme detect these errors? If the answer is yes, prove it. If no, provide a counterexample.

* (5) Activity 35.4 provided a method for finding a correct check digit using the Verhoeff scheme. Prove that this method works in general. That is, show that if the original ID number ends in 0 and has checksum c, then c^{-1} (from Table 35.2) is the correct value to use as the check digit in place of the final 0.

(6) In this exercise, we will examine the types of errors that are detected (or not detected) by the Verhoeff scheme.

(a) Show that the Verhoeff scheme detects all single digit errors. (See part (a) of Exercise (3).)

(b) Complete the following steps to show that the Verhoeff scheme detects all errors obtained by interchanging digits (for example, typing 12 instead of 21).

(i) Show by direct calculation that if $a \neq b$ in D_5, then $a \cdot \pi(b) \neq \pi(a) \cdot b$.

(ii) Extend the result of the previous part to show that if $a \neq b$ in D_5, then $\pi^{i-1}(a)\pi^i(b) \neq \pi^i(a)\pi^{i-1}(b)$ for all $i \in \mathbb{Z}^+$.

(iii) Now show that the Verhoeff scheme detects all errors obtained by interchanging digits.

Connections

Abstract algebra has many practical applications, and this investigation considered one of these applications. In particular, we saw how congruence of integers (from Investigation 2) and a dihedral group (from Investigation 21) can be used to create check digit schemes. These check digit schemes help make transfers of information more reliable and are therefore important components of our electronic world.

Investigation 36

Games: NIM and the 15 Puzzle

Focus Questions

By the end of this investigation, you should be able to give precise and thorough answers to the questions listed below. You may want to keep these questions in mind to focus your thoughts as you complete the investigation.

- How is the game of NIM related to group theory? What is a good strategy for playing NIM?
- What is the 15 Puzzle, and how can the symmetric groups tell us if a 15 Puzzle is solvable?

Games can be fun to play—and, as it turns out, to study. In fact, many games involve mathematical ideas or can be analyzed using mathematics. In this investigation, we will learn how group theory can be used to determine winning strategies in the game of NIM and to determine if a 15 Puzzle is solvable.

Preview Activity 36.1. Go to any online version of NIM and play the game a few times. Search for a winning strategy.

The Game of NIM

To play the game of NIM, one begins with a number of sets, or stacks, of objects. We can think of these stacks as piles of stones, as shown in Table 36.1. (Note that we have displayed the piles horizontally to save space.) In this example, the first pile has 6 stones, the second has 2, and the third has 3.

Table 36.1
A NIM game.

The game is played by two players alternating turns. At each turn, a player can take as many stones from a single pile as he or she wants, but the player must remove at least one stone. The object is to be the last player to remove stones.

The number of stones in each pile in a NIM game is an element in the set $\mathbb{W}$ of whole numbers. So we can think of a particular state of a NIM game with three piles as an element in the Cartesian product $\mathbb{W} \times \mathbb{W} \times \mathbb{W}$. Recall that $\mathbb{W}$ is not a group under standard addition of integers. Therefore, to relate this game to group theory, we will need to define a relevant operation under which $\mathbb{W}$ is a group. As we will see, using binary representations of whole numbers will help us to define such an operation.

To study binary representations, it will be helpful to first review how the decimal representation of a whole number works. Recall that we typically think of the digits of a whole number as representing place value. For example, in the integer 1234, the digit 4 is located in the ones place, the digit 3 is in the 10's place, 2 is in the 100's place, and 1 is in the 1000's place. In other words, the integer 1234 can also be represented by the sum

$$(1 \times 10^3) + (2 \times 10^2) + (3 \times 10^1) + (4 \times 10^0).$$

This is the *decimal representation* of the number 1234. In the decimal system, we add two whole numbers digit-by-digit, from right to left, reducing modulo 10 and carrying a 1 to the next digit whenever the sum of two digits is 10 or greater.

There is nothing particularly special about the base 10 used in the decimal representation of a number, other than the fact that it is convenient. After all, most people have 10 fingers and 10 toes, so a base 10 system is natural. However, we could just as easily replace the base 10 with any other base. In the *binary* system, we replace the base 10 with the base 2. There is an adjustment we must make, though. With the decimal system, each individual digit can be anywhere between 0 and 9 (because these integers are less than 10). With the binary system, we will only use the digits 0 and 1. For example, the binary number 10110 represents the decimal number $(1 \times 2^4) + (0 \times 2^3) + (1 \times 2^2) + (1 \times 2^1) + (0 \times 2^0) = 22$. In fact, any whole number can be represented in binary format.

To represent a given whole number in the binary system, we first look for the highest power of 2 that is less than the number. Then we subtract this highest power of 2 and repeat the process with the difference. To illustrate, let's convert 219 to binary. First note that $2^7 = 128$ and $2^8 = 256$. So 2^7 is the highest power of 2 less than 219. Now $219 - 1 \times 2^7 = 91$, and so $219 = (1 \times 2^7) + 91$. Repeating the process with the integer 91, we note that the highest power of 2 in 91 is 2^6. Since $91 - 2^6 = 27$, we see that $219 = (1 \times 2^7) + (1 \times 2^6) + 27$. We then continue reducing the differences until we no longer have any powers of 2 remaining. This leaves us with

$$\begin{aligned} 219 = {} & (1 \times 2^7) + (1 \times 2^6) + (0 \times 2^5) + (1 \times 2^4) \\ & + (1 \times 2^3) + (0 \times 2^2) + (1 \times 2^1) + (1 \times 2^0). \end{aligned}$$

Therefore, the binary representation of 219 is 11011011. The standard sum of two whole numbers $a = a_n a_{n-1} a_{n-2} \cdots a_1 a_0$ and $b = b_n b_{n-1} b_{n-2} \cdots b_1 b_0$ in binary is similar to the decimal sum: we add digit-by-digit, from right to left, reducing modulo 2 and carrying a 1 to the next digit whenever the sum of two digits is 2 or greater.

Remember that $\mathbb{W}$ is not a group under standard addition of integers. If, however, we convert each whole number into its binary representation, then we can define a special operation under which $\mathbb{W}$ is a group. Let $x = x_n x_{n-1} \ldots x_1 x_0$ and $y = y_n y_{n-1} \ldots y_1 y_0$ be whole numbers in binary form. We can assume both integers have the same number of digits in their binary representations by simply appending zeros to the left end of one number if necessary. We define the "NIM sum" of x and y to be the binary number $x \oplus y = s_n s_{n-1} \ldots s_1 s_0$, where $s_i = (x_i + y_i) \pmod 2$. Note that the NIM sum of two binary numbers is the same as the normal binary sum, except that we don't

allow carrying. For example, the NIM sum $101101 \oplus 00111$ is 101010. Of course, we can also add more than two numbers this way. For example, $10111 \oplus 1110 \oplus 111 = 11110$.

Activity 36.2. Let $\mathcal{B}$ be the set of binary representations of the whole numbers.

(a) Is the NIM sum operation $\oplus$ well-defined in $\mathcal{B}$? Explain.

(b) Is $\mathcal{B}$ closed under the NIM sum $\oplus$? Explain.

(c) Is there an identity element in $\mathcal{B}$ with respect to the NIM sum $\oplus$? If yes, what is the identity? If no, why not?

(d) Does $\mathcal{B}$ contain an inverse for each of its elements with respect to the NIM sum $\oplus$? If yes, what is the inverse of a given element? If no, why not?

(e) Is the NIM sum associative in $\mathcal{B}$? Prove your answer.

(f) What conclusion can we draw about $\mathcal{B}$?

Now that we have an operation under which the whole numbers in binary form are a group, we can also make a group out of $\mathcal{B}^n = \underbrace{\mathcal{B} \oplus \mathcal{B} \oplus \cdots \oplus \mathcal{B}}_{n \text{ factors}}$. The group $\mathcal{B}^n$ forms the playing field for all NIM games with n piles of stones—that is, each element in $\mathcal{B}^n$ corresponds to a particular stage in a NIM game. As such, we will call any element in $\mathcal{B}^n$ a *configuration*. If the i^{th} pile of stones contains N_i stones, then the NIM game has the configuration $(N_1, N_2, \ldots, N_n)$.

A *legal move* in a NIM game consists of removing some number of stones (at least one) from a single pile. Note that we can view a move as a configuration as well; in particular, the move that takes m stones from pile i can be thought of as the configuration $M = (0, 0, \ldots, 0, m, 0, \ldots, 0)$, with m in the i^{th} component. Since every element in $\mathcal{B}^n$ is its own inverse, the result of performing move M on configuration X is the configuration $X \oplus M$. For example, let $X = (011, 100, 001)$ be the configuration in $\mathcal{B}^3$ with 3 stones in the first pile, 4 in the second, and 1 in the third, as shown on the left in Table 36.2. Let $M = (010, 000, 000)$ be the move that takes 2 stones from the first pile. The result of applying the move M to the configuration X is the configuration $X \oplus M = (001, 100, 001)$, as shown on the right in Table 36.2.

Table 36.2
A NIM move.

In any NIM game, the last move will result in the configuration $(0, 0, \ldots, 0)$. This configuration has the special property that the NIM sum of all of the components is 0. In general, a configuration that satisfies this property is called an *even* configuration—that is, $(N_1, N_2, \ldots, N_n)$ is an *even* configuration if

$$\bigoplus_{i=1}^{n} N_i = 0.$$

Any other configuration is called an *odd* configuration. Note that every move is an odd configuration.

Activity 36.3. Let n be a positive integer, and let $\mathcal{E}^n$ be the subset of $\mathcal{B}^n$ consisting of the even configurations. Is $\mathcal{E}^n$ a subgroup of $\mathcal{B}^n$? Prove your answer.

We will now explore some strategy behind the game of NIM.

Activity 36.4.

(a) If $X \in \mathcal{E}^n$ is not the identity, how many nonzero components must X have?

(b) If $X \in \mathcal{E}^n$ is not the identity, how many nonzero components must its inverse have? Explain what this observation tells us about the possibility of winning a NIM game from a nonzero even configuration.

The result of Activity 36.4 reveals a strategy for playing NIM defensively. In particular, if we can always present our opponent with an even configuration, then he or she cannot win. The question now is how that can be done.

Activity 36.5. Let n be a positive integer, and let $X \in \mathcal{E}^n$. Determine and describe all moves M so that $(X \oplus M) \in \mathcal{E}^n$. Relate your answer to playing the game of NIM.

Activity 36.5 tells us that if we present our opponent with an even configuration, we will always be confronted with an odd configuration on our next turn. So the final question is whether we can convert an odd configuration into an even one. This is a bit more complicated, and so we will describe the process using a NIM game with three piles.

Let $X = (N_1, N_2, N_3)$ be an odd configuration in $\mathcal{B}^3$. We want to find a move $M = (M_1, M_2, M_3)$ with exactly one of M_1, M_2, M_3 nonzero so that $X \oplus M$ is even. (It is not true that the NIM sum of two odd configurations is even, and you should find a simple example to convince yourself of this.) Let $N_1 = a_m a_{m-1} \dots a_1 a_0$, $N_2 = b_m b_{m-1} \dots b_1 b_0$, and $N_3 = c_m c_{m-1} \dots c_1 c_0$ (all in binary). Since $N_1 \oplus N_2 \oplus N_3 \neq 0$, there is some index i so that $a_i + b_i + c_i \equiv 1 \pmod 2$. Let k be the largest index for which this happens. At least one of a_k, b_k, c_k must be 1. Without loss of generality, assume $a_k = 1$. This means that $b_k + c_k \equiv 0 \pmod 2$. For $0 \leq i < k$, let

$$a_i' = \begin{cases} 0, & \text{if } (b_i + c_i) \equiv 0 \pmod 2 \\ 1, & \text{otherwise.} \end{cases}$$

So $a_i' + b_i + c_i \equiv 0 \pmod 2$ for $i < k$. Let M be the move that takes stones from pile 1 so that $N_1' = a_m a_{m-1} \dots a_{k+1} 0 a_{k-1}' a_{k-2}' \dots a_1' a_0'$ remain. The result of applying move M to X is the configuration $X' = (X \oplus M) = (N_1', N_2, N_3)$. Recall that $(b_k + c_k) \equiv 0 \pmod 2$, and $(a_i + b_i + c_i) \equiv 0 \pmod 2$ if $i > k$ (by the definition of k). Since $a_i' + b_i + c_i \equiv 0 \pmod 2$ for $i < k$, we can therefore conclude that $N_1' \oplus N_2 \oplus N_3 = 0$, and $X' \in \mathcal{E}^n$.

In terms of the NIM game, this result tells us that if we are confronted with an odd configuration, we can always change it to an even configuration.

Activity 36.6. Apply the algorithm provided above to find a move that converts the NIM configuration in Table 36.3 to an even configuration.

Based on our work up to this point, we have proved the following theorem (with $n = 3$ for part (iii), but you are asked to extend this to $\mathcal{B}^n$ for any n in Exercise (2)).

Theorem 36.7. *Let n be a positive integer.*

(i) *If $X \in \mathcal{E}^n$ is nonzero, then there is no move M so that $X \oplus M = 0$.*

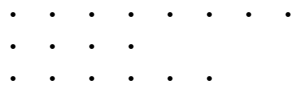

Table 36.3
A NIM configuration.

(ii) *If $X \in \mathcal{E}^n$ is nonzero, then $(X \oplus M) \notin \mathcal{E}^n$ for any move M.*

(iii) *If $X \in (\mathcal{B}^n \setminus \mathcal{E}^n)$, then there exists a move M so that $(X \oplus M) \in \mathcal{E}^n$.*

Interpreted in more natural language, Theorem 36.7 presents us with the following strategy for playing the NIM game:

(i) Since it is not possible to win from a nonzero even configuration, always present our opponent with an even configuration if possible.

(ii) If we can present our opponent with an even configuration, any move our opponent makes will present us with an odd configuration.

(iii) If we have an odd configuration at our turn, we can always turn it into an even configuration.

So our strategy is to always present our opponent with an even configuration. If the configuration is the 0 configuration, we have won. If not, then (ii) shows that our opponent cannot win because he or she will be forced to present us with an odd configuration. By part (iii), we can turn that odd configuration into an even configuration so that our opponent cannot win on the next turn. Since at least one stone is removed at each turn, our opponent will eventually have to present us with an odd configuration from which we can win.

One final note: if we are ever presented with even configuration, we cannot win the game unless our opponent makes a mistake. For this reason, it is always advantageous to be able to decide whether to move first or second after seeing the initial configuration.

We will illustrate the strategy described above with the game from Table 36.3.

Table 36.4
Our first move (left), our opponent's move (middle), and the coup de gras (right).

As we argued before, if we move first, we should remove stones from pile 1 to leave exactly 2 stones as shown on the left of Table 36.4. Now, whatever move our opponent makes, we will be left with an odd configuration. Suppose our opponent's move leaves us with the configuration shown in the middle of Table 36.4. This configuration gives us $N_1 = 2 = 10, N_2 = 4 = 100$, and $N_3 = 4 = 100$. Now $10 \oplus 100 \oplus 100 = 010$, so we must change the 1 in the 2's place. We can do this by removing all of the stones from pile 1, leaving us with the NIM sum $100 \oplus 100 = 000$, which corresponds to the configuration on the right of Table 36.4. Whatever moves our opponent makes, we can now win the game by keeping both piles of the same size. This will always result in a 0 NIM sum. Thus, the game is ours.

The 15 Puzzle

Preview Activity 36.8. Go to any on-line version of the 15 Puzzle and play the game a few times. Find one that allows you to create your own 15 Puzzle. Are all 15 Puzzles solvable? If not, search for a pattern that determines which puzzles are solvable.

The classic 15 Puzzle was made famous in the 19th century by puzzleist Sam Loyd.* The game consists of a starting position, which is a 4×4 array of the integers between 1 and 15 along with a symbol # (which we interpret as a blank space), as shown in Table 36.5. We will call each entry of the array a *cell*.

2	9	7	3
10	15	12	8
1	4	#	14
6	13	5	11

Table 36.5
15 Puzzle: Configuration 1.

The game is played with one type of legal move: interchanging the blank cell with a cell either to the left or right or the cell above or below. (The children's game is usually made of sliding tiles that are numbered 1 to 15. The interchange mentioned here is done by sliding a tile to the empty cell.) In this example, we can interchange the blank with the 12, 4, 14 or 5. Interchanging the blank and the 5 leaves us with the configuration in Table 36.6. We can then interchange the blank with the 5, 13, or 11.

2	9	7	3
10	15	12	8
1	4	5	14
6	13	#	11

Table 36.6
15 Puzzle: Configuration 2.

The object of the game is to interchange the blank with other cells and transform the starting

*In 2006, Jerry Slocum and Dic Sonneveld published their book, *The 15 Puzzle* (Slocum Puzzle Foundation), in which they write, "Sam Loyd did not invent the 15 puzzle and had nothing to do with promoting or popularizing it. The puzzle craze that was created by the 15 Puzzle began in January 1880 in the US and in April in Europe. The craze ended by July 1880 and Sam Loyd's first article about the puzzle was not published until sixteen years later, January 1896. Loyd first claimed in 1891 that he invented the puzzle, and he continued until his death a 20 year campaign to falsely take credit for the puzzle. The actual inventor was Noyes Chapman, the Postmaster of Canastota, New York, and he applied for a patent in March 1880."

position to the standard position in Table 36.7. It is important to note that each interchange is reversible. So an equivalent game is to begin with the standard position and move to obtain a specified position. It is this latter version of the game that we will analyze.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	#

Table 36.7
15 Puzzle: Standard configuration.

Permutations and the 15 Puzzle

To analyze this game, we will construct a correspondence between possible configurations of the 4×4 array and elements in the symmetric group S_{16}. To do this, we will let the number 16 represent the blank cell. Recall that any permutation can be written as a product of transpositions. We will apply a transposition $(a\ b)$ to a configuration A by interchanging the labels of the cells in *positions* a and b in A. (Note that this may be different than interchanging the cells *labeled* a and b.) For example, the transposition $(12\ 16)$ applied to the standard configuration I will correspond to Table 36.8 in which the cells in positions 12 and 16 have been interchanged.

1	2	3	4
5	6	7	8
9	10	11	#
13	14	15	12

Table 36.8
The transposition $(12\ 16)$ applied to I.

To apply a permutation $\sigma \in S_{16}$ to a configuration A, we can first write σ as a product $\tau_m \tau_{m-1} \cdots \tau_2 \tau_1$ of transpositions and then apply the transpositions, in order, to A. The resulting configuration is denoted by $\sigma(A)$. It is important to note that only certain permutations in S_{16} can be applied to a given configuration. In particular, the transpositions involved must always interchange two adjacent cells. So, for example, while we could apply the transposition $(12\ 16)$ to the standard configuration I (as we did in Table 36.8), we could not apply the transposition $(12\ 13)$ to I, since cells 12 and 13 are not adjacent. For convenience, we will refer to permutations in S_{16} that can be applied to I as *valid* permutations.

Activity 36.9. Find a permutation that converts the standard configuration to the configuration shown in Table 36.9.

We can construct a one-to-one correspondence between possible configurations and valid elements in S_{16} by assigning to each valid $\sigma \in S_{16}$ the configuration $\sigma(I)$. Because of this correspon-

1	#	2	4
5	6	3	8
9	10	7	12
13	14	11	15

Table 36.9
A target configuration.

dence, we will from this point on refer to valid elements of S_{16} as configurations, and vice versa. Note that the standard configuration is represented by the identity permutation.

Solving the 15 Puzzle

To determine if a 15 Puzzle is solvable, we are interested in answering the following equivalent questions:

- From which initial configuration can we obtain the standard configuration?
- Which configurations can be obtained starting from the standard configuration?

We will answer the second of these questions and, consequently, obtain the answer to the first question at the same time. To make our work a little easier, first observe that any configuration can be reduced to one in which the blank square is at location 16.

Activity 36.10. As an example, let A be the configuration shown in Table 36.5. Find a permutation σ for which $\sigma(A)$ produces the configuration shown in Table 36.10.

2	9	7	3
10	15	12	8
1	4	14	11
6	13	5	#

Table 36.10
Moving the blank to cell 16.

With this in mind, we need only to determine the configurations that have the blank in position 16 and can be obtained from the standard position I. The next theorem, which we will prove throughout the remainder of this investigation, tells us exactly which configurations meet these conditions.

Theorem 36.11. *Let H be the subset of S_{16} corresponding to all configurations that have the blank in position 16 and can be obtained from the standard position I. Then $H = A_{15}$.*

Note that even though S_{15} is not a subset of S_{16}, we can consider S_{15} to be contained in S_{16} as the set of all permutations that fix 16.

Activity 36.12.

(a) Explain how Theorem 36.11 tells us exactly which 15 Puzzles are solvable.

(b) Let σ_0 represent the configuration shown in Table 36.5. Can we find a sequence of allowable moves to transform σ_0 to I? Why or why not?

It is probably not surprising that H, the set of all permutations with the blank in position 16 that can be obtained from I, is a subgroup of S_{16}. This is left for you to prove in Exercise (4). It is also the case that every element in H is an even permutation. (See Exercise (5).) Thus, $H \subseteq A_{15}$. To complete the proof of Theorem 36.11, we need to show that $A_{15} \subseteq H$. Two facts will be useful in our argument:

- Lemma 24.16 (see page 342) shows that for $n \geq 3$, any permutation in A_n can be written as a product of 3-cycles.
- Exercise (16) of Investigation 22 (see page 318) shows that for any $\alpha \in S_{16}$, any $x, y, z \in \{1, 2, 3, \ldots, 16\}$, and any $n \in \mathbb{Z}^+$, we have

$$\alpha^n(x\ y\ z)\alpha^{-n} = (\alpha^n(x)\ \alpha^n(y)\ \alpha^n(z)). \tag{36.1}$$

Now we can show that H contains every 3-cycle of the form $(a\ b\ c)$ for $a, b, c \in S = \{1, 2, 3, \ldots, 15\}$. Since A_{15} is generated by three cycles, we will be able to conclude that $H = A_{15}$, as desired.

Lemma 36.13. *The group H contains every 3-cycle of the form $(a\ b\ c)$ for $a, b, c \in \{1, 2, 3, \ldots, 15\}$.*

Proof. Let $(a\ b\ c)$ be a 3-cycle with $a, b, c \in S = \{1, 2, 3, \ldots, 15\}$. If $\alpha \in S_{16}$ is in H and $(a\ b\ c) \in H$, then equation (36.1) shows that

$$\alpha^n(a\ b\ c)\alpha^{-n} = (\alpha^n(a)\ \alpha^n(b)\ \alpha^n(c)) \in H \tag{36.2}$$

for any $n \in \mathbb{Z}^+$.

To complete the proof of Lemma 36.13, we will apply this idea to some specific elements in H, as indicated in the next activity, to show the following:

(1) Every 3-cycle of the form $(11\ 7\ b)$ is in H, where $b \in S$ and $b \neq 7, 11, 16$.

(2) Every 3-cycle of the form $(a\ b\ 11)$ is in H, where $a, b \in S$, $a \neq 11, 16$, and $b \neq 7, 11, 16$.

(3) Every 3-cycle is in H.

Once we have established these facts, we will have proved the lemma. ■

Activity 36.14.

(a) Let $b \in S$ with $b \neq 7, 11, 16$. Here we will show that every 3-cycle of the form $(11\ 7\ b)$ is in H.

(i) Explain why

$$\alpha_1 = (11\ 15\ 12) = (16\ 12)(12\ 11)(11\ 15)(15\ 16)$$

(as shown in Table 36.11) is in H.

1	2	3	4
5	6	7	8
9	10	12	15
13	14	11	#

Table 36.11
α_1.

1	2	3	4
5	7	8	12
9	6	11	15
13	10	14	#

Table 36.12
α_2.

(ii) Explain why

$$\begin{aligned}\alpha_2 &= (16\ 12)(12\ 8)(8\ 7)(7\ 6)(6\ 10)(10\ 14)(14\ 15)(15\ 16)\\ &= (6\ 10\ 14\ 15\ 12\ 8\ 7)\end{aligned}$$

(as shown in Table 36.12) is an element of H. Then use equation (36.2) to show that $(11\ 7\ 6)$ is an element of H.

(iii) Next, note that the element α_3 defined by

$$\begin{aligned}\alpha_3 = (16\ 12)&(12\ 8)(8\ 4)(4\ 3)(3\ 2)(2\ 1)(1\ 5)(5\ 6)\\ &(6\ 10)(10\ 9)(9\ 13)(13\ 14)(14\ 15)(15\ 16)\\ = (1\ 5\ 6\ 10\ 4\ 13\ 14\ 15\ 12\ 8\ 4\ 3\ 2)&\end{aligned}$$

(as shown in Table 36.13) is another element of H.

2	3	4	8
1	5	7	12
10	6	11	15
9	13	14	#

Table 36.13
α_3.

Use equation (36.2) and the fact that α_3 fixes 7, 11, and 16 to complete the argument that every 3-cycle of the form $(11\ 7\ b)$, where $b \in S$ and $b \neq 7, 11, 16$, is in H.

(b) Let $a, b \in S$ with $a \neq 11, 16$ and $b \neq 7, 11, 16$. Show that every 3-cycle of the form $(a\ b\ 11)$ is in H. (Hint: Why is $(11\ b\ 7) \in H$?)

(c) Finally, show that every 3-cycle is in H. How does this show that $H = A_{15}$?

The work we did in Activity 36.14 completes our proof of Lemma 36.13. To summarize, we can determine if a particular 15 Puzzle is solvable by first transforming it to a puzzle A with the blank in position 16. Then if there is an even permutation σ so that $\sigma(I) = A$, we can conclude that our original 15 Puzzle is solvable.

Activity 36.15. Find a 15 Puzzle that is not solvable and that is also not easily seen to be unsolvable. Explain how you know your puzzle is not solvable.

One final note: Our analysis of the 15 Puzzle completely classifies which games can be won but does not tell us how to win. There are strategies for winning, but trial and error is often the best bet.

Concluding Activities

Activity 36.16. Go to any online site that has a NIM game and play it using the strategies we have described in this activity. Make sure you win!

Exercises

(1) Let m be a positive integer, and let $\mathcal{B}_m$ be the set of whole numbers, in binary, that are less than 2^m.

(a) Show that $\mathcal{B}_m$ is a subgroup of $\mathcal{B}$.

(b) As a finite Abelian group, the group $\mathcal{B}_m$ is isomorphic to a direct sum of cyclic groups. To what familiar direct sum is $\mathcal{B}_m$ isomorphic?

⋆ (2) Let n be a positive integer. Show that if $X \notin \mathcal{E}^n$, then there is a move M so that $(X + M) \in \mathcal{E}^n$.

(3) There is a story that when Sam Loyd first distributed his game, the configuration of the puzzle was the standard configuration but with the 14 and 15 pieces in reversed position. Loyd offered a prize of 1000 dollars for a correct solution to that puzzle. Did he ever pay the 1000 dollar prize? Explain.

⋆ (4) Let H be the subset of S_{16} consisting of all permutations with the blank in position 16 that can be obtained from the standard configuration I. Prove that H is a subgroup of S_{16}.

⋆ (5) Prove that if σ is in H, the set of all permutations with the blank in position 16 that can be obtained from I, then σ is an even permutation.

Connections

Abstract algebra has many applications—some serious, and some more fun. In this investigation, we saw how group theory could be used to analyze two games: NIM and the 15 Puzzle. The structure of the game of NIM is related to the group $\mathcal{B}^n$ and its subgroup $\mathcal{E}^n$ of the even configurations, while the symmetric groups are important in determining which 15 Puzzles can be solved.

Investigation 37

Groups of Order 8 and 12: Semidirect Products of Groups

Focus Questions

By the end of this investigation, you should be able to give precise and thorough answers to the questions listed below. You may want to keep these questions in mind to focus your thoughts as you complete the investigation.

- What is a semidirect product of groups? In what ways are semidirect products of groups useful?
- Up to isomorphism, how many groups are there of order 8?
- Up to isomorphism, how many groups are there of order 12?

Preview Activity 37.1. Summarize our previous work in classifying groups, and list all of the isomorphism classes of groups of orders from 1 to 7, 9 to 11, and 13 to 15. Provide whatever information you can about groups of orders 8 and 12.

Introduction

As Preview Activity 37.1 demonstrates, we have classified all groups of order 1 through 15, with the exception of groups of order 8 and 12. We have also determined all groups of prime order (see Activity 23.17 on page 326 and Theorem 26.18 on page 383), all groups of order $2p$, where p is prime (see Corollary 24.13 on page 341 and Theorem 26.16 on page 381), all groups of order pq where p and q are distinct primes and p does not divide $q - 1$ (see Activity 29.32 on page 431), and all groups of order p^2, where p is prime (see Activity 29.15 on page 425). In this investigation, we will fill in the gaps and classify all groups of order 8 and 12, which will complete our classification of groups of order less than 16. (Since there are 14 groups of order 16, this seems like a reasonable place to stop.) In the process of classifying the groups of order 8 and 12, we will introduce and use the idea of the semidirect product of two groups. We will also see that when classifying groups of a given order, it is not the order itself that determines the difficulty of the classification, but rather how large the powers of the prime divisors of the order are.

Groups of Order 8

In this section, we will determine the distinct isomorphism classes of groups of order 8. Since the Fundamental Theorem of Finite Abelian Groups tells us about the Abelian groups of order 8, we will concentrate on non-Abelian groups. Recall that we already know two non-Abelian groups of order 8—namely, D_4 and the quaternions $\mathbf{Q}$. We will now determine if there are, up to isomorphism, any other non-Abelian groups of order 8, and we will classify all such groups.

Activity 37.2. Let G be a non-Abelian group of order 8 with identity e.

(a) Use Exercise (2) of Investigation 18 (see page 274) to explain why G must contain an element b of order 4.

(b) Let $N = \langle b \rangle$. Explain why N is normal in G.

(c) Let $a \in G$ with $a \notin N$. Then $G = N \cup aN = \{e, b, b^2, b^3, a, ab, ab^2, ab^3\}$. The operation table for G will be determined once we know how to represent the elements ba and a^2 in the form $a^i b^j$, with $0 \leq i \leq 1$ and $0 \leq j \leq 3$. Given that $N \triangleleft G$, we know that $aba^{-1} \in N$, so $aba^{-1} = b^t$ for some t with $0 \leq t \leq 3$

(i) Explain why t cannot be 0 or 1.

(ii) Now explain why t cannot be equal to 2. (Hint: What is $|aba^{-1}|$?)

(iii) Assume $aba^{-1} = b^3 = b^{-1}$. Since $a \in G$ and $a \neq e$, we must have $|a| = 2$ or $|a| = 4$. What can we say about G if $|a| = 2$?

(iv) What can we say about G if $aba^{-1} = b^{-1}$ and $|a| = 4$?

(d) How many groups are there of order 8?

Activity 37.2 tells us about the groups of order 8, so we will now turn our attention to groups of order 12. It is possible (and not that difficult) to classify the groups of order 12 directly, but in the next section we will introduce a new tool, the semidirect product, that is very helpful in the general context of classifying groups of a given order. We will then use semidirect products to classify groups of order 12 and of order p^3, where p is an odd prime.

Semi-direct Products of Groups

Preview Activity 37.3. Let G be the set of ordered pairs

$$\{([a]_3, [b]_2) : [a]_3 \in \mathbb{Z}_3, [b]_2 \in \mathbb{Z}_2\}.$$

Define an operation $\cdot$ on G by

$$([a]_3, [b]_2) \cdot ([c]_3, [d]_2) = ([a + (-1)^b c]_3, [b + d]_2)$$

(a) Explain why this operation is well-defined on G.

(b) Construct the operation table for the set G with the operation defined above.

(c) Is G a group under this operation? If no, why not? If yes, to what familiar group is G isomorphic? Explain.

While we can decompose some groups into an internal direct product of normal subgroups, we cannot do this for every group. For example, if we try to write the group D_3 as an internal direct product of two proper subgroups, we run into a problem. Recall that the group $H = \langle R \rangle$ is a normal subgroup of D_3 of order 3, so we would need a normal subgroup K of order 2 to be able to write D_3 as the product $H \times K$. However, D_3 has no normal subgroup of order 2. So we cannot decompose D_3 into an internal direct product of nontrivial normal subgroups. It turns out, however, that we can decompose D_3 into what is called a semidirect product of subgroups. Activity 37.3 gives an example of such a decomposition. (Don't worry if the construction there does not seem obvious or natural to you at the moment.)

Up to this point, we have seen several different products of groups:

- The direct product $H \oplus K$ of two groups is again a group, external to both H and K.
- If H and K are subgroups of a group G with $H \triangleleft G$, then the product

  $$HK = \{hk : h \in H, K \in K\}$$

  is a subgroup of G. (See Activity 24.22 on page 346.) Moreover, if $H \cap K = \{e\}$, where e is the identity in G, then $|HK| = |H||K|$. To see this, suppose $h_1k_1 = h_2k_2$ in HK. Then $h_2^{-1}h_1 = k_2k_1^{-1} \in (H \cap K)$, so $h_2^{-1}h_1 = k_2k_1^{-1} = e$. Thus, $h_1 = h_2$, and $k_1 = k_2$. It follows that if $H \cap K = \{e\}$ and $|H||K| = |G|$, then $G = HK$.
- If H and K are normal subgroups of a group G with $H \cap K = \{e\}$, then $HK = H \times K$ is the internal direct product of H and K.

In this section, we will construct another type of product called a *semidirect product*. To understand how semidirect products work, recall that the construction of the internal product HK requires that H and K be subgroups of some group G that is already known. The question we want to answer now is if we can generalize this construction. In other words, given any two arbitrary groups H and K, can we find a group G so that G contains copies of both H and K—that is, subgroups H' and K' that are isomorphic to H and K, respectively—with $H' \triangleleft G$ and $H' \cap K' = \{e\}$, where e is the identity in G? If so, then $G = H'K'$, and we will denote this special decomposition as $H \rtimes K$. (We will say more about this notation later.)

To explore this construction more, let H and K be groups with identities e_H and e_K, respectively. We want to find a group G that contains isomorphic copies of H and K satisfying the conditions described above. A natural place to start is to let $G = \{(h, k) : h \in H, k \in K\}$. Certainly, G will contain a copy $H' = \{(h, e_K) : h \in H\}$ of H and a copy $K' = \{(e_H, k) : k \in K\}$ of K. The key to constructing the group G is to define an appropriate operation. We want to make $G = H'K'$ with $H' \triangleleft G$. This will mean that if $k \in K'$ and $h \in H'$, then $khk^{-1} \in H'$. If $a = h_1k_1$ and $b = h_2k_2$ are elements of $H'K'$, then it will follow that

$$ab = (h_1k_1)(h_2k_2) = h_1\left(k_1h_2k_1^{-1}\right)(k_1k_2) \tag{37.1}$$

is also an element of $H'K'$. So the operation we define in G will need to mimic the product in (37.1).

What makes the product in (37.1) work is that $k_1 h_2 k_1^{-1}$ is in H. In fact, conjugation by the element k_1 is an automorphism of H (an inner automorphism to be specific). We also saw this idea in Activity 37.3, where left multiplication by $(-1)^b$ for $b = 0$ or $b = 1$ is an automorphism on $\mathbb{Z}_3$. In other words, we had a mapping φ with domain $\mathbb{Z}_2$ that assigned to each $[b] \in \mathbb{Z}_2$ an automorphism on $\mathbb{Z}_3$. More specifically, we had $\varphi([0]_2)$ as the identity automorphism and $\varphi([1]_2)$ as the automorphism that sends $[a]_3$ to $[2a]_3$ for all $[a]_3 \in \mathbb{Z}_3$. Expressed another way, $\varphi : \mathbb{Z}_2 \to \text{Aut}(\mathbb{Z}_3)$ is the mapping for which $\varphi([0]_2)([a]_3) = [a]_3$ (for all $[a]_3 \in \mathbb{Z}_3$) and $\varphi([1]_2)([a]_3) = [2a_3]$ (for all $[a]_3 \in \mathbb{Z}_3$). (It turns out that φ is a homomorphism as well, and you should verify that for yourself.)

Before we proceed, a word of caution is in order: the above notation can be very confusing since we are dealing with functions whose images are functions as well. As you work through this section, it is vitally important to distinguish between the elements of a given group and the functions that act on these elements. In the next activity, we will explore these ideas in a more general context.

Activity 37.4. Let H and K be groups with identities e_H and e_K, respectively, let $\varphi : K \to \text{Aut}(H)$ be a homomorphism, and let $G = H \times K$ be the Cartesian product of H and K. Then we can define a product on G as follows:

$$(h_1, k_1)(h_2, k_2) = (h_1 \varphi(k_1)(h_2), k_1 k_2). \tag{37.2}$$

Note that this product has the same form as the products in (37.1) and Activity 37.3. In Activity 37.3, our example turned out to be a group, so it seems reasonable to ask if G will always be a group with the product defined by (37.2).

(a) Is G closed under the operation from (37.2)?

(b) Does G contain an identity element? If so, what is it? Explain.

(c) Is the operation defined by (37.2) associative? Prove your answer.

(d) Does G contain an inverse for each of its elements? If so, what is the form of an inverse of an element in G?

Activity 37.4 tells us that G, as defined above, is a group under the operation from (37.2). We can actually say more about the group G, as stated in the following theorem.

Theorem 37.5. *Let H and K be groups with identities e_H and e_K, respectively, and let $\varphi : K \to Aut(H)$ be a homomorphism. Then $G = \{(h, k) : h \in H, k \in K\}$ with the operation*

$$(h_1, k_1)(h_2, k_2) = (h_1 \varphi(k_1)(h_2), k_1 k_2)$$

is a group. Moreover,

(i) *$H' = \{(h, e_K) : h \in H\}$ is a normal subgroup of G isomorphic to H;*

(ii) *$K' = \{(e_H, k) : k \in K\}$ is a subgroup of G isomorphic to K; and*

(iii) *$H' \cap K' = \{(e_H, e_K)\}$.*

Proof. Since Activity 37.4 shows that G is a group, we will focus here on parts (i) – (iii). In particular, we will prove part (i) and leave the remaining parts for the reader in Exercise (4).

We will show that $H' = \{(h, e_K) : h \in H\}$ is a normal subgroup of G isomorphic to H. The

element (e_H, e_K) is in H', so H' contains the identity element in G. Let (h_1, e_K) and (h_2, e_K) be in H'. Then

$$(h_1, e_K)(h_2, e_K) = (h_1, \varphi(e_K)(h_2), e_K) = (h_1 h_2, e_K) \in H',$$

and so H' is closed under the operation in G.

Also,

$$(h_1, e_K)^{-1} = \left(\varphi\left(e_K^{-1}\right)\left(h_1^{-1}\right), e_K^{-1}\right) = \left(h_1^{-1}, e_K\right) \in H',$$

and so H' is a subgroup of G by the Subgroup Test. (See page 279.)

To show that H' is a normal subgroup of G, let $(h, e_K) \in H'$ and let $g = (a, b) \in G$. Then

$$\begin{aligned}(a,b)^{-1}(h,e_K)(a,b) &= \left(\varphi\left(b^{-1}\right)\left(a^{-1}\right), b^{-1}\right)(h\varphi(e_K)(a), b) \\ &= \left(\varphi\left(b^{-1}\right)\left(a^{-1}\right)\varphi\left(b^{-1}\right)(ha), b^{-1}b\right) \\ &= \left(\varphi\left(b^{-1}\right)\left(a^{-1}\right)\varphi\left(b^{-1}\right)(ha), e_K\right)\end{aligned}$$

and $(a, b)^{-1}(h, e_K)(a, b) \in H'$. Therefore, $g^{-1}H'g = H'$ and $H' \triangleleft G$.

That H' is isomorphic to H can be shown by considering the mapping $\alpha : H \to H'$ defined by $\alpha(h) = (h, e_K)$. Let $h_1, h_2 \in H$. Then

$$\alpha(h_1 h_2) = (h_1 h_2, e_K) = (h_1, e_K)(h_2, e_K) = \alpha(h_1)\alpha(h_2),$$

and α is a homomorphism. If $\alpha(h_1) = \alpha(h_2)$, then $(h_1, e_K) = (h_2, e_K)$ and $h_1 = h_2$. Thus, α is a monomorphism. If $(x, e_K) \in H'$, then $\alpha(x) = (x, e_K)$, and so α is an epimorphism. We have therefore shown that α is an isomorphism and $H \cong H'$.

■

The group G described in Theorem 37.5 is called the **semidirect product** of H and K and is denoted $H \rtimes_\varphi K$ to indicate its dependence on the particular homomorphism $\varphi : K \to \text{Aut}(H)$. When the homomorphism φ is clear from the context, we will simply write $H \rtimes K$.

Definition 37.6. Let H and K be groups, and let $\varphi : K \to \text{Aut}(H)$ be a homomorphism. The **semidirect product** of H and K with respect to φ is the group

$$H \rtimes_\varphi K = \{(h, k) : h \in H, k \in K\}$$

with the operation

$$(h_1, k_1)(h_2, k_2) = (h_1 \varphi(k_1)(h_2), k_1 k_2).$$

Why are semidirect products important? First, the direct product $H \oplus K$ is an example of a semidirect product (see Exercise (3)), so we can think of the semidirect product as an extension of the direct product. (This is also the main motivation for the notation $\rtimes$ for the semidirect product, as it is more general than the internal direct product.) Second, if φ is a nontrivial homomorphism, then the semidirect product $H \rtimes_\varphi K$ is a non-Abelian group (see Exercise (6)), so semidirect products provide a method for constructing non-Abelian groups. Semidirect products are also useful in classifying groups. For example, if H and K are subgroups of a group G so that H is normal in G, $H \cap K$ is trivial, and $|HK| = |G|$, then G will be isomorphic to a semidirect product $H \rtimes_\varphi K$ for some φ. The next activity makes this last point clear.

Activity 37.7. Let G be a group with identity e and subgroups H and K such that

(i) $H \triangleleft G$ and

(ii) $H \cap K = \{e\}$.

Let $\varphi : K \to \text{Aut}(H)$ be the mapping that sends each element $k \in K$ to the inner automorphism defined by conjugation by k. That is, let $\varphi(k) = \pi_k$, where $\pi_k(h) = khk^{-1}$. There is a natural function $\Phi : HK \to (H \rtimes_\varphi K)$. Define this function Φ and show that it is an isomorphism. Then explain how we have proved the following theorem:

Theorem 37.8. *Let G be a group with identity e and subgroups H and K with*

(i) $H \triangleleft G$ *and*

(ii) $H \cap K = \{e\}$.

Then $HK \cong (H \rtimes_\varphi K)$ for some homomorphism $\varphi : K \to Aut(H)$.

In the next section, we will use Theorem 37.8 to classify all groups of order 12 and all groups of order p^3, where p is a prime.

Groups of Order 12 and p^3

We will begin this section with groups of order 12. The Fundamental Theorem of Finite Abelian Groups tells us that the groups $\mathbb{Z}_{12}$ and $\mathbb{Z}_6 \oplus \mathbb{Z}_2$ are the distinct Abelian groups of order 12. We already know at least three non-Abelian groups of order 12—namely, D_6, A_4, and T, where T is described in Exercise (18) of Investigation 22 (see page 318) as having presentations $\langle s, t \mid s^6 = 1, s^3 = t^2, sts = t\rangle$ and $\langle x, y \mid x^4 = y^3 = 1, yxy = x\rangle$. We will now show that these three groups are, up to isomorphism, the only non-Abelian groups of order 12.

Let G be a non-Abelian group of order 12 with identity e. Since $12 = 2^2 \times 3$, G has a Sylow 3-subgroup H of order 3 and a Sylow 2-subgroup K of order 4. Since H is cyclic, $H = \langle h\rangle$ for some $h \in G$.

Define $\varphi : G \to P(G/H)$ by $\varphi(a) = \pi_a$ where $\pi_a(gH) = (ag)H$. Recall that G/H denotes the collection of left cosets of H in G (even if H is not normal in G) and $P(G/H)$ denotes the group of permutations of G/H. In Exercise (7) of Investigation 30 (see page 442), we showed that φ is a homomorphism. Now $[G : H] = 4$, and so $P(G/H) \cong S_4$. If φ is a monomorphism, then G is isomorphic to a subgroup of order 12 in S_4. The only such subgroup is A_4, and so $G \cong A_4$ in this case.

Now assume that $|\text{Ker}(\varphi)| > 1$. We will next show that $\text{Ker}(\varphi) \subseteq H$. Let $a \in \text{Ker}(\varphi)$. Then $\varphi(a) = \pi_a$ is the identity permutation in $P(G/H)$, so $H = \pi_a(H) = aH$ and $a \in H$. Thus, $\text{Ker}(\varphi) \subseteq H$. Because $|H| = 3$ and $|\text{Ker}(\varphi)| > 1$, it follows that $\text{Ker}(\varphi) = H$, and so $H \triangleleft G$.

We know that $H \cap K = \{e\}$ and that $|G| = |H||K|$, so $G = HK$. Theorem 37.8 shows us that $G \cong (H \rtimes_\varphi K)$ for some φ. Next we will determine the different groups of this form.

Since $|H| = 3$, we know that $H \cong \mathbb{Z}_3$. Recall from Exercise (34) of Investigation 26 (see page 390) that $\text{Aut}(\mathbb{Z}_3) \cong U_3$, so $|\text{Aut}(H)| = 2$. The two automorphisms of H are the identity automorphism π_0 and the automorphism π_1 defined by $\pi_1(h) = h^{-1} = h^2$. Since $|K| = 4$, there are two possibilities for K: $K \cong \mathbb{Z}_4$ and $K \cong (\mathbb{Z}_2 \oplus \mathbb{Z}_2)$.

Case 1: $K \cong \mathbb{Z}_4$**.** In this case, we can identify K with $\mathbb{Z}_4$, and so any homomorphism φ from K to

Aut(H) is determined by its action on $[1]$. Thus, there are two possibilities: $\varphi([1]) = \pi_0$ and $\varphi([1]) = \pi_1$. In the first case, G will be Abelian and isomorphic to $\mathbb{Z}_3 \oplus \mathbb{Z}_4$. In the second case, the product in HK will have the form

$$(h_1k_1)(h_2k_2) = (h_1h_2^{-1})(k_1k_2).$$

In particular, we will have

$$(h^ik^j)(h^uk^v) = h^{i-j}k^{u+v}$$

for all i, j, u, and v. More specifically, $hkh = k$. So G has the presentation

$$\langle h, k \mid h^3 = k^4 = 1, hkh = k\rangle,$$

and $G \cong T$.

Case 2: $K \cong (\mathbb{Z}_2 \oplus \mathbb{Z}_2)$**.** In this case, we will identify K with $\mathbb{Z}_2 \oplus \mathbb{Z}_2$, and so any homomorphism from K to Aut(H) will be determined by its actions on the generators $([1], [0])$ and $([0], [1])$.

- Let $\varphi_0 : K \to \text{Aut}(H)$ be defined by

  $$\varphi_0(([1], [0])) = \pi_0 \text{ and } \varphi_0(([0], [1])) = \pi_0.$$

  Then G is Abelian and isomorphic to $\mathbb{Z}_3 \oplus (\mathbb{Z}_2 \oplus \mathbb{Z}_2)$.
- Let $\varphi_1 : K \to \text{Aut}(H)$ be defined by

  $$\varphi_1(([1], [0])) = \pi_0 \text{ and } \varphi_1(([0], [1])) = \pi_1.$$

  Then

  $$\varphi_1(([a], [b])) = \varphi_1(a([1], [0]) + b([0], [1])) = \pi_0^a\pi_1^b = \pi_1^b.$$

  Note that $\pi_1^b(t) = t^{2^b}$. In this case, the product on HK is

  $$\begin{aligned}(h^i([a], [b]))(h^u([c], [d])) &= (h^i\pi_1^b(h^u))([a + c], [b + d])\\ &= h^{i+u2^b}([a + c], [b + d]).\end{aligned}$$

  Let $x = ([1], [0])$ in K. Then $(hx)^2 = (hx)(hx) = h^2$, $(hx)^3 = x$, $(hx)^4 = h$, $(hx)^5 = h^2x$, and $(hx)^6 = e$. So $|hx| = 6$, and $N = \langle hx\rangle$ is a subgroup of G of order 6. Since $[G : N] = 2$, it follows that $N \triangleleft G$. (See Exercise (19) on page 349.) Let $y = ([0], [1]) \in K$. Then $y^2 = e$, and so $G = N\langle y\rangle$. Note that

  $$y(hx)y^{-1} = y(hx)y = h^2([1], [1])y = h^2x = (hx)^{-1}.$$

  So G has presentation $\langle hx, y \mid (hx)^6 = y^2 = 1, y(hx)y^{-1} = (hx)^{-1}\rangle$. But this is exactly the presentation for D_6, and so $G \cong D_6$ in this case.
- Let $\varphi_2 : K \to \text{Aut}(H)$ be defined by

  $$\varphi_2(([1], [0])) = \pi_1 \text{ and } \varphi_2(([0], [1])) = \pi_0.$$

  In this case, we also have $G \cong D_6$, which is left as an exercise for you to verify. (See Exercise (5).)
- Let $\varphi_3 : K \to \text{Aut}(H)$ be defined by

  $$\varphi_3(([1], [0])) = \pi_1 \text{ and } \varphi_3(([0], [1])) = \pi_1.$$

  In this case, we again have $G \cong D_6$, which is left as an exercise for you to verify. (See Exercise (5).)

We can therefore conclude that the only non-Abelian groups of order 12 are D_6, A_4, and T.

Our classification of groups of order 12 demonstrates that two groups $H \rtimes_\alpha K$ and $H \rtimes_\beta K$ can be isomorphic even if $\alpha \neq \beta$. In general, it can be difficult to determine if two semidirect products with the same underlying groups are isomorphic or not. One tool is given in the next lemma (which we will use in our classification of groups of order p^3), and others are presented in Activity 37.10.

Lemma 37.9. *Let K be a finite cyclic group, and let H be any group. Let $\varphi_1 : K \to Aut(H)$ and $\varphi_2 : K \to Aut(H)$ be homomorphisms so that $\varphi_1(K)$ and $\varphi_2(K)$ are conjugate subgroups of $Aut(H)$. Then $(H \rtimes_{\varphi_1} K) \cong (H \rtimes_{\varphi_2} K)$.*

Proof. Since K is cyclic, there is an element k so that $K = \langle k \rangle$. The fact that $\varphi_1(K)$ and $\varphi_2(K)$ are conjugate subgroups of $\mathrm{Aut}(H)$ means that there exists $\delta \in \mathrm{Aut}(H)$ so that $\delta^{-1}\varphi_2(K)\delta = \varphi_1(K)$. So

$$\varphi_1(k) = \delta^{-1}\varphi_2(k^t)\delta$$

for some integer t. Then

$$\varphi_1(k)^i = \left(\delta^{-1}\varphi_2(k^t)\delta\right)^i,$$

and so

$$\begin{aligned}\varphi_1(k^i) &= \delta^{-1}\varphi_2(k^t)^i\delta \\ &= \delta^{-1}\varphi_2(k^i)^t\delta\end{aligned}$$

for any integer i. Thus,

$$\varphi_1(x) = \delta\varphi_2(x^t)\delta^{-1} \tag{37.3}$$

for all $x \in K$.

Since $K = \langle k \rangle$, it follows that $\mathrm{Im}(\varphi_1) = \langle \varphi_1(k) \rangle$ and $\mathrm{Im}(\varphi_2) = \langle \varphi_2(k) \rangle$. Since $\varphi_1(k)$ and $\varphi_2(k)$ are conjugate, it also follows that $|\mathrm{Im}(\varphi_1)| = |\mathrm{Im}(\varphi_2)|$. Thus, $|\varphi_1(k)| = |\varphi_2(k)|$. Equation (37.3) shows that $\varphi_1(k) = \delta^{-1}\varphi_2(k)^t\delta$, and so $|\varphi_2(k)| = |\varphi_1(k)| = |\varphi_2(k)^t|$. Thus, $\gcd(|\varphi_1(K)|, t) = 1$. Since $|\varphi_1(K)|$ divides $|K|$, Exercise (12) shows that there is an integer t' with $t' \equiv t \pmod{|\varphi_1(K)|}$ and $\gcd(t', |K|) = 1$. Since $\varphi_2(x^t) = \varphi_2(x)^t = \varphi_2(x)^{t'} = \varphi_2(x^{t'})$ for all $x \in K$, we can replace t with t' in (37.3). This allows us to assume without loss of generality that $\gcd(t, |K|) = 1$, and so there exist integers x and y with $x|K| + yt = 1$.

Define $\Psi : (H \rtimes_{\varphi_1} K) \to (H \rtimes_{\varphi_2} K)$ by $\Psi((a, b)) = (\delta(a), b^t)$. To show that Ψ is a homomorphism, let (a_1, b_1) and (a_2, b_2) be in $H \times K$. Let $\cdot_1$ denote the operation in $H \rtimes_{\varphi_1} K$, and let $\cdot_2$ denote the operation in $H \rtimes_{\varphi_2} K$. Then

$$\begin{aligned}\Psi((a_1, b_1) \cdot_1 (a_2, b_2)) &= \Psi((a_1\varphi_1(b_1)(a_2), b_1b_2)) \\ &= (\delta(a_1\varphi_1(b_1)(a_2)), (b_1b_2)^t) \\ &= (\delta(a_1(\delta^{-1}\varphi_2(b_1^t)\delta)(a_2)), b_1^tb_2^t) \\ &= (\delta(a_1)\delta((\delta^{-1}\varphi_2(b_1^t)\delta)(a_2)), b_1^tb_2^t) \\ &= (\delta(a_1)\varphi_2(b_1^t)(\delta(a_2)), b_1^tb_2^t) \\ &= (\delta(a_1), b_1^t) \cdot_2 (\delta(a_2), b_2^t) \\ &= \Psi((a_1, b_1)) \cdot_2 \Psi((a_2, b_2)),\end{aligned}$$

and Ψ is a homomorphism.

To show that Ψ is a monomorphism, suppose $(a, k^r) \in \mathrm{Ker}(\Psi)$. Let e_H be the identity in H, and let e_K be the identity in K. Then

$$(e_H, e_K) = \Psi((a, k^r)) = (\delta(a), k^{rt}),$$

and so $\delta(a) = e_H$ and $k^{rt} = e_K$. Since $\delta \in \text{Aut}(H)$, we can conclude that $a = e_H$. That $k^{rt} = e_K$ means $|K|$ divides rt. But $\gcd(t, |K|) = 1$ implies that $|K|$ divides r. Thus, $k^r = e_K$, and so (a, k^r) is the identity in $H \rtimes_{\varphi_1} K$. Therefore, $\text{Ker}(\Psi)$ is trivial and Ψ is a monomorphism.

Finally, we will demonstrate that Ψ is an epimorphism. Let $(w, z) \in (H \rtimes_{\varphi_2} K)$. Since δ^{-1} is a surjection, there exists $a \in H$ such that $\delta^{-1}(a) = w$. Recall that $x|K| + yt = 1$, so

$$z = z^{x|K|+yt} = \left(z^{|K|}\right)^x (z^y)^t = (z^y)^t .$$

Then

$$\Psi\left((\delta^{-1}(a), z^y)\right) = (w, z),$$

and we have shown that Ψ is an epimorphism. Therefore, Ψ is an isomorphism, and $(H \rtimes_{\varphi_1} K) \cong (H \rtimes_{\varphi_2} K)$.

■

We will end this investigation by classifying all groups of order p^3, where p is a prime. Doing so will illustrate the fact that it is not necessarily the size of a group that makes classification difficult, but rather how large the powers of the prime divisors of the order are. In working out the classification, we will leave a number of details for you to complete in the exercises. Since we have already classified the groups of order 8, we will restrict ourselves to odd primes.

Let G be a group of order p^3, where p is an odd prime. We know the Abelian groups of order p^3 by the Fundamental Theorem of Finite Abelian Groups: $\mathbb{Z}_{p^3}$, $\mathbb{Z}_{p^2} \oplus \mathbb{Z}_p$, and $\mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_p$. Therefore, we will now focus on the non-Abelian groups of order p^3.

Let G be a non-Abelian group of order p^3. Since G is a p-group, we know that $Z = Z(G)$ is nontrivial. Since G is non-Abelian, there are two possibilities for $|Z|$: $|Z| = p$ or $|Z| = p^2$. If $|Z| = p^2$, then $|G/Z| = p$ and so G/Z is cyclic. It follows then that G is Abelian (see Theorem 24.10 on page 338), a contradiction. We can therefore conclude that $|Z| = p$. Thus, $|G/Z| = p^2$, and so $G/Z \cong \mathbb{Z}_{p^2}$ or $G/Z \cong (\mathbb{Z}_p \oplus \mathbb{Z}_p)$. Since G/Z is not cyclic, it must be that $G/Z \cong (\mathbb{Z}_p \oplus \mathbb{Z}_p)$.

Recall from Exercise (9) of Investigation 21 (see page 306) that the commutator of elements $x, y \in G$ is the element $[x, y] = x^{-1}y^{-1}xy$. Since G is non-Abelian, the subgroup G' generated by the commutators of pairs of elements is nontrivial. Let $x, y \in G$. Then $xZ, yZ \in G/Z \cong (\mathbb{Z}_p \oplus \mathbb{Z}_p)$ (an Abelian group), and so

$$Z = (xZ)^{-1}(yZ)^{-1}(xZ)(yZ) = (x^{-1}y^{-1}xy)Z.$$

This, however, implies that $x^{-1}y^{-1}xy \in Z$. Thus, $G' \subseteq Z$, and so $G' = Z$.

Define $\psi : G \to G$ by $\psi(g) = g^p$. That ψ is a homomorphism can be demonstrated as follows. Let $a, b \in G$. Since $G' = Z$, the commutators commute with every element in G. Exercise (8) and the fact that every nonidentity element in Z has order p show that

$$\begin{aligned}
\psi(a)\psi(b) &= a^p b^p \\
&= (ab)^p [a, b]^{p(p-1)/2} \\
&= (ab)^p \left([a, b]^{(p-1)/2}\right)^p \\
&= (ab)^p \\
&= \psi(ab),
\end{aligned}$$

and ψ is a homomorphism. Moreover, we can again use the fact that every nonidentity element in G/Z has order p to see that

$$Z = (aZ)^p = a^p Z,$$

and so $a^p \in Z$ for every $a \in G$. Thus, $\text{Im}(\psi) \subseteq Z$.

Now consider $\text{Ker}(\psi)$. If every nonidentity element in G has order p, then $\text{Ker}(\psi) = G$. Otherwise, G contains an element of order p^2. We will consider each of these cases. Let e be the identity in G.

Case 1: There is an element h of order p^2 in G. Let $H = \langle h \rangle$. It follows that $H \cap Z = \langle h^p \rangle$. By Exercise (7), we have that H is normal in G. If we can find a subgroup K of G so that $G = HK$, then we will have $G \cong (H \rtimes_\varphi K$ for some φ.

Since $h^p \neq e$, we have $\text{Ker}(\psi) \neq G$. So $|\text{Im}(\psi)| > 1$, and it follows that $\text{Im}(\psi) = Z$. Thus, $|\text{Ker}(\psi)| = p^2$. Since $h \notin \text{Ker}(\psi)$, we have that $\text{Ker}(\psi) \neq H$. Let k be any element of $\text{Ker}(\psi)$ such that $k \notin H$, and let $K = \langle k \rangle$. Note that $e = \psi(k) = k^p$, so $|k| = p$. Since every nonidentity element in K generates K, it follows that $H \cap K = \{e\}$, and so $G = HK$. Thus, $G \cong (\mathbb{Z}_{p^2} \rtimes_\varphi \mathbb{Z}_p)$ for some nontrivial $\varphi : \mathbb{Z}_p \to \text{Aut}(\mathbb{Z}_{p^2})$. Now

$$\text{Aut}(\mathbb{Z}_{p^2}) \cong U_{p^2} \cong \mathbb{Z}_{p^2-p}$$

by Theorem 31.9 (see page 452), so $\text{Aut}(\mathbb{Z}_{p^2})$ contains a unique subgroup of order p. (In fact, $\text{Aut}(\mathbb{Z}_{p^2}) = \langle \gamma \rangle$, where $\gamma([x]) = (1+p)[x]$.) Lemma 37.9 then shows that the mappings $\varphi : \mathbb{Z}_p \to \text{Aut}(\mathbb{Z}_{p^2})$ will all produce isomorphic groups, and so there is just one non-Abelian group G of order p^3 if G contains an element of order p^2.

Case 2: Every nonidentity element in G has order p. Every p-group contains normal subgroups of any order dividing the order of the group (see Exercise (25) on page 433 of Investigation 29), so let H be a normal subgroup of G of order p^2. Since no element in G has order p^2, it must be the case that $H \cong (\mathbb{Z}_p \oplus \mathbb{Z}_p)$. Let a and b be generators for H so that $|a| = |b| = p$ and $H = \langle a \rangle \langle b \rangle$. Let $k \in G$ with $k \notin H$, and let $K = \langle k \rangle$. Again, $H \cap K = \{e\}$ and so $G = HK$. Thus, $G \cong (H \rtimes_\varphi K)$ for some nontrivial $\varphi : K \to \text{Aut}(H)$. (Note that $|k| = p$.) Now $\text{Aut}(H) \cong \text{Aut}(\mathbb{Z}_p \oplus \mathbb{Z}_p) \cong \text{GL}_2(\mathbb{Z}_p)$, and $|\text{Aut}(H)| = p^4 - p^3 - p^2 + p = p(p^3 - p^2 - p + 1)$. (See Exercise (10).) The Sylow p-subgroups of $\text{Aut}(H)$ all have order p, so any two subgroups of $\text{Aut}(H)$ of order p are conjugate.

Define $\gamma \in \text{Aut}(H) = \langle a \rangle \langle b \rangle$ by $\gamma(a) = ab$ and $\gamma(b) = b$. Since every non-identity element in H has order p, it follows that $|\gamma| = p$, and so any Sylow p-subgroup of $\text{Aut}(H)$ is conjugate to $\gamma(K)$. (We can also represent γ as an element in $\text{GL}_2(\mathbb{Z}_p)$ as $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.) Lemma 37.9 again shows that there is exactly one non-Abelian group of order p^3 of this type. This group is the Heisenberg group we introduced in Exercise (9) of Investigation 26. (See page 387.)

So the groups of order p^3, for p an odd prime, are:

- $\mathbb{Z}_{p^3}$;
- $\mathbb{Z}_{p^2} \oplus \mathbb{Z}_p$;
- $\mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_p$;
- $\mathbb{Z}_{p^2} \rtimes_\varphi \mathbb{Z}_p$, where $\varphi([1]) = \gamma$ with $\gamma([x]) = (1+p)[x]$; and
- $(\mathbb{Z}_p \oplus \mathbb{Z}_p) \rtimes_\varphi \mathbb{Z}_p$, where $\varphi([1]) = \gamma$ with γ represented by the matrix transformation $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.

Concluding Activities

Activity 37.10. Let H and K be finite groups, and let φ_1 and φ_2 be homomorphisms from K to Aut(H). It can be a difficult task to determine if $H \rtimes_{\varphi_1} K$ is isomorphic to $H \rtimes_{\varphi_2} K$.

(a) If φ_1 and φ_2 act in a similar way on elements of K, then we might expect the corresponding semidirect products to be isomorphic. Show that if $\varphi_1 = \varphi_2\theta$ for some $\theta \in$ Aut(K), then $(H \rtimes_{\varphi_1} K) \cong (H \rtimes_{\varphi_2} K)$.

(b) If φ_1 and φ_2 are significantly different in some way, we should expect that $H \rtimes_{\varphi_1} K$ is not isomorphic to $H \rtimes_{\varphi_2} K$. In this part we will show that if $\gcd(|H|, |K|) = 1$ and Ker$(\varphi_1) \ncong$ Ker(φ_2), then $H \rtimes_{\varphi_1} K$ is not isomorphic to $H \rtimes_{\varphi_2} K$. Let e_H be the identity in H, and let e_K be the identity in K. Then let $H' = \{(h, e_K) : h \in H\}$ and $K' = \{(e_H, k) : k \in K\}$ be subgroups of $H \rtimes_{\varphi_1} K$ that are copies of H and K. For $h \in H$, let $h' = (h, e_K)$ and for $k \in K$, let $k' = (e_H, k)$.

(i) Show that Ker$(\varphi_1) = \{x \in K : x'h'(x')^{-1} = h' \text{ for all } h' \in H'\}$. In other words, Ker$(\varphi_1) \cong C_{K'}(H')$.

(ii) Assume that $\gcd(|H|, |K|) = 1$ throughout the remainder of this exercise. Show that H' is the only subgroup of order $|H|$ in $G_1 = H \rtimes_{\varphi_1} K$.

(iii) Prove that if $H \rtimes_{\varphi_1} K$ is isomorphic to $H \rtimes_{\varphi_2} K$, then Ker$(\varphi_1) \cong$ Ker(φ_2). This will show that if Ker$(\varphi_1) \ncong$ Ker(φ_2), then $(H \rtimes_{\varphi_1} K) \ncong (H \rtimes_{\varphi_2} K)$.

Exercises

(1) Let $H = \mathbb{Z}_n$ and $K = \mathbb{Z}_2$. Let $\varphi : K \to$ Aut(H) be defined by $\varphi([1]_2) = \pi$, where $\pi([x]) = -[x]$ for all $x \in H$. Show that $(H \rtimes_\varphi K) \cong D_n$.

(2) Can the group $\mathbf{Q} = \{\pm 1, \pm i, \pm j, \pm k\}$ of quaternions be written as a semidirect product of proper subgroups? Explain.

⋆ (3) **Direct sums and semidirect products.** The direct sum of two groups is a special case of the semidirect product, as this exercise illustrates. Let H and K be groups with identities e_H and e_K, respectively, and let $\varphi : K \to$ Aut(H) be a homomorphism. Show that the following are equivalent.

(a) The map $\Phi : (H \rtimes_\varphi K) \to (H \oplus K)$ defined by $\Phi((h, k)) = hk$ is an isomorphism.

(b) The homomorphism φ is the trivial homomorphism.

(c) The subgroup $K' = \{(e_H, k) : k \in K\}$ is normal in $H \rtimes_\varphi K$.

⋆ (4) Prove the remaining items from Theorem 37.5. That is, prove that

(ii) $K' = \{(e_H, k) : k \in K\}$ is a subgroup of G isomorphic to K and

(iii) $H' \cap K' = \{(e_H, e_K)\}$.

⋆ (5) Complete the classification of groups of order 12 by showing that $(\mathbb{Z}_3 \rtimes_{\varphi_2} (\mathbb{Z}_2 \oplus \mathbb{Z}_2)) \cong D_3$ and that $(\mathbb{Z}_3 \rtimes_{\varphi_3} (\mathbb{Z}_2 \oplus \mathbb{Z}_2)) \cong D_3$, where φ_2 and φ_3 are as described in that section.

⋆ (6) When is $H \rtimes_\varphi K$ an Abelian group? Prove your answer. (Hint: Exercise (3) might be helpful.)

⋆ (7) Let G be a finite group of order n, and let p be the smallest prime divisor of n. Prove that any subgroup of index p in G is normal in G. (Hint: Consider the homomorphism $\pi : G \to P(G/N)$ defined by $\pi(a)(gN) = (ag)N$, where N is a subgroup of G of index p.)

⋆ (8) Let G be any group, and let $x, y \in G$ so that x and y commute with $[x, y] = x^{-1}y^{-1}xy$. Prove that

$$x^n y^n = (xy)^n [x, y]^{n(n-1)/2}$$

for every nonnegative integer n.

(9) **Groups of order pq.** We have previously shown that any group of order pq is cyclic if p and q are distinct primes and p does not divide $q - 1$. We will now classify the rest of the groups of order pq. Let p and q be primes with $p < q$ so that p divides $q - 1$, and let G be a group of order pq. Determine all of the groups to which G could be isomorphic.

⋆ (10) Let p be a prime. Explain why $\mathrm{Aut}(\mathbb{Z}_p \oplus \mathbb{Z}_p) \cong \mathrm{GL}_2(\mathbb{Z}_p)$. Then show that the order of $\mathrm{GL}_2(\mathbb{Z}_p)$ is $p^4 - p^3 - p^2 + p$. (Hint: Use the result from linear algebra that a 2×2 matrix is invertible if and only if no row is a multiple of the other.)

(11) Our definition of a semidirect product requires two groups. In this exercise, we will introduce a useful construction of a semidirect product that uses only one group. Let H be any group, and let $K = \mathrm{Aut}(H)$. Define $\varphi : K \to \mathrm{Aut}(H)$ by $\varphi(\pi) = \pi$. The resulting semidirect product $H \rtimes_\varphi K$ is called the **holomorph** of H and is denoted $\mathrm{Hol}(H)$. Holomorphs provide a context in which to study elements of a group and their automorphisms together.

(a) Let $n \geq 2$ be an integer. Find $|\mathrm{Hol}(\mathbb{Z}_n)|$ in terms of the prime factors of n. (Hint: See Exercise (13) on page 461 of Investigation 31.)

(b) Create the operation table for $\mathrm{Hol}(\mathbb{Z}_3)$. To which familiar group is $\mathrm{Hol}(\mathbb{Z}_3)$ isomorphic?

(c) Let $H = \mathbb{Z}_2 \oplus \mathbb{Z}_2$, $K = \mathrm{Aut}(H)$, and $G = \mathrm{Hol}(H)$. Show that $G \cong S_4$ using the following steps.

(i) Determine the order of $\mathrm{Hol}(\mathbb{Z}_2 \oplus \mathbb{Z}_2)$

(ii) For ease of notation, we will identify K with the subgroup of G that is isomorphic to K. Note that $[G : K] = 4$, so the permutation group $P(G/K)$ of the left cosets of K in G is isomorphic to S_4. Now define $\theta : G \to P(G/K)$ by $\theta(g)(aK) = (ga)K$. We have shown that θ is a homomorphism, so if we can show that θ is a bijection, then we will have that $G \cong S_4$. Prove that θ is a monomorphism. How can we conclude that $G \cong S_4$?

⋆ (12) Let t, m, and n be integers such that m divides n and $\gcd(t, m) = 1$. Prove that there exists an integer t' such that $t' \equiv t \pmod{m}$ and $\gcd(t', n) = 1$.

(13) (a) There are two distinct semidirect products of the form $\mathbb{Z}_4 \rtimes \mathbb{Z}_2$. What are they? Construct the operation table for the non-Abelian one. To which familiar group is this non-Abelian semidirect product isomorphic?

(b) Let $n \geq 3$ be an integer. Show that $D_n \cong (\mathbb{Z}_n \rtimes_\varphi \mathbb{Z}_2)$ for some appropriate choice of φ.

(14) **Groups of order pq^2.**

(a) Construct a non-Abelian group of order 75.

(b) Determine all non-Abelian groups of order 75.

(c) Determine all groups of order pq^2, where p and q are primes with $p < q$ such that p does not divide $q - 1$.

(15) Classify all groups of order 20.

(16) **Groups of order $2p^2$.** Let p be an odd prime. In this exercise, we work on the general problem of classifying all groups of order $2p^2$.

(a) Use the steps suggested below to determine the conjugacy classes of the elements of order 2 in $\mathrm{GL}_2(\mathbb{F}_p)$.

(i) Show that every element of order 2 or less in $\mathrm{GL}_2(\mathbb{F}_p)$ is conjugate to a diagonal matrix with 1's and/or -1's along the diagonal.

(ii) If A is an element of order 2 in $\mathrm{GL}_2(\mathbb{F}_p)$, show that A is conjugate to either $-I$ or the matrix $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

(b) Classify all groups of order $2p^2$. Note that there are three non-Abelian groups, and you may not be able to easily distinguish them. You might try to show that they have different centers. Part (a) of this exercise and Activity 37.10 should be useful.

Connections

This investigation continued our classification of groups of various orders that we began in Investigation 26. The tools we used in this investigation were mostly familiar, with the exception of the new construction of the semidirect product, an extension of the direct product first discussed in Investigation 25.

Part I

Appendices

Appendix A

Functions

Focus Questions

By the end of this investigation, you should be able to give precise and thorough answers to the questions listed below. You may want to keep these questions in mind to focus your thoughts as you complete the investigation.

- What is a function?
- What does it mean to say that a function is an injection? How can we prove that a function is (or is not) an injection?
- What does it mean to say that a function is a surjection? How can we prove that a function is (or is not) a surjection?
- What is a bijection?
- What is the composition of two functions, and what is a composite function? What are some important theorems about composite functions?
- What is the inverse of a function? Under what conditions is the inverse of a function $f : A \to B$ a function from B to A?
- What are some important theorems about functions and their inverses?

Functions are frequently used in mathematics to define and describe certain relationships between sets and other mathematical objects. In this appendix, we will first study special types of functions known as injections and surjections. Before defining these types of functions, we will review the definition of a function and explore certain functions with finite domains.

Definition A.1. A function f from a set A to a set B is a collection of ordered pairs

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

such that for each element a in A, there is one and only one element in B such that (a, b) is in f.

There is a special notation, called *functional notation*, that is commonly used to describe functions and the way they act on sets. In particular, if (a, b) is in the function f, we write $f(a) = b$ (read as "f of a equals b"). It is important to note the dual use of the symbol f here; we use f to represent a collection of ordered pairs and also to describe an action (pairing a with b in $f(a) = b$). In general practice, we use functional notation and think of a function as assigning to an element a a unique element b. In this context, we think of the elements in A as the input of the assignment and the elements in B as the output. In this way, we can consider f as a *mapping* from A to B and write

$f : A \to B$ to indicate this mapping action. How we read this notation depends on the context in which it appears. For instance, the statement

Consider the function $f : A \to B$

would be read as, "Consider the function f from A to B." On the other hand, if we write

Let $f : A \to B$,

then this statement would be read as "Let f be a function from A to B" or "Let f map from A to B."

There is some familiar terminology and notation associated with functions. Let f be a function from a set A to a set B.

- The set A is called the **domain** of f, and we write $\text{dom}(f) = A$.
- The set B is called the **codomain** of f, and we write $\text{codom}(f) = B$.
- The subset $\{f(a) : a \in A\}$ of B is called the **range** of f, which we denote by $\text{range}(f)$. Note that the range of f could equivalently be defined as follows:

$$\text{range}(f) = \{y \in B \mid y = f(x) \text{ for some } x \in A\}.$$

- If $a \in A$, then $f(a)$ is the **image** of a under f.
- If $b \in B$ and $b = f(a)$ for some $a \in A$, then a is called a **pre-image** of b.

Notice that, according to these definitions, $\text{range}(f) \subseteq \text{codom}(f)$, but it is not necessarily the case that $\text{range}(f) = \text{codom}(f)$. Whether we have this set equality or not depends on the function f, as we will see in the next section.

Special Types of Functions: Injections and Surjections

Preview Activity A.2. Let $A = \{1, 2, 3\}$, $B = \{a, b, c, d\}$, and $C = \{s, t\}$. Define

$f : A \to B$ by	$g : A \to B$ by	$h : A \to C$ by
$f(1) = a$	$g(1) = a$	$h(1) = s$
$f(2) = b$	$g(2) = b$	$h(2) = t$
$f(3) = c$	$g(3) = a$	$h(3) = s$

(a) Consider the following property, defined for an arbitrary function F:

For all $x, y \in \text{dom}(F)$, if $x \neq y$, then $F(x) \neq F(y)$.

Which of the functions defined above satisfy this property?

(b) Which of the functions defined above satisfy the following property (defined in terms of an arbitrary function F)?

For all $x, y \in \text{dom}(F)$, if $F(x) = F(y)$, then $x = y$.

(c) Determine the range of each of the functions f, g, and h.

(d) Which of these functions have their range equal to their codomain?

(e) Which of the these functions satisfy the following property (again, defined in terms of an arbitrary function F)?

For all y in the codomain of F, there exists an $x \in \text{dom}(F)$ such that $F(x) = y$.

(f) Let F be a function from a set S to a set T.

(i) Is it possible to have two elements x_1 and x_2 in S with $x_1 \neq x_2$ and $F(x_1) = F(x_2)$? If no, explain why not. If yes, give an example and explain why this does not violate the definition of a function.

(ii) Are the range and codomain of a function the same or different? If they are the same, explain why. If different, give an example to illustrate the difference and explain any relationships that must exist between the two sets.

Preview Activity A.3. Let A and B be nonempty sets, and let $f : A \to B$. In Preview Activity A.2, we determined whether or not certain functions satisfied some specified properties. These properties were written in the form of statements, and we will now examine these statements in more detail.

(a) Consider the following statement:

For all $x, y \in A$, if $x \neq y$, then $f(x) \neq f(y)$.

(i) Write the contrapositive of this conditional statement.

(ii) Write the negation of this conditional statement.

(b) Now consider the following statement:

For all $y \in B$, there exists an $x \in A$ such that $f(x) = y$.

Write the negation of this statement.

(c) Let $g : \mathbb{R} \to \mathbb{R}$ be defined by $g(x) = 5x + 3$, for all $x \in \mathbb{R}$. Complete the proofs of the following propositions about the function g.

Proposition 1. For all $a, b \in \mathbb{R}$, if $g(a) = g(b)$, then $a = b$.

Proof. Let $a, b \in \mathbb{R}$, and assume that $g(a) = g(b)$. We will prove that $a = b$. Since $g(a) = g(b)$, we know that

$$5a + 3 = 5b + 3.$$

(Now prove that in this situation, $a = b$.)

Proposition 2. For all $b \in \mathbb{R}$, there exists an $a \in \mathbb{R}$ such that $g(a) = b$.

Proof. Let $b \in \mathbb{R}$. We will prove that there exists an $a \in \mathbb{R}$ such that $g(a) = b$ by constructing such an a in $\mathbb{R}$. In order for this to happen, we need $g(a) = 5a + 3 = b$.

(Now solve the equation for a, and then show that for this real number a, $g(a) = b$.)

Injections

We have now seen examples of functions for which there exist different inputs that produce the same output. Using more formal notation, this means that there are functions $f : A \to B$ for which there exist $x_1, x_2 \in A$ with $x_1 \neq x_2$ and $f(x_1) = f(x_2)$. The work in the preview activities was intended to motivate the following definition.

Definition A.4. Let $f : A \to B$ be a function from the set A to the set B. The function f is called an **injection** provided that

$$\text{for all } x_1, x_2 \in A, \text{ if } x_1 \neq x_2, \text{ then } f(x_1) \neq f(x_2).$$

When f is an injection, we also say that f is a **one-to-one function**, or that f is an **injective function**.

Notice that the condition that specifies that a function f is an injection is given in the form of a conditional statement. As we will see, in proofs it is usually easier to use the contrapositive of this conditional statement. Although we did not define the term then, we have already written the contrapositive for the conditional statement in the definition of an injection in part (a) of Preview Activity A.3. In that preview activity, we also wrote the negation of the definition of an injection. The box below summarizes this work by giving the conditions that are equivalent to f being an injection or not being an injection.

Let $f : A \to B$.

"The function f is an injection" means that

- for all $x_1, x_2 \in A$, if $x_1 \neq x_2$, then $f(x_1) \neq f(x_2)$; or
- for all $x_1, x_2 \in A$, if $f(x_1) = f(x_2)$, then $x_1 = x_2$.

"The function f is not an injection" means that

- there exist $x_1, x_2 \in A$ such that $x_1 \neq x_2$ and $f(x_1) = f(x_2)$.

Activity A.5. Now that we have defined what it means for a function to be an injection, we can see that in part (c) of Preview Activity A.3, we proved that the function $g : \mathbb{R} \to \mathbb{R}$, where $g(x) = 5x+3$ for all $x \in \mathbb{R}$, is an injection. Use the definition (or its negation) to determine whether or not the following functions are injections.

(a) $k : A \to B$, where $A = \{a, b, c\}$, $B = \{1, 2, 3, 4\}$, and $k(a) = 4$, $k(b) = 1$, and $k(c) = 3$

(b) $f : A \to C$, where $A = \{a, b, c\}$, $C = \{1, 2, 3\}$, and $f(a) = 2$,$f(b) = 3$, and $f(c) = 2$

(c) $F : \mathbb{Z} \to \mathbb{Z}$ defined by $F(m) = 3m + 2$ for all $m \in \mathbb{Z}$

(d) $h : \mathbb{R} \to \mathbb{R}$ defined by $h(x) = x^2 - 3x$ for all $x \in \mathbb{R}$

(e) $s : \mathbb{Z}_5 \to \mathbb{Z}_5$ defined by $s(x) = x^3$ for all $x \in \mathbb{Z}_5$

Surjections

In previous mathematics courses and in Preview Activity A.2, we have seen that there exist functions $f : A \to B$ for which the codomain and range of f are equal—that is, range$(f) = B$. This means that every element of B is an output of the function f for some input from the set A. Using quantifiers, this means that for every $y \in B$, there exists an $x \in A$ such that $f(x) = y$. One of the objectives of the preview activities was to motivate the following definition:

Definition A.6. Let $f : A \to B$ be a function from the set A to the set B. The function f is called a **surjection** provided that the range of f equals the codomain of f. This means that

$$\text{for every } y \in B, \text{ there exists an } x \in A \text{ such that } f(x) = y.$$

When f is a surjection, we also say that f is an **onto function**, that f maps A **onto** B, or that f is a **surjective function**.

Note that the main condition defining what it means for a function f to be a surjection is given in the form of a universally quantified statement. Although we did not define the term then, we have already written the negation of this statement defining a surjection in part (b) of Preview Activity A.3. The box below summarizes the conditions for f being a surjection or not being a surjection.

Let $f : A \to B$.

"The function f is a surjection" means that

- range$(f) =$ codom$(f) = B$; or
- for every $y \in B$, there exists an $x \in A$ such that $f(x) = y$.

"The function f is not a surjection" means that

- range$(f) \neq$ codom(f); or
- there exists a $y \in B$ such that for all $x \in A$, $f(x) \neq y$.

Activity A.7. Now that we have defined what it means for a function to be a surjection, we can see that in part (c) of Preview Activity A.3, we proved that the function $g : \mathbb{R} \to \mathbb{R}$, where $g(x) = 5x+3$ for all $x \in \mathbb{R}$, is a surjection. Determine whether or not the following functions are surjections. Are any of these functions injections?

(a) $k : A \to B$, where $A = \{a, b, c\}$, $B = \{1, 2, 3, 4\}$, and $k(a) = 4$, $k(b) = 1$, and $k(c) = 3$.

(b) $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 3x + 2$ for all $x \in \mathbb{R}$.

(c) $F : \mathbb{Z} \to \mathbb{Z}$ defined by $F(m) = 3m + 2$ for all $m \in \mathbb{Z}$.

(d) $s : \mathbb{Z}_5 \to \mathbb{Z}_5$ defined by $s(x) = x^3$ for all $x \in \mathbb{Z}_5$.

Another important class of functions are those that are both injective and surjective. Any such function is called a *bijection*.

Definition A.8. A **bijection** is a function that is both an injection and a surjection. If the function f is a bijection, we also say that f is **one-to-one and onto** and that f is a **bijective function**.

Activity A.9. Which of the functions in Activity A.7 are bijections?

The Importance of the Domain and Codomain

The functions in the next activity will illustrate why the domain and the codomain are just as important as the rule defining the outputs when we are trying to determine if a given function is injective and/or surjective.

Activity A.10. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2 + 1$. Notice that

$$f(2) = 5 \text{ and } f(-2) = 5.$$

This observation is enough to prove that the function f is not an injection since we can see that there exist two different inputs that produce the same output.

Since $f(x) = x^2 + 1$, we know that $f(x) \geq 1$ for all $x \in \mathbb{R}$. This implies that the function f is not a surjection. For example, -2 is in the codomain of f and $f(x) \neq -2$ for all x in the domain of f.

(a) Now let $T = \{y \in \mathbb{R} \mid y \geq 1\}$, and define $F : \mathbb{R} \to T$ by $F(x) = x^2 + 1$. Notice that the function F uses the same formula as the function f and has the same domain as f, but has a different codomain than f.

  (i) Explain why F is not an injection.

  (ii) Is F a surjection? Justify your conclusion.

(b) Let $\mathbb{Z}^* = \{x \in \mathbb{Z} \mid x \geq 0\} = \mathbb{N} \cup \{0\}$. Define $g : \mathbb{Z}^* \to \mathbb{N}$ by $g(x) = x^2 + 1$. (Notice that this is the same formula used in part (a).)

  (i) Calculate $g(0), g(1), g(2), g(3), g(4)$, and $g(5)$. Based on this information, does the function g appear to be an injection? Does the function g appear to be a surjection?

  (ii) Is the function g an injection? Justify your conclusion with a proof or a counterexample.

  (iii) Is the function g a surjection? Justify your conclusion with a proof of a counterexample.

In Activity A.10, the same mathematical formula was used to determine the outputs for the functions. However:

- One of the functions was neither an injection nor a surjection.
- Another one of the functions was not an injection but was a surjection.
- The third function was an injection but was not a surjection.

This illustrates the important fact that whether a function is injective or surjective not only depends on the formula that defines the output of the function but also on the domain and codomain of the function.

Composition of Functions

The basic idea of function composition is that, when possible, the output of a function f is used as the input of a function g. The resulting function can be referred to as "f followed by g" and is called the composition of f with g. For example, if

$$f(x) = 3x^2 + 2 \quad \text{and} \quad g(x) = \sin(x),$$

then we can compute $g(f(x))$ as follows:

$$\begin{aligned} g(f(x)) &= g\big(3x^2+2\big) \\ &= \sin\big(3x^2+2\big)\,. \end{aligned}$$

In this case, $f(x)$, the output of the function f, was used as the input for the function g. This idea motivates the formal definition of the composition of two functions.

Definition A.11. Let A, B, and C be nonempty sets, and let $f : A \to B$ and $g : B \to C$ be functions. The **composition of f and g** is the function $g \circ f : A \to C$ defined by

$$(g \circ f)(x) = g\,(f(x))$$

for all $x \in A$. We often refer to the function $g \circ f$ as a **composite function**.

Activity A.12. Let $A = \{1, 2, 3\}$, $B = \{a, b, c, d\}$, and $C = \{s, t\}$. Define $f : A \to B$ by

$$f(1) = a, f(2) = b, f(3) = c,$$

$g : A \to B$ by

$$g(1) = c, g(2) = d, g(3) = c,$$

and $h : B \to C$ by

$$h(a) = s, h(b) = s, h(c) = t, h(d) = s.$$

(a) Find the images of the elements in A under the function $h \circ f$.

(b) Find the images of the elements in A under the function $h \circ g$.

(c) Is $h \circ f$ an injection? Is $h \circ f$ a surjection? Explain.

(d) Is $h \circ g$ an injection? Is $h \circ g$ a surjection? Explain.

In Activity A.12, we asked questions about whether certain composite functions were injections and/or surjections. In mathematics, it is typical to explore whether certain properties of an object transfer to related objects. In particular, we might want to know whether or not the composite of two injective functions is also an injection. (Of course, we could ask a similar question for surjections.) These types of questions are explored in the next activity.

Activity A.13. Let the sets A, B, C, and D be as follows:

$$A = \{a, b, c\}, \quad B = \{p, q, r\}, \quad C = \{u, v, w, x\}, \quad \text{and} \quad D = \{u, v\}.$$

(a) Construct a function $f : A \to B$ that is an injection and a function $g : B \to C$ that is an injection. In this case, is the composite function $g \circ f : A \to C$ an injection? Explain.

(b) Construct a function $f : A \to B$ that is a surjection and a function $g : B \to D$ that is a surjection. In this case, is the composite function $g \circ f : A \to D$ a surjection? Explain.

(c) Construct a function $f : A \to B$ that is a bijection and a function $g : B \to A$ that is a bijection. In this case, is the composite function $g \circ f : A \to A$ a bijection? Explain.

In Activity A.13, we explored some properties of composite functions related to injections, surjections, and bijections. The following theorem summarizes the results that these explorations were intended to illustrate.

Theorem A.14. *Let A, B, and C be nonempty sets, and assume that $f : A \to B$ and $g : B \to C$.*

(i) *If f and g are both injections, then $(g \circ f) : A \to C$ is an injection.*

(ii) *If f and g are both surjections, then $(g \circ f) : A \to C$ is a surjection.*

(iii) *If f and g are both bijections, then $(g \circ f) : A \to C$ is a bijection.*

The proof of part (i) is Exercise 4, and part (iii) is a direct consequence of the first two parts. Therefore, we will focus here on constructing a proof of part (ii). Our goal is to prove that $g \circ f$ is a surjection. Since $g \circ f : A \to C$, this is equivalent to proving that

$$\text{for all } c \in C, \text{ there exists an } a \in A \text{ such that } (g \circ f)(a) = c.$$

Thus, we need to find an $a \in A$ such that $(g \circ f)(a) = c$.

Now we can look at the hypotheses. In particular, we are assuming that both $f : A \to B$ and $g : B \to C$ are surjections. Since we have chosen $c \in C$, and $g : B \to C$ is a surjection, we know that there exists a $b \in B$ such that $g(b) = c$. Now, $b \in B$ and $f : A \to B$ is a surjection. Therefore, there exists an $a \in A$ such that $f(a) = b$. If we now compute $(g \circ f)(a)$, we will see that

$$(g \circ f)(a) = g\left(f(a)\right) = g(b) = c.$$

We can now write the complete proof as follows:

Proof of Theorem A.14, part (ii). Let A, B, and C be nonempty sets, and assume that $f : A \to B$ and $g : B \to C$ are both surjections. We will prove that $g \circ f : A \to C$ is a surjection.

Let c be an arbitrary element of C. We will prove there exists an $a \in A$ such that $(g \circ f)(a) = c$. Since $g : B \to C$ is a surjection, it follows that there exists a $b \in B$ such that $g(b) = c$. Now $b \in B$ and $f : A \to B$ is a surjection. Hence, there exists an $a \in A$ such that $f(a) = b$. We now see that

$$\begin{aligned}(g \circ f)(a) &= g\left(f(a)\right) \\ &= g(b) \\ &= c.\end{aligned}$$

We have therefore shown that for every $c \in C$, there exists an $a \in A$ such that $(g \circ f)(a) = c$. This proves that $g \circ f$ is a surjection. ■

Inverse Functions

Now that we have studied composite functions, we will move on to consider another important idea: the inverse of a function. In order to study inverse functions, we will need to use the concept of the

Cartesian product of two sets A and B, denoted by $A \times B$, which is the set of all ordered pairs (x, y) where $x \in A$ and $y \in B$. That is,

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

In previous mathematics courses, you probably learned that the exponential function (with base e) and the natural logarithm functions are inverses of each other. You may have seen this relationship expressed as follows:

$$\begin{gathered}\text{For each } x \in \mathbb{R} \text{ with } x > 0 \text{ and for each } y \in \mathbb{R}, \\ y = \ln(x) \text{ if and only if } x = e^y.\end{gathered}$$

Notice that x is the input and y is the output for the natural logarithm function if and only if y is the input and x is the output for the exponential function. In essence, the inverse function (in this case, the exponential function) reverses the action of the original function (in this case, the natural logarithm function). In terms of ordered pairs (input-output pairs), this means that if (x, y) is an ordered pair for a function, then (y, x) is an ordered pair for its inverse. The idea of reversing the roles of the first and second coordinates is the basis for our definition of the inverse of a function.

Definition A.15. Let $f : A \to B$ be a function. The **inverse** of *f*, denoted by f^{-1}, is the set of ordered pairs $\{(b, a) \in B \times A \mid f(a) = b\}$. That is,

$$f^{-1} = \{(b, a) \in B \times A : f(a) = b\}.$$

If we use the ordered pair representation for *f*, we could also write

$$f^{-1} = \{(b, a) \in B \times A : (a, b) \in f\}.$$

Notice that this definition does not state that f^{-1} is a function. Rather, f^{-1} is simply a subset of $B \times A$. In Activity A.16, we will explore the conditions under which the inverse of a function $f : A \to B$ is itself a function from B to A.

Activity A.16. Let $A = \{a, b, c\}$, $B = \{a, b, c, d\}$, and $C = \{p, q, r\}$. Define

$f : A \to C$ by	$g : A \to C$ by	$h : B \to C$ by
$f(a) = r$	$g(a) = p$	$h(a) = p$
$f(b) = p$	$g(b) = q$	$h(b) = q$
$f(c) = q$	$g(c) = p$	$h(c) = r$
		$h(d) = q$

(a) Determine the inverse of each function as a set of ordered pairs.

(b) (i) Is f^{-1} a function from C to A? Explain.

(ii) Is g^{-1} a function from C to A? Explain.

(iii) Is h^{-1} a function from C to B? Explain.

(c) Make a conjecture about what conditions on a function $F : S \to T$ will ensure that its inverse is a function from T to S.

We will now consider a general argument suggested by the explorations in Activity A.16. By definition, if $f : A \to B$ is a function, then f^{-1} is a subset of $B \times A$. However, f^{-1} may or may

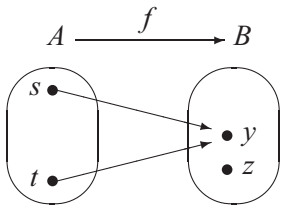

Figure A.1
The inverse is not a function.

not be a function from B to A. For example, suppose that $s, t \in A$ with $s \neq t$ and $f(s) = f(t)$ (as illustrated in Figure A.1).

In this case, if we try to reverse the arrows, we will not get a function from B to A. This is because $(y, s) \in f^{-1}$ and $(y, t) \in f^{-1}$ with $s \neq t$. Consequently, f^{-1} is not a function. This observation suggests that if f is not an injection, then f^{-1} is not a function.

Also, if f is not a surjection, then there exists a $z \in B$ such that $f(a) \neq z$ for all $a \in A$, as in the diagram in Figure A.1. In other words, there is no ordered pair in f with z as the second coordinate. This means that there would be no ordered pair in f^{-1} with z as a first coordinate. Consequently, f^{-1} cannot be a function from B to A.

Theorem A.17 formalizes these observations. In the proof of the theorem, we will use both the input-output representation and the ordered pair representation of a function. The idea is that if $G : S \to T$ is a function, then for $s \in S$ and $t \in T$,

$$G(s) = t \text{ if and only if } (s, t) \in G.$$

When we use the ordered pair representation of a function, we will also use the ordered pair representation of its inverse. In this case, we know that

$$(s, t) \in G \text{ if and only if } (t, s) \in G^{-1}.$$

Theorem A.17. *Let A and B be nonempty sets, and let $f : A \to B$. The inverse of f is a function from B to A if and only if f is a bijection.*

Proof. Let A and B be nonempty sets, and let $f : A \to B$. We will first assume that f is a bijection and prove that f^{-1} is a function from B to A. To do this, we will show that f^{-1} satisfies the conditions of Definition A.1.

Let $b \in B$. Since the function f is a surjection, there exists an $a \in A$ such that $f(a) = b$. This implies that $(a, b) \in f$ and hence that $(b, a) \in f^{-1}$. Thus, each element of B is the first coordinate of an ordered pair in f^{-1}. We must now prove that each element of B is the first coordinate of exactly one ordered pair in f^{-1}. So let $b \in B$, $a_1, a_2 \in A$ and assume that

$$(b, a_1) \in f^{-1} \text{ and } (b, a_2) \in f^{-1}.$$

This means that $(a_1, b) \in f$ and $(a_2, b) \in f$. We can then conclude that

$$f(a_1) = b \text{ and } f(a_2) = b.$$

But this means that $f(a_1) = f(a_2)$. Since f is a bijection, f is by definition an injection, and we can conclude that $a_1 = a_2$. This proves that b is the first element of only one ordered pair in f^{-1}. Consequently, we have proved that f^{-1} satisfies the conditions of Definition A.1 and hence f^{-1} is a function from B to A.

We will now assume that f^{-1} is a function from B to A and prove that f is a bijection. First, to prove that f is an injection, we will assume that $a_1, a_2 \in A$ and that $f(a_1) = f(a_2)$. We wish to show that $a_1 = a_2$. If we let $b = f(a_1) = f(a_2)$, we can conclude that

$$(a_1, b) \in f \text{ and } (a_2, b) \in f.$$

But this means that

$$(b, a_1) \in f^{-1} \text{ and } (b, a_2) \in f^{-1}.$$

Since we have assumed that f^{-1} is a function, we can conclude that $a_1 = a_2$. Hence, f is an injection.

Now to prove that f is a surjection, we will choose an arbitrary $b \in B$ and show that there exists an $a \in A$ such that $f(a) = b$. Since f^{-1} is a function, b must be the first coordinate of some ordered pair in f^{-1}. Consequently, there exists an $a \in A$ such that

$$(b, a) \in f^{-1}.$$

Now this implies that $(a, b) \in f$, and so $f(a) = b$. This proves that f is a surjection. Since we have also proved that f is an injection, we can conclude that f is a bijection, as desired. ■

Theorems about Inverse Functions

In the situation where $f : A \to B$ is a bijection and f^{-1} is a function from B to A, we can write $f^{-1} : B \to A$. In this case, we frequently say that f is an **invertible function**, and we usually do not use the ordered pair representation for either f or f^{-1}. Instead of writing $(a, b) \in f$, we write $f(a) = b$, and instead of writing $(b, a) \in f^{-1}$, we write $f^{-1}(b) = a$. Using the fact that $(a, b) \in f$ if and only if $(b, a) \in f^{-1}$, we can now write $f(a) = b$ if and only if $f^{-1}(b) = a$. Theorem A.18 formalizes this observation.

Theorem A.18. *Let A and B be nonempty sets, and let $f : A \to B$ be a bijection. Then $f^{-1} : B \to A$ is a function, and for every $a \in A$ and $b \in B$,*

$$f(a) = b \text{ if and only if } f^{-1}(b) = a.$$

The next two results are two important theorems about inverse functions. The first can be considered to be a corollary of Theorem A.18.

Corollary A.19. *Let A and B be nonempty sets, and let $f : A \to B$ be a bijection. Then*

(i) *For every x in A, $\left(f^{-1} \circ f\right)(x) = x$.*

(ii) *For every y in B, $\left(f \circ f^{-1}\right)(y) = y$.*

Activity A.20. Prove Corollary A.19. For the first part, let $x \in A$, write $f(x) = y$, and then use the result in Theorem A.18.

We will now consider the case where $f : A \to B$ and $g : B \to C$ are both bijections. In this case, $f^{-1} : B \to A$ and $g^{-1} : C \to B$. Figure A.2 illustrates this situation.

By Theorem A.14, $g \circ f : A \to C$ is also a bijection. Hence, by Theorem A.17, $(g \circ f)^{-1}$ is a function and, in fact, $(g \circ f)^{-1} : C \to A$. Notice that we can also form the composition of g^{-1} followed by f^{-1} to get $f^{-1} \circ g^{-1} : C \to A$. Figure A.2 helps illustrate the result of the next theorem.

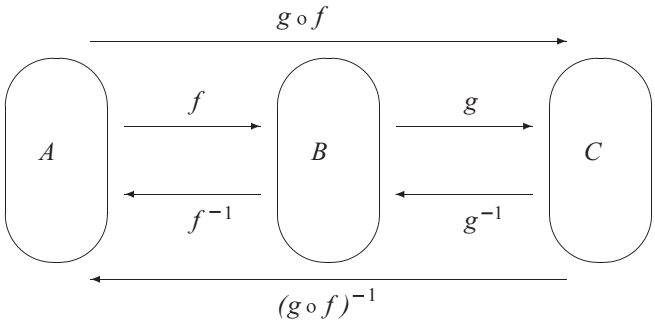

Figure A.2
Composition of two bijections.

Theorem A.21. *Let $f : A \to B$ and $g : B \to C$ be bijections. Then $g \circ f$ is a bijection and $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.*

Proof. Let $f : A \to B$ and $g : B \to C$ be bijections. Then $f^{-1} : B \to A$ and $g^{-1} : C \to B$. Hence, $f^{-1} \circ g^{-1} : C \to A$. Also, by Theorem A.14, $g \circ f : A \to C$ is a bijection, and hence $(g \circ f)^{-1} : C \to A$. We will now prove that for each $z \in C$, $(g \circ f)^{-1}(z) = (f^{-1} \circ g^{-1})(z)$.

Let $z \in C$. Since the function g is a surjection, there exists a $y \in B$ such that

$$g(y) = z. \tag{A.1}$$

Also, since f is a surjection, there exists an $x \in A$ such that

$$f(x) = y. \tag{A.2}$$

Now equations (A.1) and (A.2) can be written in terms of the respective inverse functions as

$$g^{-1}(z) = y \text{ and} \tag{A.3}$$
$$f^{-1}(y) = x. \tag{A.4}$$

Using equations (A.3) and (A.4), we see that

$$\begin{aligned} \left(f^{-1} \circ g^{-1}\right)(z) &= f^{-1}\left(g^{-1}(z)\right) \\ &= f^{-1}(y) \\ &= x. \end{aligned} \tag{A.5}$$

Using equations (A.1) and (A.2) again, we see that $(g \circ f)(x) = z$. However, in terms of the inverse function, this means that

$$(g \circ f)^{-1}(z) = x. \tag{A.6}$$

Comparing equations (A.5) and (A.6), we have shown that for all $z \in C$, $(g \circ f)^{-1}(z) = (f^{-1} \circ g^{-1})(z)$. This proves that $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$. ■

Concluding Activities

Activity A.22. Prove the following:

If $f : A \to B$ is a bijection, then $f^{-1} : B \to A$ is also a bijection.

Exercises

(1) For each of the following functions, determine if the function is an injection, a surjection, a bijection, or none of these. Justify all of your conclusions.

(a) $F : \mathbb{R} \to \mathbb{R}$ defined by $F(x) = 5x + 3$, for all $x \in \mathbb{R}$.

(b) $G : \mathbb{Z} \to \mathbb{Z}$ defined by $G(x) = 5x + 3$, for all $x \in \mathbb{Z}$.

(c) $f : (\mathbb{R} - \{4\}) \to \mathbb{R}$ defined by $f(x) = \dfrac{3x}{x-4}$, for all $x \in (\mathbb{R} - \{4\})$.

(d) $g : (\mathbb{R} - \{4\}) \to (\mathbb{R} - \{3\})$ defined by $g(x) = \dfrac{3x}{x-4}$, for all $x \in (\mathbb{R} - \{4\})$.

(2) Define $f : \mathbb{N} \to \mathbb{Z}$ as follows: For each $n \in \mathbb{N}$,

$$f(n) = \frac{1 + (-1)^n(2n-1)}{4}.$$

Is the function f an injection? Is the function f a surjection? Justify your conclusions.

Suggestions: Start by calculating several outputs for the function before you attempt to write a proof. In exploring whether or not the function is an injection, it might be a good idea to use cases based on whether the inputs are even or odd. In exploring whether f is a surjection, consider using cases based on whether the output is positive or less than or equal to zero.

(3) An operation $*$ on a set S is a function from $S \times S$ to S that assigns to the pair $(x, y) \in S \times S$ the element $x * y$ in S. For example, addition of integers can be defined as a function $f : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ that maps the pair $(a, b) \in \mathbb{Z} \times \mathbb{Z}$ to the integer $f(a, b) = a + b$.

(a) Is the function f an injection? Justify your conclusion.

(b) Is the function f a surjection? Justify your conclusion.

(4) Prove Part (i) of Theorem A.14:

Let A, B, and C be nonempty sets, and let $f : A \to B$ and $g : B \to C$. If f and g are both injections, then $g \circ f : A \to C$ is an injection.

(5) Suppose $f : A \to B$ and $g : B \to C$ are functions.

(a) Is it true that if $g \circ f$ is an injection, then both f and g are injections? If the answer is no, are there any conditions that f or g must satisfy to make $g \circ f$ an injection? Prove your answers.

(b) Is it true that if $g \circ f$ is a surjection, then both f and g are surjections? If the answer is no, are there any conditions that f or g must satisfy to make $g \circ f$ a surjection? Prove your answers.

(6) Is composition of functions a commutative operation? Prove your answer.

(7) Is composition of functions an associative operation? Prove your answer.

(8) (a) Define $f : \mathbb{Z}_5 \to \mathbb{Z}_5$ by $f([x]) = \left[x^2 + 4\right]$ for all $[x] \in \mathbb{Z}_5$. Write the inverse of f as a set of ordered pairs, and explain why f^{-1} is not a function.

(b) Define $g : \mathbb{Z}_5 \to \mathbb{Z}_5$ by $f([x]) = \left[x^3 + 4\right]$ for all $[x] \in \mathbb{Z}_5$. Write the inverse of g as a set of ordered pairs, and explain why g^{-1} is a function.

(c) Is it possible to write a formula for $g^{-1}([y])$, where $[y] \in \mathbb{Z}_5$? The answer to this question depends on whether or not it is possible to define a cube root of elements of $\mathbb{Z}_5$. Recall that for a real number x, we define the cube root of x to be the real number y such that $y^3 = x$. That is,

$$y = \sqrt[3]{x} \text{ if and only if } y^3 = x.$$

Using this idea, is it possible to define the cube root of each element of $\mathbb{Z}_5$? If so, what is $\sqrt[3]{[0]}$, $\sqrt[3]{[1]}$, $\sqrt[3]{[2]}$, $\sqrt[3]{[3]}$, and $\sqrt[3]{[4]}$.

(d) Now answer the question posed at the beginning of part (c). If possible, determine a formula for $g^{-1}([y])$ where $g^{-1} : \mathbb{Z}_5 \to \mathbb{Z}_5$.

Appendix B

Mathematical Induction and the Well-Ordering Principle

Focus Questions

By the end of this investigation, you should be able to give precise and thorough answers to the questions listed below. You may want to keep these questions in mind to focus your thoughts as you complete the investigation.

- What does the Principle of Mathematical Induction say? What do we need to verify in order to prove a statement using the Principle of Mathematical Induction?
- How do the extended and strong forms of induction differ from the Principle of Mathematical Induction? How are all of these different versions of induction similar?
- What does the Well-Ordering Principle say? What do we need to verify in order to prove a statement using the Well-Ordering Principle?
- How are the Principle of Mathematical Induction, the Extended Principle of Mathematical Induction, the Strong Form of Mathematical Induction, and the Well-Ordering Principle all related?

Preview Activity B.1. Suppose you are on a game show called *Let's Make a Great Deal*. You have reached the final round and will be asked one question. If you answer the question correctly, you will win a key to open door number 1. Behind door number 1 is a prize and a key to open door number 2. Behind door number 2 is a prize and a key to open door number 3. Behind door number 3 is a prize and a key to open door number 4, and so on.

(a) How many prizes will you win if you fail to answer the question correctly?

(b) Which prizes will you win if you answer the question correctly?

Introduction

Mathematical induction is an important tool in mathematics. Induction helps us prove that certain types of statements are true for *all* positive integers. This is quite a feat, since there are infinitely

many positive integers! Mathematical induction comes in more than one flavor. There is the basic principle, the extended principle, and the strong (or second, or complete) principle. An equivalent version of the Principle of Mathematical Induction is the Well-Ordering Principle. We will study each of these principles in this investigation.

The Principle of Mathematical Induction

Activity B.1 demonstrated the basic idea behind induction. Just as no prize comes for free (in our game, we needed to answer the question in order to win anything), to verify a statement using induction, we will need to prove something. In other words, we will need to unlock the door that has our first statement behind it. But unlocking the first door is not enough to unlock every other door, unless we are able to establish—as was specified in the rules of our game show—that each door, when opened, contains the key to the next door. If this condition also holds, then once we open the first door—that is, once we prove the first statement—we will be able to open every other door, thus proving our statement for every positive integer.

To illustrate this process in a more concrete way, consider the example in the following activity.

Activity B.2. Let n be a positive integer. Complete Table B.2. What do you notice?

n	$1+2+3+\cdots+n$	$\frac{n(n+1)}{2}$
1		
2		
3		
4		
5		

The calculations in Activity B.2 show that

$$1+2+3+\cdots+n=\frac{n(n+1)}{2} \tag{B.1}$$

for all integers n between 1 and 5. A few more calculations might convince you that equation (B.1) is actually true for many more integers, and perhaps for all positive integers. Although we cannot physically evaluate both sides of equation (B.1) to determine if it is true for every positive integer, we can use mathematical induction to accomplish the same goal.

To return to our game show analogy, think of verifying equation (B.1) as the goal of the game. Each door corresponds to one instance of equation (B.1). The first door corresponds to equation (B.1) with $n=1$, the second door to equation (B.1) with $n=2$, and so on. In general, for each positive integer m, the m^{th} door corresponds to equation (B.1) with $n=m$.

In this context, the question we need to answer to open door number one is whether equation (B.1) is true when $n=1$.

Activity B.3. Is equation (B.1) is true when $n=1$? Why?

Opening the first door is important, but it does not complete the problem of proving equation (B.1) for *all* positive integers. In the game, behind each door was a key to opening the next door. Of course, these keys were a critical part of the game. If one of the doors did not have a key to the next, then we wouldn't necessarily win all of the prizes just by opening the first door. In the same way, to complete our verification of equation (B.1), we will need to show that the first door ($n = 1$) contains the key to the second door ($n = 2$), the second door ($n = 2$) contains the key to the third door ($n = 3$), and so on.

Our calculations in Activity B.2 show that equation (B.1) is true for n from 1 to 5, but they don't demonstrate that each holds the key to the next. In other words, if equation (B.1) is true for $n = 1$, must it also be true when $n = 2$? And if equation (B.1) is true when $n = 2$, must it also be true when $n = 3$? And so on. In a nutshell, what we need to demonstrate is that if equation (B.1) is true for some arbitrary positive integer n, then it must also be true for the integer $n + 1$. This shows that each instance when equation (B.1) is true for a given positive integer n provides the key to proving that the equation is also true for the integer $n + 1$.

As an example, let's show that if equation (B.1) is true for $n = 1$, then it must also be true for $n = 2$. To do so, we will assume equation (B.1) is true when $n = 1$. That is, we will assume that

$$1 = \frac{(1)(2)}{2}. \tag{B.2}$$

Assuming equation (B.2) is true, we need to prove that equation (B.1) is true when $n = 2$, or that

$$1 + 2 = \frac{(2)(3)}{2}.$$

Since we are assuming equation (B.1) to be true when $n = 1$, we can begin with the true statement (B.2). Adding 2 to both sides of equation (B.2) yields

$$\begin{aligned} 1 + 2 &= \frac{(1)(2)}{2} + 2 \\ &= \frac{(1)(2) + 2(2)}{2} \\ &= \frac{(2)(1 + 2)}{2} \\ &= \frac{(2)(3)}{2}, \end{aligned}$$

which shows that equation (B.1) is true when $n = 2$ (assuming that the same equation is true when $n = 1$).

Activity B.4. To complete our proof of equation (B.1) for all positive integers n, we need to verify that whenever equation (B.1) for some arbitrary positive integer n, then it is also true for the integer $n + 1$.

(a) Continue, as above, to show that if equation (B.1) is true for $n = 2$, then it is also true for $n = 3$.

(b) Of course, we cannot continue showing each specific implication in turn, as that would take an infinite amount of time. To be more efficient, we really want to show that if n is any positive integer and equation (B.1) is true for n, then equation (B.1) is also true for $n+1$. To do this, we let n be an arbitrary positive integer and assume that equation (B.1) is true for n. That is, we assume that

$$1 + 2 + 3 + \cdots + n = \frac{(n)(n + 1)}{2}.$$

Use this assumption to show that equation (B.1) is true for $n + 1$.

To summarize, we needed to prove two things to show that equation (B.1) is true for all positive integers n:

(1) Equation (B.1) is true when $n = 1$; and

(2) whenever equation (B.1) is true for a positive integer n, then it is also true for the integer $n + 1$.

Step 1 is equivalent to answering the question in our game show and opening the first door. The prize is that equation (B.1) is true when $n = 1$. Step 2 verifies that each door holds the key to opening the next—that is, if equation (B.1) is true for the integer n, then it is also true for $n + 1$. Completing both steps shows that equation (B.1) is true for every positive integer n.

Let's now formalize the ideas from the previous example. In doing so, we will develop the Principle of Mathematical Induction, which can be used when we have a family of statements, one for each positive integer, that we want to prove. For example, for each positive integer n, let $P(n)$ be the statement that

$$1 + 2 + 2^2 + 2^3 + \cdots + 2^{n-1} = 2^n - 1 \tag{B.3}$$

To prove that $P(n)$ is true for all $n \in \mathbb{N}$, we have seen that we need to prove that:

(1) $P(1)$ is true (we call this the *base case*); and

(2) for every positive integer n, if $P(n)$ is true, then $P(n + 1)$ is also true. (This second step is called the *inductive step*.)

When we prove the inductive step, we assume $P(n)$ is true for some arbitrary positive integer n. This assumption is called the *induction hypothesis* or *inductive hypothesis*. We then show, using this assumption, that $P(n + 1)$ is also true.

Rephrasing this process in a slightly different form leads us to the formal statement of the Principle of Mathematical Induction. Let S be the set of positive integers for which $P(n)$ is true. Proving $P(1)$ is true is the same as showing that 1 is in S. Likewise, showing that $P(n)$ implies $P(n + 1)$ is equivalent to showing that $n + 1 \in S$ whenever $n \in S$. Combining these two observations, we arrive at the following axiom:

Axiom B.5 (Principle of Mathematical Induction)**.** *Let S be a subset of the set of natural numbers $\mathbb{N}$. If*

(i) *S contains 1 and*

(ii) *S contains the positive integer $n + 1$ whenever S contains n,*

then $S = \mathbb{N}$.

In essence, the Principle of Mathematical Induction tells us that if we have a set $S \subseteq \mathbb{N}$ containing 1, and if S contains the integer $n + 1$ whenever S contains n, then S must contain $1 + 1 = 2$. But then S must also contain $2 + 1 = 3$, and $3 + 1 = 4$, and so on. Therefore, S will contain *all* natural numbers. By using the Principle of Mathematical Induction, we can prove infinitely many statements in only two steps.

To illustrate, let's formally apply the Principle of Mathematical Induction to establish equation (B.3). Let

$$S = \{n \in \mathbb{N} : 1 + 2 + 2^2 + 2^3 + \cdots + 2^{n-1} = 2^n - 1\}.$$

To use the Principle of Mathematical Induction, we need to show that $1 \in S$ and that $n+1 \in S$ whenever $n \in S$. First we will show that $1 \in S$ (the base case). Notice that when $n = 1$, we have

$$1 + 2 + 2^2 + 2^3 + \cdots + 2^{n-1} = 2^0 = 1$$

and

$$2^n - 1 = 2^1 - 1 = 2 - 1 = 1.$$

So equation (B.3) is true when $n = 1$, which means that $1 \in S$. For the inductive step, we need to show that $n+1 \in S$ whenever $n \in S$. To do so, we will assume that $n \in S$ for some integer $n \geq 1$ (the inductive hypothesis). In this case, we will assume that

$$1 + 2 + 2^2 + 2^3 + \cdots + 2^{n-1} = 2^n - 1. \tag{B.4}$$

We then need to prove that $n+1 \in S$. So we need to show that

$$1 + 2 + 2^2 + 2^3 + \cdots + 2^{n-1} + 2^{(n+1)-1} = 2^{n+1} - 1,$$

or, equivalently,

$$1 + 2 + 2^2 + 2^3 + \cdots + 2^{n-1} + 2^n = 2^{n+1} - 1. \tag{B.5}$$

To prove (B.5), we can substitute from (B.4) in the left hand side of (B.5) to obtain

$$\begin{aligned} 1 + 2 + 2^2 + 2^3 + \cdots + 2^{n-1} + 2^n &= \left(1 + 2 + 2^2 + 2^3 + \cdots + 2^{n-1}\right) + 2^n \\ &= (2^n - 1) + 2^n \\ &= 2(2^n) - 1 \\ &= 2^{n+1} - 1, \end{aligned}$$

which shows that $n+1 \in S$. Therefore, $S = \mathbb{N}$ by the Principle of Mathematical Induction, which means that (B.3) is true for all $n \in \mathbb{N}$.

Activity B.6. Let r be a real number with $r > 1$. Consider the statements

$$1 + r + r^2 + r^3 + \cdots r^{n-1} = \frac{r^n - 1}{r - 1} \tag{B.6}$$

for all $n \geq 1$. (Note that you have probably seen equation (B.6) in a previous class. Do you remember where it comes from?) We will use induction to prove that (B.6) is true for all $n \in \mathbb{N}$.

(a) Identify an appropriate set S on which to apply our induction argument.

(b) What is the base case? Give a precise statement, and then verify your statement.

(c) What is the induction hypothesis? What is the inductive step?

(d) Complete the inductive step to verify that equation (B.6) is true for all $n \in \mathbb{N}$.

The Extended Principle of Mathematical Induction

Preview Activity B.7. Let's now return to our *Let's Make a Great Deal* game. Suppose you fail to correctly answer the question that allows you to open the first door. But suppose also that, after

doing so, the host gives you the opportunity to answer a second question. If you correctly answer question number 2, then you win a key to open door number 2. As before, behind door number 2 is a prize and a key to open door number 3. Behind door number 3 is a prize and a key to open door number 4, and so on.

(a) Which prizes would you win if you answered the second question correctly?

(b) Suppose you fail to answer the second question correctly. The host then gives you an option of answering a third question. If you correctly answer question number 3, then you win a key to open door number 3. Which prizes would you win if you answered the third question correctly?

(c) Suppose you fail to answer the third question correctly. The host then gives you an option of answering a fourth question. If you correctly answer question number 4, then you win a key to open door number 4. Which prizes would you win if you answered the fourth question correctly?

(d) You probably see the pattern by now. Suppose you fail to answer the first $m - 1$ questions correctly for some integer $m \geq 2$. The host then gives you an option of answering an m^{th} question. If you correctly answer question number m, then you win a key to open door number m. Which prizes would you win if you answered the m^{th} question correctly?

(e) Compare this version of the game to the version described in Activity B.1. How are the two versions alike, and how are they different?

The Principle of Mathematical Induction that we stated in the previous section is a method for proving an entire family of statements, one for each positive integer n. There are, however, instances when we need to modify our induction arguments slightly. For example, consider the statement

$$2^n < n!.$$

If we try to prove this statement for all $n \in \mathbb{N}$, we immediately encounter a problem in that the statement is not true for $n = 1$. In fact, the statement is also false for $n = 2$ and $n = 3$. However, the statement does appear to be true for $n \geq 4$. If we want to use the Principle of Mathematical Induction to prove that this statement is true for $n \geq 4$, we will need to somehow translate it to an equivalent statement that is true for *all* positive integers. One way to do this is by re-indexing the statement to say

$$2^{n+3} < (n + 3)!$$

for all positive integers n. Although this would solve the problem, the resulting statement is more complicated, and if we started with a more involved result, re-indexing could potentially make the situation appear more difficult than it really is. This is where the result of Activity B.7 is useful; in fact, proving that $2^n < n!$ for all $n \geq 4$ is analogous to answering the 4^{th} question correctly. The point of Activity B.7 is that it really shouldn't make any difference what our starting point is, as long as we have one. If we can open door number n_0 for some integer n_0, then we will be able to open all doors with numbers higher than n_0 as well—even though we may not be able to open the doors numbered lower than n_0. This is the idea behind the *Extended Principle of Mathematical Induction*.

Axiom B.8 (Extended Principle of Mathematical Induction)**.** *Let S be a subset of the set of the integers $\mathbb{Z}$. If there is an integer n_0 such that*

(i) *S contains n_0 and*

(ii) *for all $n \geq n_0$, S contains the positive integer $n + 1$ whenever S contains n,*

then S contains every integer greater than or equal to n_0.

When applying the Extended Principle of Mathematical Induction, our base case is when $n = n_0$ instead of $n = 1$, but the inductive step is still the same. Note that our goal in this situation is to show that S contains every integer greater than or equal to n_0, which establishes that our statement is true for all $n \geq n_0$. Doing so doesn't rule out the possibility that S could contain other integers as well, but we are only interested in the integers greater than or equal to n_0.

Activity B.9. To illustrate the Extended Principle of Mathematical Induction, we will continue with our example of proving that $2^n < n!$ for all $n \geq 4$.

(a) State and verify the base case for this inductive proof.

(b) What is the inductive hypothesis in this proof? Give a precise statement, and then complete the inductive step.

(c) What conclusion can you draw from your work in parts (a) and (b)?

Notice that the only difference between the Principle of Mathematical Induction and the Extended Principle of Mathematical Induction is the base case. In fact, letting $n_0 = 1$ in the Extended Principle yields the original Principle of Mathematical Induction. A bit later, we will see that these two forms of induction are actually equivalent.

Before we move one, one additional comment about the format of induction proofs is in order. Generally, when we construct an induction argument, we do not set up the set S as we have done in our previous examples. Instead, if we are trying to prove a family $\{P(n)\}$ of statements, one for each integer greater than or equal to some integer n_0, we simply prove that $P(n_0)$ is true, and then prove that if $P(n)$ is true for some integer $n \geq n_0$, then $P(n+1)$ is also true. In the induction proofs throughout the rest of this appendix (and throughout the remainder of the text), we will follow this simplified format.

The Strong Form of Mathematical Induction

Preview Activity B.10. Let's return once more to our *Let's Make a Great Deal* game. We will keep the rules the same and suppose in addition that behind each door is not only a prize and a key to open the next door, but also keys to open all of the *preceding* doors. Compare and contrast this game to the previous versions of the game we have studied. How is it similar, and how is it different? Do the outcomes of the game change?

The version of the *Let's Make a Great Deal* game from Activity B.10 may seem a bit silly; after all, why do we need all of those extra keys? But let's examine how this version of the game translates to an induction proof. We still need to answer some question (prove some base case n_0) to begin. In our previous inductive steps, we then showed that $n \in S$ implies $n + 1 \in S$—that is, each door contains the key to the next door. In our new version of the game, the idea of having all of the keys to the preceding doors is analogous to assuming not only that $n \in S$, but also that n_0, $n_0 + 1$, $n_0 + 2$, . . ., n are all in S. In other words, we can assume the validity of all of the previous statements, not just the n^{th} statement. To see why this might be useful, consider the statement that

every nonnegative integer n has a binary representation—that is, there exists $r \geq 0$ and integers a_r, a_{r-1}, …, a_1, a_0, all either 0 or 1, such that

$$n = a_r 2^r + a_{r-1} 2^{r-1} + \cdots + a_2 2^2 + a_1 2 + a_0.$$

For our base case ($n = 0$) we have that $0 = 0$, and so we are done (letting $r = 0$ and $a_0 = 0$). For the inductive step, it is a bit complicated to add 1 to n (in binary) to show that $n + 1$ has a binary representation. (We would have to worry about all the possible carries.) However, if $n + 1$ is even, then $k = (n + 1)/2$ is smaller than $n + 1$. If k has a binary representation

$$k = b_s 2^s + b_{s-1} 2^{s-1} + \cdots + b_2 2^2 + b_1 2 + b_0,$$

then

$$n + 1 = 2k = b_s 2^{s+1} + b_{s-1} 2^s + \cdots + b_2 2^3 + b_1 2^2 + b_0 2,$$

and so $n + 1$ has a binary representation. If $n + 1$ is odd, then $k = n/2$ is smaller than $n + 1$. If k has a binary representation

$$k = b_s 2^s + b_{s-1} 2^{s-1} + \cdots + b_2 2^2 + b_1 2 + b_0,$$

then

$$n + 1 = 2k + 1 = b_s 2^{s+1} + b_{s-1} 2^s + \cdots + b_2 2^3 + b_1 2^2 + b_0 2 + 1,$$

and so $n+1$ has a binary representation in this case as well. By assuming that *all* of the nonnegative integers less than or equal to n have binary representations, we can fairly easily prove that $n + 1$ also has a binary representation.

Being able to assume the statement we want to prove for *all* integers less than or equal to n is at times necessary for us to carry out an induction proof. The axiom that allows us to use such a method is called the *Strong Form of Mathematical Induction*.

Axiom B.11 (Strong Form of Mathematical Induction)**.** *Let S be a subset of $\mathbb{Z}$ containing some integer n_0. Suppose that for all $n \geq n_0$, S contains $n + 1$ whenever S contains each integer m with $n_0 \leq m \leq n$. Then S contains all integers greater than or equal to n_0.*

To reiterate, the Strong Form of Mathematical Induction allows us to assume much more in our inductive hypothesis than the previous two versions do. Strong induction is useful in a variety of settings, including proving results involving certain recursively defined sequences like the Fibonacci sequence.

Recall that the Fibonacci sequence is defined by the recurrence relation

$$f_n = f_{n-1} + f_{n-2} \tag{B.7}$$

for all $n \geq 3$, with $f_1 = f_2 = 1$. The recurrence relation (B.7) is very time consuming to use to compute f_n for large values of n. However, it turns out that there is a fascinating formula that gives the n^{th} term of the Fibonacci sequence directly, without using the relation from (B.7).

Let $\varphi = \frac{1+\sqrt{5}}{2}$ and $\overline{\varphi} = \frac{1-\sqrt{5}}{2}$. We will show that

$$f_n = \frac{\varphi^n - \overline{\varphi}^n}{\sqrt{5}}. \tag{B.8}$$

Formula (B.8) is called *Binet's Formula.*[*] The number $\varphi = \frac{1+\sqrt{5}}{2}$ is intimately related to the Fibonacci sequence. This number also occurs often in other areas of mathematics. It was an important

[*] If you wonder where a formula like this comes from, the quantities φ and $\overline{\varphi}$ are eigenvalues for a certain matrix that we can use to generate the Fibonacci sequence. This formula follows in a straightforward manner.

number to the ancient Greek mathematicians who felt that the most aesthetically pleasing rectangles had sides in the ratio of $\varphi : 1$. The Greeks called φ the *golden mean* or *golden ratio*. Formula (B.8) provides a fascinating relationship between the Fibonacci numbers and the golden ratio. It is also surprising (and not at all obvious) that the expression on the right hand side of (B.8) is an integer for each positive integer n.

To prove formula (B.8), we will use mathematical induction. Note that since f_1 and f_2 are defined independent of the recursion relation, it will be necessary to verify our statement in both the $n = 1$ and $n = 2$ cases. First we will make a few observations. Note that the golden ratio φ and its conjugate $\overline{\varphi}$ are the solutions (check this!) to the quadratic equation

$$x^2 = x + 1.$$

In addition,

$$\varphi + \overline{\varphi} = 1 \qquad \text{and} \qquad \varphi - \overline{\varphi} = \sqrt{5}.$$

Therefore,

$$\varphi^{n+1} = \varphi^2\varphi^{n-1} = (\varphi + 1)\varphi^{n-1} = \varphi^n + \varphi^{n-1}.$$

Similarly,

$$\overline{\varphi}^{n+1} = \overline{\varphi}^2\overline{\varphi}^{n-1} = (\overline{\varphi} + 1)\overline{\varphi}^{n-1} = \overline{\varphi}^n + \overline{\varphi}^{n-1}.$$

We will use these last two identities in our proof of Binet's Formula. We will proceed by mathematical induction on n. When $n = 1$, we have

$$\frac{1}{\sqrt{5}}(\varphi + \overline{\varphi}) = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2} - \frac{1-\sqrt{5}}{2}\right) = \frac{\sqrt{5}}{\sqrt{5}} = 1 = f_1.$$

So equation (B.8) is true when $n = 1$. When $n = 2$, we have

$$\begin{aligned}
\frac{1}{\sqrt{5}}(\varphi^2 + \overline{\varphi}^2) &= \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^2 - \left(\frac{1-\sqrt{5}}{2}\right)^2\right] \\
&= \frac{1}{\sqrt{5}}\left(\frac{1}{4}(1 + 2\sqrt{5} + 5) - \frac{1}{4}(1 - 2\sqrt{5} + 5)\right) \\
&= \frac{\sqrt{5}}{\sqrt{5}} \\
&= 1 \\
&= f_1.
\end{aligned}$$

So equation (B.8) is true when $n = 2$.

Since each term in the Fibonacci sequence depends on the preceding two terms, we will need to use the Strong Form of Mathematical Induction in our proof. Therefore, assume equation (B.8) is true for all positive integers m less than a given $n \geq 2$. We must show that

$$f_{n+1} = \frac{\varphi^{n+1} - \overline{\varphi}^{n+1}}{\sqrt{5}}.$$

This follows by observing that

$$\begin{aligned}\frac{\varphi^{n+1}-\overline{\varphi}^{n+1}}{\sqrt{5}} &= \frac{1}{\sqrt{5}}\left[\left(\varphi^n+\varphi^{n-1}\right)-\left(\overline{\varphi}^n+\overline{\varphi}^{n-1}\right)\right]\\ &= \frac{1}{\sqrt{5}}\left[\left(\varphi^n-\overline{\varphi}^n\right)+\left(\varphi^{n-1}-\overline{\varphi}^{n-1}\right)\right]\\ &= f_n+f_{n-1}\\ &= f_{n+1}.\end{aligned}$$

Thus, by induction, Binet's Formula is true for all $n \in \mathbb{N}$.

Note that, with Binet's Formula, we can easily compute f_n for very large values of n. For example, f_{500} is equal to

$$\begin{aligned}&139423224561697880139724382870407283950070256587697 3\\ &072641089629483255716228632906915576588762225212941 25.\end{aligned}$$

The Well-Ordering Principle

Preview Activity B.12.

(a) Which of the following sets contains a smallest element? Explain.

(i) $A = \{1, 2, 3, 4\}$

(ii) $B = \{n \in \mathbb{N} \mid n > 4\}$

(iii) $C = \{x \in \mathbb{Z} \mid x > 4\}$

(iv) $D = \{x \in \mathbb{Z} \mid x < 4\}$

(b) Do you believe the following statement is true or false?

Every nonempty subset of $\mathbb{N}$ has a least element.

No proof is required if you believe the statement is true, but if you believe it is false, you should be able to give a counterexample.

(c) Do you believe the following statement is true or false?

Every nonempty subset of $\mathbb{Z}$ has a least element.

No proof is required if you believe the statement is true, but if you believe it is false, you should be able to give a counterexample.

(d) Let's return to our *Let's Make a Great Deal* game. Suppose a friend of yours has won the game. Is it possible to determine which question your friend answered correctly to win? Explain.

We will begin our discussion of the Well-Ordering Principle with the following familiar proof that $\sqrt{2}$ is not a rational number:

> Assume to the contrary that $\sqrt{2} = \frac{m}{n}$ for some positive integers m and $n \neq 0$ so that $\frac{m}{n}$ is in reduced form (that is, the greatest common divisor of m and n is 1). Then $n\sqrt{2} = m$, and so $2n^2 = m^2$. Thus, the prime 2 divides m^2 and so 2 also divides m. This means that $m = 2k$ for some integer k. Then $4k^2 = m^2 = 2n^2$, and so $2k^2 = n^2$. From this we see that 2 divides n, contradicting the fact that m and n have no common factors greater than 1. We can therefore conclude that $\sqrt{2}$ is not a rational number.

All of the results that we used in this proof—including those that you may not have seen before—are verified in our investigations, with one exception. This exception illustrates how easy it is to take certain mathematical results for granted. The exception in the proof is that we can always find a rational number *in reduced form* that is equal to $\frac{m}{n}$. How do we know we can do this? Recall that the rational numbers $\frac{a}{b}$ and $\frac{c}{d}$ are equal if $ad = bc$. To show that we can find a rational number in reduced form that is equal to $\frac{m}{n}$, we might consider all rational numbers $\frac{a}{b}$ that are equal to $\frac{m}{n}$ and prove that there is one so that the greatest common divisor (or gcd) of a and b is 1. In other words, let

$$S = \left\{ \gcd(a, b) : \frac{a}{b} = \frac{m}{n} \right\}$$

be the set of all greatest common divisors of the numerators and denominators of fractions that are equal to $\frac{m}{n}$. Certainly S is not empty, since $\gcd(m, n)$ is in S. Also, since $\gcd(x, y) \geq 1$ for any integers x and y, not both 0, it follows that the integers in S are all greater than or equal to 1. We need to actually show that 1 is in S. If 1 is in S, then 1 will have to be the smallest element in S.

This is where our conclusion from Activity B.12 is helpful. If we assume that every nonempty subset of $\mathbb{N}$ contains a smallest integer, then S must contain a smallest integer d. Now all we have to do is show that $d = 1$. This is because if $d = 1$, then there must be a fraction $\frac{a}{b}$ equal to $\frac{m}{n}$ with $\gcd(a, b) = 1$. Since $d \in S$, there exists a rational number $\frac{a}{b}$ that is equal to $\frac{m}{n}$ with $\gcd(a, b) = d$. Now d divides both a and b, so let $da' = a$ and $db' = b$ for some positive integers a' and b'. Since $d = \gcd(a, b)$, it follows that $\gcd(a', b') = 1$. But

$$\frac{m}{n} = \frac{a}{b} = \frac{a'd}{b'd} = \frac{a'}{b'},$$

and so $\gcd(a', b')$ is in S. However, this can happen only if $d = 1$, and so 1 is the smallest element in S, and there is a fraction in reduced form that is equal to $\frac{m}{n}$.

The key to our proof that every rational number is equal to a rational number in reduced form was the assumption that the set S contained a smallest element. Based on Activity B.12, this seems like a reasonable assumption to make. The principle that allows us to make it is called the *Well-Ordering Principle*.

To thoroughly understand the Well-Ordering Principle, we first need to discuss well-ordered sets. But to talk about well-ordered sets, we need to understand ordered sets in general. This leads us to the idea of binary relations.

A *binary relation* on a set (or *relation* for short) is simply a way to compare elements in the set. For example, consider the set consisting of the citizens of the state of Michigan. We might say that one person is related to another if the two have at least one parent in common. Note that some people in this set are related and others are not. This observation illustrates the fact that a relation on a set does not need to compare *every* pair of elements in the set.

As a smaller example, let $S = \{1, 2, 3, 4\}$, and say that a and b are related in S if a divides b. In this case we have that 1 is related to 2, 3, and 4, while 2 is related only to 4. To clearly identify related pairs of elements in S, we might list all of the related elements as ordered pairs. For this relation, the resulting pairs are $(1, 2)$, $(1, 3)$, $(1, 4)$, and $(2, 4)$. The general definition of a relation on a set follows this example.

Definition B.13. A **relation** on a set S is a subset R of the Cartesian product $S \times S$. In other words, a relation on S is a set of ordered pairs, where both coordinates of each pair are elements of S.

For example, the subset $R = \{(a, a) : a \in \mathbb{Z}\}$ of $\mathbb{Z} \times \mathbb{Z}$ is the relation we call *equals*. If R is a relation on a set S, we usually suppress the set notation and write $a \sim b$, read "a is related to b," if $(a, b) \in R$. In this case, we often refer to $\sim$ as the relation instead of the set R. Sometimes we use familiar symbols for special relations. For example, we write $a = b$ if (a, b) is in the set $R = \{(a, a) : a \in \mathbb{Z}\}$.

There are several properties that relations may satisfy. For example:

- A relation $\sim$ on a set S is *reflexive* if $a \sim a$ for all $a \in S$.
- A relation $\sim$ on a set S is *symmetric* if whenever $a \sim b$ (for any a, $b \in S$), we also have $b \sim a$.
- A relation $\sim$ on a set S is *transitive* if whenever $a \sim b$ and $b \sim c$ (for any a, b, $c \in S$), we also have $a \sim c$.
- A relation $\sim$ on a set S is *antisymmetric* if whenever $a \sim b$ and $b \sim a$ (for any a, $b \in S$), then $a = b$.

Activity B.14. Determine whether each of the given relations on $\mathbb{Z}$ is reflexive, symmetric, transitive, and/or antisymmetric. Give reasons to support your answers.

(a) $R = \{(a, b) : a > b\}$

(b) $R = \{(a, b) : a^2 = b^2\}$

(c) $R = \{(a, b) : ab \geq 0\}$

(d) $R = \{(a, b) : a$ and b leave the same remainder when divided by $3\}$

Some relations, like the relation $\leq$ on $\mathbb{R}$, give us a way of organizing the elements in the sets on which they are defined in a specified way (e.g., on the number line). The next definition formalizes this idea.

Definition B.15. A set S is a **partially ordered set** (or **poset**) if there is a relation, which we will denote by $\leq$, on S such that for all $x, y, z \in S$:

(i) $x \leq x$ ($\leq$ is a reflexive relation);

(ii) if $x \leq y$ and $y \leq x$, then $x = y$ ($\leq$ is an antisymmetric relation); and

(iii) if $x \leq y$ and $y \leq z$, then $x \leq z$ ($\leq$ is a transitive relation).

A partially ordered set S is **totally ordered** if it also satisfies

(iv) either $x \leq y$, $y \leq x$, or $x = y$ ($\leq$ satisfies the trichotomy property).

What makes a partially ordered set totally ordered is that *any* two elements are related somehow, which is not necessary in a partially ordered set. An example of a partially ordered set that is not totally ordered is the set of positive integers, where a is related to b if a divides b. Examples of totally ordered sets are $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$, using the standard "less than or equal to" relation ($\leq$).

The Well-Ordering Principle tells us that any subset of $\mathbb{Z}$ that is bounded below contains a smallest element. To make this all precise, we need to explain what we mean by a smallest element in a set and also what bounded below means. These definitions should not be surprising.

Definition B.16. Let S be a totally ordered set, and let A be a subset of S.

- An element $m \in S$ is a **lower bound** for A if $m \leq a$ for all $a \in A$. The set A is **bounded below** if A has a lower bound in S.
- An element $a \in A$ is a **least** or **smallest** element in A if $a \leq a'$ for all $a' \in A$.

It is important to note the difference between a lower bound and a smallest element. The integer -2 is a lower bound for $\mathbb{N}$, but is not a smallest element in $\mathbb{N}$ since it is not an element of $\mathbb{N}$. Every smallest element in a set is also a lower bound for the set. However, not every set is bounded below or contains a least element. For example, the set of even integers is not bounded below. In addition, a set can be bounded below but not contain a least element. For example, the open interval $(0, 1) = \{x \in \mathbb{R} : 0 < x < 1\}$ is bounded below by 0 but does does not have a smallest element (since there is no smallest positive real number).

We have one more step before stating the Well-Ordering Principle.

Definition B.17. A totally ordered set S is **well-ordered** if every nonempty subset A of S contains a least element.

We can now formally state the Well-Ordering Principle.

Axiom B.18 (The Well-Ordering Principle)**.** *Every nonempty subset of $\mathbb{Z}$ that is bounded below is well-ordered.*

The Well-Ordering Principle is often stated within the specific context of the natural numbers, where it implies that every nonempty subset of $\mathbb{N}$ contains a smallest element. Our version is somewhat more general and is equivalent to the following:

Axiom B.19 (The Well-Ordering Principle)**.** *Every nonempty subset of $\mathbb{Z}$ that is bounded below contains a smallest element.*

Note that, in general, a set can have a smallest element without being well-ordered. Consider, for example, the set $\mathbb{R}^*$ of all nonnegative real numbers. Note that $\mathbb{R}^*$ has a smallest element—namely, 0—but is not well-ordered, since it contains a nonempty subset (the positive reals, for example) that does not have a smallest element. The equivalence of the two forms of the Well-Ordering Principle, as we have stated them, stems from the fact that both are universally quantified—that is, both refer to *every* nonempty subset of $\mathbb{Z}$ that is bounded below. The first is really saying that if S is a nonempty subset of $\mathbb{Z}$ that is bounded below, then every nonempty subset of S contains a smallest element. But a nonempty subset of S is still a nonempty subset of $\mathbb{Z}$ that is bounded below. For this reason, the second version of the Well-Ordering Principle is equivalent to the first.

As an example of the use of the Well-Ordering Principle, we will prove the following theorem, which we also proved in Investigation 1 as part of the Fundamental Theorem of Arithmetic. (See page 10.)

Theorem. *Every integer greater than 1 is either prime or can be factored into a product of primes.*

Proof. To use the Well-Ordering Principle, we need to define a nonempty subset of $\mathbb{Z}$ that is bounded below. To do so, we will proceed by contradiction and assume that there is an integer greater than 1 that is not prime and cannot be written as a product of primes. Let

$$S = \{n \in \mathbb{N} : n \text{ is not prime and cannot be written as a product of primes}\}.$$

Then S is nonempty by hypothesis and is bounded below (by 1). The Well-Ordering Principle tells

us that S contains a smallest element m. By definition, m is not prime, so there exist integers a and b with $1 < a, b < m$ such that $m = ab$. Since m is the smallest element in S, it follows that a and b are not in S. Thus, a and b are either prime or can be written as a product of primes. Therefore, there exist positive integers r and s and primes $p_1, p_2, \ldots, p_r$ and $q_1, q_2, \ldots, q_s$ such that

$$a = p_1 p_2 \cdots p_r \quad \text{and} \quad b = q_1 q_2 \cdots q_s.$$

But then

$$m = ab = p_1 p_2 \cdots p_r q_1 q_2 \cdots q_s$$

is a product of primes, which is a contradiction, since we assumed that m could not be written as a product of primes. We can therefore conclude that every integer greater than 1 is either prime or can be factored into a product of primes. ■

You may want to compare the above proof to the proof of the Fundamental Theorem of Arithmetic from Investigation 1, which used induction instead of the Well-Ordering Principle. It is no coincidence that both methods can be used to establish similar results. In fact, as we will see in the next section, the Well-Ordering Principle and all three different forms of the Principle of Mathematical Induction are logically equivalent.

It is also important to note that we have labeled both the principles of mathematical induction and the Well-Ordering Principle as axioms and not theorems. That is because we cannot prove any one of these principles (although, as noted above, we can prove that they are equivalent to each other), but they seem evident enough that we will assume them to be true.

The Equivalence of the Well-Ordering Principle and the Principles of Mathematical Induction

In this section, we will prove that the principles of mathematical induction and the Well-Ordering Principle are equivalent. That is, any one of these principles implies any of the others. It is important to note that Theorem B.20 does not prove any of these principles, but says that if we assume one of them to be valid, then all of the others are valid as well.

Theorem B.20. *The following are equivalent:*

(i) *The Principle of Mathematical Induction*

(ii) *The Extended Principle of Mathematical Induction*

(iii) *The Strong Form of Mathematical Induction*

(iv) *The Well-Ordering Principle.*

A word on the proof of Theorem B.20: To prove this string of equivalences, we will show that (i) implies (ii), (ii) implies (iii), (iii) implies (iv), and then (iv) implies (i). This will demonstrate that any one of the four statements implies any of the others, as can be seen by following an appropriate string of implications. (For example, to see that (iii) implies (ii), we can simply note that (iii) implies (iv), (iv) implies (i), and (i) implies (ii).) Also, the proofs of each equivalence are subtle in that both the hypotheses and conclusions are complicated statements. Because of this, we will have to be

very careful about our assumptions in each case. We will begin by showing that the Principle of Mathematical Induction implies the Extended Principle of Mathematical Induction. The steps to complete this proof are outlined in the next activity.

Activity B.21. To prove that the Principle of Mathematical Induction implies the Extended Principle of Mathematical Induction, we will assume that the Principle of Mathematical Induction is true. This means that any subset S of $\mathbb{N}$ that contains 1 and has the property that $n+1 \in S$ whenever $n \in S$ must be equal to $\mathbb{N}$.

We need to prove that the Extended Principle of Mathematical Induction is true. So we will assume that n_0 is an integer and T is a subset of $\mathbb{Z}$ such that $n_0 \in T$ and $n+1 \in T$ whenever $n \geq n_0$ and $n \in T$. We need to prove that $\{n \in \mathbb{Z} \mid n \geq n_0\} \subseteq T$.

In order to use the Principle of Mathematical Induction, we need to construct some subset of $\mathbb{N}$ that is related to T but contains 1 as its smallest element. To do so, we can shift or re-index the elements of T so that n_0 corresponds to 1. In particular, we will define S to be the set

$$S = \{k - n_0 + 1 \in \mathbb{N} \mid k \in T\}.$$

(a) Use the assumption that $n_0 \in T$ to prove that $1 \in S$.

(b) We now need to prove that if $n \geq 1$ is in S, then $n+1 \in S$. Let $n \geq 1$ be in S. There is a corresponding element k in T. Write down a formula for k in terms of n and n_0. Explain your reasoning.

(c) Based on our assumptions about T, what integer besides k must also be in T?

(d) Now use the result of part (c) to conclude that $n+1 \in S$.

This proves that S contains 1 and that $n+1 \in S$ whenever $n \in S$. Therefore, by the Principle of Mathematical Induction, $S = \mathbb{N}$. We will now use this fact to prove that $\{n \in \mathbb{Z} \mid n \geq n_0\} \subseteq T$.

By assumption, $n_0 \in T$. So we need to prove that if $x \in \mathbb{Z}$ with $x > n_0$, then $x \in T$. To this end, assume that $x \in \mathbb{Z}$ and $x > n_0$.

(e) Prove that $x - n_0 \in \mathbb{N}$ and therefore $x - n_0 \in S$.

(f) Show that part (e) implies that $x - 1 \in T$. Then explain how we can conclude that $x \in T$.

(g) Explain why we have now completed the proof that the Extended Principle of Mathematical Induction is implied by the Principle of Mathematical Induction.

We will now consider the other implications in Theorem B.20.

Proof of Theorem B.20. (ii) $\rightarrow$ (iii): We will assume that the Extended Principle of Mathematical Induction is true. We will then prove that the Strong Form of Mathematical Induction must also be true.

To prove the Strong Form, we let n_0 be an integer and assume T is a subset of $\mathbb{Z}$ such that

(1) $n_0 \in T$, and

(2) for every $n \in \mathbb{Z}$ with $n \geq n_0$, if $\{n_0, n_0+1, \ldots, n\} \subseteq T$, then $(n+1) \in T$.

We then need to prove that T contains all integers greater than or equal to n_0—or, equivalently, that

$$\{x \in \mathbb{Z} \mid x \geq n_0\} \subseteq T.$$

We will use the Extended Principle of Mathematical Induction to prove the following statement:

For each natural number k with $k \geq n_0$, $\{n_0, n_0 + 1, \ldots, k\} \subseteq T$.

Since we have assumed that $n_0 \in T$, we know that $\{n_0\} \subseteq T$. Hence the base case ($k = n_0$) is true.

For the inductive step, let $k \in \mathbb{N}$ with $k \geq n_0$, and assume that

$$\{n_0, n_0 + 1, \ldots, k\} \subseteq T.$$

By what we assumed about the set T, we can conclude that $k + 1 \in T$, and therefore

$$\{n_0, n_0 + 1, \ldots, k, k + 1\} \subseteq T.$$

This proves that if $\{n_0, n_0 + 1, \ldots, k\} \subseteq T$, then $\{n_0, n_0 + 1, \ldots, k, k + 1\} \subseteq T$ is true; hence, the inductive step has been established. By the Extended Principle of Mathematical Induction, we can conclude that for each natural number k with $k \geq n_0$, $\{n_0, n_0 + 1, \ldots, k\} \subseteq T$. This proves that T contains all integers greater than or equal to n_0, which is what we needed to prove to show that the Strong Form of Mathematical Induction is true.

We have therefore shown that if the Extended Principle of Mathematical Induction is true, then the Strong Form of Mathematical Induction is also true.

(iii) $\rightarrow$ (iv): We will now show that the Strong Form of Mathematical Induction implies the Well-Ordering Principle. We will assume the Strong Form of Mathematical Induction. That is, whenever we have a subset U of $\mathbb{Z}$ such that

- $n_0 \in U$ for some integer n_0; and
- whenever $k \in U$ for all $n_0 \leq k \leq n$, then $n + 1 \in U$,

then U contains the set of all integers greater than or equal to n_0. To prove the Well-Ordering Principle, we must show that any nonempty subset of $\mathbb{Z}$ that is bounded below contains a smallest element. We will proceed by contradiction and assume there is a nonempty subset T of $\mathbb{Z}$ that is bounded below and does not contain a least element. Let m be a lower bound for T. If $m \in T$, then T contains a smallest element, namely m. So m cannot be an element of T. Let S be the set of all strict lower bounds for T—that is,

$$S = \{n \in \mathbb{Z} : n < t \text{ for all } t \in T\}.$$

Since m is a lower bound for T and $m \notin T$ it follows that $m < t$ for all $t \in T$. So $m \in S$. Suppose $n \geq m$ so that m, $m + 1$, $m + 2$, …, n are all in S. We will show $n + 1 \in S$. Since $n \in S$ we must have $n < t$ for all $t \in T$. This, however, implies that

$$n + 1 \leq t \tag{B.9}$$

for all $t \in T$. If $n + 1 = t$ for some $t \in T$, then $n + 1$ must be the smallest element in T, which cannot happen. Therefore,

$$n + 1 \neq t \tag{B.10}$$

for all $t \in T$. Combining (B.9) and (B.10) shows $n + 1 < t$ for all $t \in T$, and so $n + 1 \in S$. By the Strong Form of Mathematical Induction, we can then conclude that S contains all integers greater than or equal to m. It follows that every integer is a strict lower bound for T, and so $T = \emptyset$, a contradiction. Therefore, no such set T exists, which means that every nonempty subset of $\mathbb{Z}$ that is bounded below contains a smallest element. We have therefore shown that the Strong Form of Mathematical Induction implies the Well-Ordering Principle.

(iv) $\rightarrow$ (i): This is left to the reader in Activity B.22.

■

Concluding Activities

Activity B.22. Complete the proof of Theorem B.20 by proving that the Well-Ordering Principle implies the Principle of Mathematical Induction. (Hint: If S is a subset of $\mathbb{N}$ that contains 1 and also contains $n + 1$ whenever S contains n, consider the set T of all natural numbers that are not in S.)

Exercises

(1) In calculus, we often use the fact that $\frac{d}{dx}x^n = nx^{n-1}$ for every positive integer n, but we usually don't provide a rigorous proof of this result. Use induction to verify this derivative formula. Assume the product rule if you need it.

(2) Prove that

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

for every positive integer n.

(3) In the mid 20th century, the mathematician George Pólya suggested the apparent paradox that all girls have eyes of the same color.[†] His induction argument to verify this statement is as follows:

> For $n = 1$ the statement is obviously (or "vacuously") true. It remains to pass from n to $n + 1$. For the sake of concreteness, I shall pass from 3 to 4 and leave the general case for you.
>
> Let me introduce you to any four girls, Ann, Berthe, Carol, and Dorothy, or A, B, C, and D, for short. Allegedly ($n = 3$) the eyes of A, B, and C are of the same color. Consequently, the eyes of all four girls A, B, C, and D, must be of the same color; for the sake of full clarity, you may look at the diagram:
>
> $$\overbrace{A + \underbrace{B + C}} + D.$$

[†]This appears in Pólya's 1954 work *Induction and Analogy in Mathematics*, volume 1 of *Mathematics and Plausible Reasoning*, Princeton University Press.

This proves the point for $n+1=4$, and the passage from 4 to 5, for example, is, obviously, not more difficult.

A quick glance into the eyes of several girls will show that not all girls have the same eye color, so there must be a flaw in the argument. Find and explain the flaw.

(4) Consider the conjecture

$$1+2+3+4+\cdots+n=\frac{1}{2}\left(n+\frac{1}{2}\right)^2. \tag{B.11}$$

(a) Assume (B.11) is true for some positive integer n. That is, assume

$$1+2+3+4+\cdots+n=\frac{1}{2}\left(n+\frac{1}{2}\right)^2.$$

Show that (B.11) is true for the integer $n+1$.

(b) For which integers n is (B.11) a true statement? Explain. What does this exercise tell us about the importance of establishing a base case in an induction proof?

(5) (a) Experiment and conjecture a simple closed form for the sum

$$s_n=\frac{1}{1\times 3}+\frac{1}{3\times 5}+\frac{1}{5\times 7}+\cdots+\frac{1}{(2n-1)(2n+1)}$$

that is valid for every positive integer n.

(b) Use induction to prove your formula from part (a). Be explicit about which version of induction you are using.

(6) Experiment and conjecture a simple closed form for

$$\left(1-\frac{1}{4}\right)\left(1-\frac{1}{9}\right)\left(1-\frac{1}{16}\right)\cdots\left(1-\frac{1}{n^2}\right)$$

that is valid for every positive integer $n\geq 2$. Prove your conjecture.

(7) (a) Let $a_1=5$, $a_2=7$ and $a_n=3a_{n-1}-2a_{n-2}$ for $n\geq 3$. Experiment and conjecture a simple closed form for a_n that is valid for every positive integer n. (Hint: Compare a_n to 2^n.)

(b) Use induction to prove your formula from part (a). Be explicit about which version of induction you are using.

(8) Recall that the Fibonacci numbers f_n are defined by $f_1=1$, $f_2=1$, and $f_n=f_{n-1}+f_{n-2}$ for all $n\geq 3$. Show that every fifth Fibonacci number is divisible by 5. (In fact, something stronger is true: for any prime p, every p^{th} Fibonacci number is divisible by p.)

(9) Prove that for every positive integer n,

$$1(1!)+2(2!)+3(3!)+\cdots+n(n!)=(n+1)!-1. \tag{B.12}$$

(10) Prove that for every $n\in\mathbb{N}$, the number of subsets of a set with n elements in 2^n.

(11) Is the following statement true or false?

For all $n \in \mathbb{N}$, $(1 + 2 + 3 + \cdots + n)^2 = 1^3 + 2^3 + 3^3 + \cdots n^3$

If the statement is true, prove it. If it is false, find a counterexample.

(12) For which positive integers is $n!$ less than n^n? Prove your assertion.

(13) In this exercise, we will compare exponential functions to factorials. Let $a \geq 2$ be a positive integer.

(a) Show that if $a^n < n!$ for some positive integer n, then $a^{n+1} < (n+1)!$.

(b) To show that $a^n < n!$ for all n larger than some fixed integer, it remains to demonstrate that $a^n < n!$ for some positive integer n. This is a challenging problem. It is conjectured that, for $a > 3$, the sequence

$$s(a) = \text{round}\left(ae - \left(\frac{1}{2}\right)\log(2a\pi) - \frac{1}{a}\right)$$

gives the smallest positive integer n so that $a^n < n!$.[‡] (The function *round* means to round to the nearest integer.) Verify this formula for $a = 4$, $a = 5$, and $a = 6$.

(14) **Round Robin Tournaments.** Consider a tournament involving m players in which each player plays every other player just once and there are no ties. A *cycle* in the tournament is a set $\{P_1, P_2, \ldots, P_n\}$ of players so that player P_1 beats player P_2, player P_2 beats player P_3, and so on, and player P_n beats player P_1. Show that if there is a cycle in the tournament, then there is a cycle consisting of exactly three players.

(15) In this investigation, we proved that

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$$

for every positive integer n. Then, in Exercise (2), you were asked to show that

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

for every positive integer n. Mathematical induction is a useful tool for verifying such formulas, but how do we actually find the formulas in the first place? In this exercise, we will consider ways to answer this question.

(a) Let's next determine a formula for the sum of cubes. Our starting place is the expansion of $(x-1)^4$. Note that

$$(x-1)^4 = x^4 - 4x^3 + 6x^2 - 4x + 1.$$

Then

$$x^4 - (x-1)^4 = 4x^3 - 6x^2 + 4x - 1.$$

Next, we will calculate each side of the previous equation as x ranges from 1 to n:

$$\begin{array}{rcccccccc}
n^4 - (n-1)^4 & = & 4n^3 & - & 6n^2 & + & 4n & - & 1 \\
(n-1)^4 - (n-2)^4 & = & 4(n-1)^3 & - & 6(n-1)^2 & + & 4(n-1) & - & 1 \\
(n-2)^4 - (n-3)^4 & = & 4(n-2)^3 & - & 6(n-2)^2 & + & 4(n-2) & - & 1 \\
\vdots & & & & \vdots & & & & \\
4^4 - 3^4 & = & 4(4)^3 & - & 6(4)^2 & + & 4(4) & - & 1 \\
3^4 - 2^4 & = & 4(3)^3 & - & 6(3)^2 & + & 4(3) & - & 1 \\
2^4 - 1^4 & = & 4(2)^3 & - & 6(2)^2 & + & 4(2) & - & 1 \\
1^4 - 0^4 & = & 4(1)^3 & - & 6(1)^2 & + & 4(1) & - & 1
\end{array}$$

[‡]By Benoit Cloitre; see sequence A086824 in the On-Line Encyclopedia of Integer Sequences (`https://oeis.org/`).

Now we can add the entries on each side to find a formula for

$$1^3 + 2^3 + \cdots + n^3.$$

Complete this process, and then prove your formula by induction.

(b) Repeat the process from part (a) to find a formula for

$$1^4 + 2^4 + \cdots + n^4.$$

Then prove your formula.

(16) **The Towers of Hanoi.** In an ancient city in India, so the legend goes, monks in a temple have to move a pile of 64 sacred disks from one location to another. The disks are fragile; only one can be carried at a time. A disk may not be placed on top of a smaller, less valuable disk. In addition, there is only one other location in the temple (besides the original and destination locations) sacred enough that a pile of disks can be placed there.

So the monks begin moving disks back and forth, between the original pile, the pile at the new location, and the intermediate location, always keeping the piles in order (largest on the bottom, smallest on the top). The legend is that, before the monks make the final move to complete the pile in the new location, the temple will turn to dust and the world will end.

Generalize this problem to show that if there were n disks to move, it would take a total of $2^n - 1$ moves to complete the transfer from one location to another. Should we be worried about the world coming to an end?

(17) The usual total ordering given by $\leq$ on $\mathbb{Z}$ behaves nicely with respect to addition. Show that there is *no* total ordering of $\mathbb{Z}_n$ that behaves nicely with respect to addition in $\mathbb{Z}_n$. That is, show that there is no total ordering on $\mathbb{Z}_n$ such that, for all $[a]$, $[b]$, $[c] \in \mathbb{Z}_n$, if $[a] \leq [b]$, then $([a] + [c]) \leq ([b] + [c])$. (Hint: If there is such an ordering with $[0] \leq [1]$, use transitivity to show that $[0] \leq [n-1]$, and explain why this leads to a contradiction. Then think about what similar argument needs to be made to complete the proof.)

Appendix C

Methods of Proof

Preliminaries

This section is meant primarily for review and to clearly state the definitions that will be used in proofs throughout the appendix. For those who are familiar with this material, it is not necessary to read this section. It is included primarily for reference for the discussion of proofs in subsequent sections.

Definitions

Definitions play a very important role in mathematics. A direct proof of a proposition in mathematics is often a demonstration that the proposition follows logically from certain definitions and previously proven propositions. A **definition** is an agreement that a particular word or phrase will stand for some object, property, or other concept that we expect to refer to often. In many elementary proofs, the answer to the question, "How do we prove a certain proposition?", is often answered by means of a definition. For mathematical proofs, we need very precise and carefully worded definitions.

Definition C.1.

- The set of **natural numbers**, denoted $\mathbb{N}$, contains the counting numbers (1, 2, 3, and so on); that is,
$$\mathbb{N} = \{1, 2, 3, \ldots\}.$$
- The set of **whole numbers**, denoted $\mathbb{W}$, contains the counting numbers and zero; that is,
$$\mathbb{W} = \{0, 1, 2, 3, \ldots\}.$$
- The set of **integers**, denoted $\mathbb{Z}$, contains the whole numbers and their opposites (or negatives); that is,
$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}.$$

Definition C.2. An integer a is an **even integer** provided that there exists an integer n such that $a = 2n$. An integer a is an **odd integer** provided there exists an integer n such that $a = 2n + 1$.

Definition C.3. A nonzero integer m **divides** an integer n provided that there is an integer q such that $n = m \cdot q$.

- If a and b are integers and $a \neq 0$, we frequently use the notation $a \mid b$ as a shorthand for "a divides b."

- If a and b are integers and $a \neq 0$ and a divides b, we also ay that m is a **divisor** of n, m is a **factor** of n, and n is a **multiple** of m.
- The integer 0 is not a divisor of any integer and is a multiple of every integer.

Definition C.4. A natural number p is a **prime number** provided that it is greater than 1 and the only natural numbers that are factors of p are 1 and p. A natural number other than 1 that is not a prime number is a **composite number**. The number 1 is neither prime nor composite.

Definition C.5. Let $n \in \mathbb{N}$. If a and b are integers, then we say that $\boldsymbol{a}$ **is congruent to** $\boldsymbol{b}$ **modulo** $\boldsymbol{n}$ provided that n divides $a - b$.

Note: A standard notation for "a is congruent to b modulo n is $a \equiv b \pmod{n}$. This is read as "a is congruent to b modulo n" or "a is congruent to b mod n."

Definitions Involving Sets

Definition C.6. Two sets, A and B, are **equal** when they have precisely the same elements.

The set A is a **subset** of a set B provided that each element of A is an element of B.

- When sets A and B are equal, we write $A = B$ and when they are not equal, we write $A \neq B$.
- When the set A is a subset of the set B, we write $A \subseteq B$ and also say that A is contained in B. When A is not a subset of B, we write $A \nsubseteq B$.

Definition C.7. Let A and B be two sets contained in some universal set U. The set A is a **proper subset** of B provided that $A \subseteq B$ and $A \neq B$.

Note: When a set A is a proper subset of a set B, we write $A \subset B$.

Definition C.8. Let A and B be subsets of some universal set U. The **intersection** of A and B, written $A \cap B$ and read "A intersect B," is the set of all elements that are in both A and B. That is,

$$A \cap B = \{x \in U \mid x \in A \text{ and } x \in B\}.$$

The **union** of A and B, written $A \cup B$ and read "A union B," is the set of all elements that are in A or in B. That is,

$$A \cup B = \{x \in U \mid x \in A \text{ or } x \in B\}.$$

Definition C.9. Let A and B be subsets of some universal set U. The **set difference** of A and B, or **relative complement** of B with respect to A, written $A - B$ and read "A minus B" or "the complement of B with respect to A," is the set of all elements in A that are not in B. That is,

$$A - B = \{x \in U \mid x \in A \text{ and } x \notin B\}.$$

The **complement** of the set A, written A^c and read "the complement of A," is the set of all elements of U that are not in A. That is,

$$A^c = \{x \in U \mid x \notin A\}.$$

Useful Logic for Constructing Proofs

A **statement** is a declarative sentence that is either true or false but not both. A **compound statement** is a statement that contains one or more operators. Because some operators are used so frequently in logic and mathematics, we give them names and use special symbols to represent them.

- The **conjunction** of the statements P and Q is the statement "P **and** Q" and is denoted by $P \wedge Q$. The statement $P \wedge Q$ is true only when both P and Q are true.
- The **disjunction** of the statements P and Q is the statement "P **or** Q" and is denoted by $P \vee Q$. The statement $P \vee Q$ is true only when at least one of P or Q is true.
- The **negation (of a statement)** of the statement P is the statement "**not** P" and is denoted by $\neg P$. The negation of P is true only when P is false, and $\neg P$ is false only when P is true.
- The **implication** or **conditional** is the statement "**If** P **then** Q" and is denoted by $P \to Q$. The statement $P \to Q$ is often read as "P **implies** Q. The statement $P \to Q$ is false only when P is true and Q is false.
- The **biconditional statement** is the statement "***P* if and only if *Q***" and is denoted by $P \leftrightarrow Q$. The statement $P \leftrightarrow Q$ is true only when both P and Q have the same truth values.

Definition C.10. Two expressions X and Y are **logically equivalent** provided that they have the same truth value for all possible combinations of truth values for all variables appearing in the two expressions. In this case, we write $X \equiv Y$ and say that X and Y are logically equivalent.

The following theorem states some of the most frequently used logical equivalencies used when writing mathematical proofs.

Theorem C.11 (Important Logical Equivalencies).
For statements P, Q, and R,

De Morgan's Laws	$\neg(P \wedge Q) \equiv \neg P \vee \neg Q$ $\neg(P \vee Q) \equiv \neg P \wedge \neg Q$
Conditional Statements	$P \to Q \equiv \neg Q \to \neg P$ *(contrapositive)* $P \to Q \equiv \neg P \vee Q$ $\neg(P \to Q) \equiv P \wedge \neg Q$
Biconditional Statement	$(P \leftrightarrow Q) \equiv (P \to Q) \wedge (Q \to P)$
Double Negation	$\neg(\neg P) \equiv P$
Distributive Laws	$P \vee (Q \wedge R) \equiv (P \vee Q) \wedge (P \vee R)$ $P \wedge (Q \vee R) \equiv (P \wedge Q) \vee (P \wedge R)$
Conditionals with Disjunctions	$P \to (Q \vee R) \equiv (P \wedge \neg Q) \to R$ $(P \vee Q) \to R \equiv (P \to R) \wedge (Q \to R)$

Direct Proofs

In order to prove that a conditional statement $P \to Q$ is true, we only need to prove that Q is true whenever P is true. This is because the conditional statement is true whenever the hypothesis is false. So in a direct proof of $P \to Q$, we assume that P is true, and using this assumption, we proceed through a logical sequence of steps to arrive at the conclusion that Q is true. Unfortunately, it is often not easy to discover how to start this logical sequence of steps or how to get to the conclusion that Q is true. We will describe a method of exploration that often can help in discovering the steps of a proof. This method will involve working forward from the hypothesis, P, and backward from the conclusion, Q. We will illustrate this "forward-backward" method with the following proposition.

Using the Definitions of Congruence and Divides

We will consider the following proposition and try to determine if it is true or false.

Proposition C.12. For all integers a and b, if $a \equiv 5 \pmod 8$ and $b \equiv 6 \pmod 8$, then $(a + b) \equiv 3 \pmod 8$.

Before we try to prove a proposition, it is a good idea to try some examples for which the hypothesis is true and then determine whether or not the conclusion is true for these examples. The idea is to convince ourselves that this proposition at least appears to be true. On the other hand, if we find an example where the hypothesis is true and the conclusion is false, then we have found a **counterexample** for the proposition and we would have proven the proposition to be false. The following table summarizes four examples that suggest this proposition is true.

a	b	$a+b$	Is $(a+b) \equiv 3 \pmod 8$?
5	6	11	Yes since $11 \equiv 3 \pmod 8$
13	22	35	Yes since $35 \equiv 3 \pmod 8$
-3	14	11	Yes since $11 \equiv 3 \pmod 8$
-11	-2	-13	Yes since $-13 \equiv 3 \pmod 8$

We will now attempt to construct a proof of this proposition. We will start with the backwards process. Please keep in mind that it is a good idea to write all of this down on paper. We should not try to construct a proof in our heads. Writing helps.

We know that the goal is to prove that $(a + b) \equiv 3 \pmod 8$. (We label this as statement Q.) We then ask a "backwards question" such as, "How do we prove $(a + b) \equiv 3 \pmod 8$?" We may be able to answer this question in different ways depending on whether or not we have some previously proven results, but we can always use the definition. So an answer to this question is, "We can prove that 8 divides $(a + b) - 3$." (We label this as statement $Q1$). We now ask, "How can we prove that 8 divides $(a + b) - 3$?" Again, we can use the definition and answer that we can prove that there exists an integer k such that $(a + b) - 3 = 8k$. (This is statement $Q2$.) Here is what we should have written down.

- Q: $(a + b) \equiv 3 \pmod 8$.

- $Q1$: 8 divides $(a + b) - 3$.
- $Q2$: There exists an integer k such that $(a + b) - 3 = 8k$.

The idea is that if we can prove that $Q2$ is true, then we can conclude that $Q1$ is true, and then we can conclude that Q is true. $Q2$ is a good place to stop the backwards process since it involves proving that something exists and we have an equation with which to work. So we start the forward process. We start by writing down the assumptions stated in the hypothesis of the proposition and label it statement P. We then make conclusions based on these assumptions. While doing this, we look at the items in the backward process and try to find ways to connect the conclusions in the forward process to the backward process. From statement P, we conclude that 8 divides $a - 5$ and 8 divides $b - 6$. (This becomes statement $P1$.) We make a conclusion based on statement $P1$, which becomes statement $P2$. The forward process can be summarized as follows:

- P: a and b are integers and $a \equiv 5 \pmod 8$ and $b \equiv 6 \pmod 8$.
- $P1$: 8 divides $a - 5$ and 8 divides $b - 6$.
- $P2$: There exists an integer m such that $a - 5 = 8m$ and there exists an integer n such that $b - 6 = 8n$.

It now seems that there is a way to connect the forward part $(P2)$ to the backward part $(Q2)$ using the existence of m and n (which have been proven to exist) and the equations in $P2$ and $Q2$.

Solving the two equations in $P2$ for a and b, we obtain $a = 8m + 5$ and $b = 8n + 6$. We can now use these in $Q2$.

Important Note: In the proof, we cannot use the integer k in $Q2$ since we have not proven that such an integer exists. This is why we used the letter m in statement $P2$. The goal is to prove that the integer k exists.

We can now proceed as followings:

$$\begin{aligned}(a + b) - 3 &= (8m + 5) + (8n + 6) - 3\\ &= 8m + 8n + 8\\ &= 8(m + n + 3)\end{aligned}$$

Since the integers are closed under addition, we conclude that $(m + n + 3)$ is an integer and so the last equation implies that 8 divides $(a + b) - 3$. We can now write a proof.

Proposition C.12. *For all integers a and b, if $a \equiv 5 \pmod 8$ and $b \equiv 6 \pmod 8$, then $(a + b) \equiv 3 \pmod 8$.*

Proof. We assume that a and b are integers and that $a \equiv 5 \pmod 8$ and $b \equiv 6 \pmod 8$. We will prove that $(a + b) \equiv 3 \pmod 8$. From the assumptions, we conclude that

$$8 \text{ divides } (a - 5) \text{ and } 8 \text{ divides } (b - 6).$$

So there exist integers m and n such that

$$a - 5 = 8m \text{ and } b - 6 = 8n.$$

Solving these equations for a and b, we obtain $a = 8m + 5$ and $b = 8n + 6$. We can now substitute

for a and b in the expression $(a + b) - 3$. This gives

$$\begin{aligned}(a + b) - 3 &= (8m + 5) + (8n + 6) - 3 \\ &= 8m + 8n + 8 \\ &= 8(m + n + 3)\end{aligned}$$

Since the integers are closed under addition, we conclude that $(m + n + 3)$ is an integer and so the last equation implies that 8 divides $(a + b) - 3$. So by the definition of congruence, we can conclude that $(a + b) \equiv 3 \pmod 8$. This proves that for all integers a and b, if $a \equiv 5 \pmod 8$ and $b \equiv 6 \pmod 8$, then $(a + b) \equiv 3 \pmod 8$. ■

Note: This shows a typical way to construct and write a direct proof of a proposition or theorem. We will not be going into this much detail on the construction process in all of the results proved in this book. In fact, most textbooks do not do this. What they most often show is only the final product as shown in the preceding proof. Do not be fooled that this is the way that proofs are constructed. Constructing a proof often requires trial and error and because of this, it is always a good idea to write down what is being assumed and what it is we are trying to prove. Then be willing to work backwards from what it is to be proved and work forwards from the assumptions. The hard part is often connecting the forward process to the backward process. This becomes extremely difficult if we do not write things down and try to work only in our heads.

We sometimes think that a proposition is true and attempt to write a proof. If we get stuck, we need to consider that a possible reason for this is that the proposition is actually false. Consider the following proposition.

Proposition. *For each integer n, if 7 divides $\left(n^2 - 4\right)$, then 7 divides $(n - 2)$.*

If we think about starting a proof, we would let n be an integer, assume that 7 divides $\left(n^2 - 4\right)$ and from this assumption, try to prove that 7 divides $(n - 2)$. That is, we would assume that there exists an integer k such that $n^2 - 4 = 7k$ and try to prove that there exists an integer m such that $n - 2 = 7m$. From the assumption, we can using factoring and conclude that

$$(n - 2)(n + 2) = 7k.$$

There does not seem to be a direct way to prove that there is an integer m such that $n - 2 = 7m$. So we start looking for examples of integers n such that 7 divides $\left(n^2 - 4\right)$ and see if 7 divides $(n - 2)$ for these examples. After trying a few examples, we find that for $n = 5$, 7 divides $\left(n^2 - 4\right)$. (There are many other such values for n.) For $n = 5$, we see that

$$n^2 - 4 = 21 = 7 \cdot 3 \quad \text{and} \quad n - 2 = 3.$$

However, 7 does not divide 3. This shows that for $n = 5$, the hypothesis of the proposition is true and the conclusion is false. This is a counterexample for the proposition and proves that the proposition is false.

Direct Proofs Involving Sets

One of the most basic types of proofs involving sets is to prove that one set is a subset of another set. If S and T are both subsets of some universal set U, to prove that S is a subset of T, we need to prove that

For each element x in U, if $x \in S$, then $x \in T$.

When we have to prove something that involves a universal quantifier, we frequently use a method that can be called the **choose-an-element method.** To prove that a set S is a subset of a set T, the key is that we have to prove something about all elements in S. We can then add something to the forward process by choosing an arbitrary element from the set S. This does not mean that we can choose a specific element of S. Rather, we must give the arbitrary element a name and use only the properties it has by being a member of the set S.

The truth of the next proposition may be clear, but it is included to illustrate the process of proving one set is a subset of another set. In this proposition, the set S is the set of all integers that are a multiple of 6. So when we "choose" an element from S, we are not selecting a specific element in S (such as 12 or 24), but rather we are selecting an arbitrary element of S and so the only thing we can assume is that the element is a multiple of 6.

Proposition C.13. *Let S be the set of all integers that are multiples of 6, and let T be the set of all even integers. Then S is a subset of T.*

Proof. Let S be the set of all integers that are multiples of 6, and let T be the set of all even integers. We will show that S is a subset of T by showing that if an integer x is an element of S, then it is also an element of T.

Let $x \in S$. (**Note**: The use of the word "let" is often an indication that the we are choosing an arbitrary element.) This means that x is a multiple of 6. Therefore, there exists an integer m such that

$$x = 6m.$$

Since $6 = 2 \cdot 3$, this equation can be written in the form

$$x = 2(3m).$$

By closure properties of the integers, $3m$ is an integer. Hence, this last equation proves that x must be even. Therefore, we have shown that if x is an element of S, then x is an element of T, and hence that $S \subseteq T$. ■

One way to prove that two sets are equal is to prove that each one is a subset of the other one. This is illustrated in the next proposition.

Proposition C.14. *Let A and B be subsets of some universal sets. Then $A - B = A \cap B^c$.*

Proof. Let A and B be subsets of some universal set. We will prove that $A - B = A \cap B^c$ by proving that each set is a subset of the other set. We will first prove that $A - B \subseteq A \cap B^c$. Let $x \in A - B$. We then know that $x \in A$ and $x \notin B$. However, $x \notin B$ implies that $x \in B^c$. Hence, $x \in A$ and $x \in B^c$, which means that $x \in A \cap B^c$. This proves that $A - B \subseteq A \cap B^c$.

To prove that $A \cap B^c \subseteq A - B$, we let $y \in A \cap B^c$. This means that $y \in A$ and $y \in B^c$, and hence, $y \in A$ and $y \notin B$. Therefore, $y \in A - B$ and this proves that $A \cap B^c \subseteq A - B$. Since we have proved that each set is a subset of the other set, we have proved that $A - B = A \cap B^c$. ■

Using Logical Equivalencies in Proofs

It is sometimes difficult to construct a direct proof of a conditional statement. Fortunately, there are certain logical equivalencies in Theorem C.11 on page 89 that can be used to justify some other

methods of proof of a conditional statement. Knowing that two expressions are logically equivalent tells us that if we prove one, then we have also proven the other. In fact, once we know the truth value of a statement, then we know the truth value of any other statement that is logically equivalent to it.

Using the Contrapositive

One of the most useful logical equivalencies to prove a conditional statement is that a conditional statement $P \to Q$ is logically equivalent to its contrapositive, $\neg Q \to \neg P$. This means that if we prove the contrapositive of the conditional statement, then we have proven the conditional statement. The following are some important points to remember.

- A conditional statement is logically equivalent to its contrapositive.
- To prove the statement $P \to Q$ we can use a direct proof to prove the equivalent statement that $\neg Q \to \neg P$ is true.
- Caution: One difficulty with this type of proof is in the formation of correct negations. (We need to be very careful doing this.)
- We might consider using a proof by contrapositive when the statements P and Q are stated as negations.

We will use the following proposition to illustrate how the contrapositive of a conditional statement can be used in a proof.

Proposition C.15. *For each integer n, if n^2 is an even integer, then n is an even integer.*

Proof. We will prove this result by proving the contrapositive of the statement, which is

> For each integer n, if n is an odd integer, then n^2 is an odd integer.

So we assume that n is an odd integer and prove that n^2 is an odd integer. Since n is odd, there exists an integer k such that $n = 2k + 1$. Hence,

$$\begin{aligned} n^2 &= (2k+1)^2 \\ &= 4k^2 + 4k + 1 \\ &= 2\left(2k^2 + 2k\right) + 1 \end{aligned}$$

Since the integers are closed under addition and multiplication, $\left(2k^2 + 2k\right)$ is an integer and so the last equation proves that n^2 is an odd integer. This proves that for all integers n, if n is an odd integer, then n^2 is an odd integer. Since this is the contrapositive of the proposition, we have completed a proof of the proposition. ■

Using Other Logical Equivalencies

There are many logical equivalencies, but fortunately, only a small number are frequently used when trying to construct and write proofs. Most of these are listed in Theorem C.11 on page 89. We will illustrate the use of one of these logical equivalencies with the following proposition:

For all real numbers a and b, if $a \neq 0$ and $b \neq 0$, then $ab \neq 0$.

First, notice that the hypothesis and the conclusion of the conditional statement are stated in the form of negations. This suggests that we consider the contrapositive. Care must be taken when we negate the hypothesis since it is a conjunction. We use one of De Morgan's Laws as follows:

$$\neg\left(a \neq 0 \ \wedge \ b \neq 0\right) \equiv (a = 0) \vee (b = 0)\,.$$

So the contrapositive is:

For all real numbers a and b, if $ab = 0$, then $a = 0$ or $b = 0$.

The contrapositive is a conditional statement in the form $X \to (Y \vee Z)$. The difficulty is that there is not much we can do with the hypothesis $(ab = 0)$ since we know nothing else about the real numbers a and b. However, if we knew that a was not equal to zero, then we could multiply both sides of the equation $ab = 0$ by $\dfrac{1}{a}$. This suggests that we consider using the following logical equivalency based on a result in Theorem C.11 on page 89:

$$X \to (Y \vee Z) \equiv (X \wedge \neg Y) \to Z.$$

Proposition C.16. *For all real numbers a and b, if $a \neq 0$ and $b \neq 0$, then $ab \neq 0$.*

Proof. We will prove the contrapositive of this proposition, which is

For all real numbers a and b, if $ab = 0$, then $a = 0$ or $b = 0$.

This contrapositive, however, is logically equivalent to the following:

For all real numbers a and b, if $ab = 0$ and $a \neq 0$, then $b = 0$.

To prove this, we let a and b be real numbers and assume that $ab = 0$ and $a \neq 0$. We can then multiply both sides of the equation $ab = 0$ by $\dfrac{1}{a}$. This gives

$$\frac{1}{a}(ab) = \frac{1}{a} \cdot 0.$$

We now use the associative property on the left side of this equation and simplify both sides of the equation to obtain

$$\begin{aligned} \left(\frac{1}{a} \cdot a\right) b &= 0 \\ 1 \cdot b &= 0 \\ b &= 0 \end{aligned}$$

Therefore, $b = 0$ and this proves that for all real numbers a and b, if $ab = 0$ and $a \neq 0$, then $b = 0$. Since this statement is logically equivalent to the contrapositive of the proposition, we have proved the proposition. ■

Proofs of Biconditional Statements

One of the logical equivalencies in Theorem C.11 on page 89 is the following one for biconditional statements.

$$(P \leftrightarrow Q) \equiv (P \rightarrow Q) \wedge (Q \rightarrow P)\,.$$

This logical equivalency suggests one method for proving a biconditional statement written in the form "P if and only if Q." This method is to construct separate proofs of the two conditional statements $P \rightarrow Q$ and $Q \rightarrow P$.

We will illustrate this with a proposition about right triangles.

Recall that the **Pythagorean Theorem** for right triangles states that if a and b are the lengths of the legs of a right triangle and c is the length of the hypotenuse, then $a^2 + b^2 = c^2$. We also know that the area of any triangle is one-half the base times the altitude. So for the right triangle we have described, the area is $A = \frac{1}{2}ab$.

Proposition C.17. *Suppose that a and b are the lengths of the legs of a right triangle and c is the length of the hypotenuse. This right triangle is an isosceles triangle if and only if the area of the right triangle is $\frac{1}{4}c^2$.*

Proof. We assume that we have a right triangle where a and b are the lengths of the legs of a right triangle and c is the length of the hypotenuse. We will prove that this right triangle is an isosceles triangle if and only if the area of the right triangle is $\frac{1}{4}c^2$ by proving the two conditional statements associated with this biconditional statement.

We first prove that if this right triangle is an isosceles triangle, then the area of the right triangle is $\frac{1}{4}c^2$. So we assume the right triangle is an isosceles triangle. This means that $a = b$, and consequently, $A = \frac{1}{2}a^2$. Using the Pythagorean Theorem, we see that

$$c^2 = a^2 + a^2 = 2a^2.$$

Hence, $a^2 = \frac{1}{2}c^2$, and we obtain $A = \frac{1}{2}a^2 = \frac{1}{4}c^2$. This proves that if this right triangle is an isosceles triangle, then the area of the right triangle is $\frac{1}{4}c^2$.

We now prove the converse of the first conditional statement. So we assume the area of this isosceles triangle is $A = \frac{1}{4}c^2$, and will prove that $a = b$. Since the area is also $\frac{1}{2}ab$, we see that

$$\begin{aligned} \frac{1}{4}c^2 &= \frac{1}{2}ab \\ c^2 &= 2ab \end{aligned}$$

We now use the Pythagorean Theorem to conclude that $a^2 + b^2 = 2ab$. So the last equation can be rewritten as follows:

$$\begin{aligned} a^2 - 2ab + b^2 &= 0 \\ (a-b)^2 &= 0. \end{aligned}$$

The last equation implies that $a = b$ and hence the right triangle is an isosceles triangle. This proves that if the area of this right triangle is $A = \frac{1}{4}c^2$, then the right triangle is an isosceles triangle.

Since we have proven both conditional statements, we have proven that this right triangle is an isosceles triangle if and only if the area of the right triangle is $\frac{1}{4}c^2$. ■

Proof by Contradiction

Explanation and an Example

Another method of proof that is frequently used in mathematics is a **proof by contradiction.** This method is based on the fact that a statement X can only be true or false (and not both). The idea is to prove that the statement X is true by showing that it cannot be false. This is done by assuming that X is false and proving that this leads to a contradiction. (The contradiction often has the form $(R \wedge \neg R)$, where R is some statement.) When this happens, we can conclude that the assumption that the statement X is false is incorrect and hence X cannot be false. Since it cannot be false, then X must be true.

A logical basis for the contradiction method of proof is the tautology

$$[\neg X \to C] \to X,$$

where X is a statement and C is a contradiction. The following truth table establishes this tautology.

X	C	$\neg X$	$\neg X \to C$	$(\neg X \to C) \to X$
T	F	F	T	T
F	F	T	F	T

This tautology shows that if $\neg X$ leads to a contradiction, then X must be true. The previous truth table also shows that the statement $\neg X \to C$ is logically equivalent to X. This means that if we have proved that $\neg X$ leads to a contradiction, then we have proved statement X. So if we want to prove a statement X using a proof by contradiction, we assume that $\neg X$ is true and show that this leads to a contradiction.

When we try to prove the conditional statement, "If P then Q" using a proof by contradiction, we must assume that $P \to Q$ is false and show that this leads to a contradiction. Since we are assuming the conditional statement is false, we are assuming its negation is true. According to Theorem C.11 on page 89,

$$\neg (P \to Q) \equiv P \wedge \neg Q.$$

We will illustrate the process of a proof by contradiction with the following proposition.

Proposition C.18. *For each real number x, if $0 < x < 1$, then* $\dfrac{1}{x(1-x)} \geq 4$.

Proof. We will use a proof by contradiction. So we assume that the proposition is false, or that there

exists a real number x such that $0 < x < 1$ and

$$\frac{1}{x(1-x)} < 4. \tag{C.1}$$

We note that since $0 < x < 1$, we can conclude that $x > 0$ and that $(1-x) > 0$. Hence, $x(1-x) > 0$ and if we multiply both sides of inequality (C.1) by $x(1-x)$, we obtain

$$1 < 4x(1-x).$$

We can now use algebra to rewrite the last inequality as follows:

$$\begin{aligned} 1 &< 4x - 4x^2 \\ 4x^2 - 4x + 1 &< 0 \\ (2x-1)^2 &< 0 \end{aligned}$$

However, $(2x-1)$ is a real number and the last inequality says that a real number squared is less than zero. This is a contradiction since the square of any real number must be greater than or equal to zero. Hence, the proposition cannot be false, and we have proved that for each real number x, if $0 < x < 1$, then $\dfrac{1}{x(1-x)} \geq 4$. ■

Proving that Something Does Not Exist

In mathematics, we sometimes need to prove that something does not exist or that something is not possible. Instead of trying to construct a direct proof, it is sometimes easier to use a proof by contradiction so that we can assume that the something exists.

Proposition C.19. *For all integers x and y, if x and y are odd integers, then there does not exist an integer z such that $x^2 + y^2 = z^2$.*

Proof. We will use a proof by contradiction. So we assume that the proposition is false or that there exist integers x and y such that x and y are odd and there exists an integer z such that $x^2 + y^2 = z^2$. Since x and y are odd, there exist integers m and n such that $x = 2m + 1$ and $y = 2n + 1$. So we get

$$\begin{aligned} x^2 + y^2 &= (2m+1)^2 + (2n+1)^2 \\ &= 4m^2 + 4m + 1 + 4n^2 + 4n + 1 \\ &= 2\left(2m^2 + 2m + 2n^2 + 2n + 1\right) \end{aligned} \tag{C.1}$$

Since the integers are closed under addition and multiplication, we see that
$2\left(2m^2 + 2m + 2n^2 + 2n + 1\right)$ is an integer, and so the last equation shows that $x^2 + y^2$ is an even integer. Hence, z^2 is even since $z^2 = x^2 + y^2$. So using the result in Proposition C.15 on page 94, we can conclude that z is even and that there exists an integer k such that $z = 2k$. Now, using equation (1) above, we see that

$$\begin{aligned} z^2 &= 2\left(2m^2 + 2m + 2n^2 + 2n + 1\right) \\ (2k)^2 &= 2\left(2m^2 + 2m + 2n^2 + 2n + 1\right) \\ 4k^2 &= 2\left(2m^2 + 2m + 2n^2 + 2n + 1\right) \end{aligned}$$

Dividing both sides of the last equation by 2, we obtain

$$4k^2 = 2\left(2m^2 + 2m + 2n^2 + 2n + 1\right)$$
$$2k^2 = 2\left(m^2 + m + n^2 + n\right) + 1$$

However, the left side of the last equation is an even integer and the right side is an odd integer. This is a contradiction, and so the proposition cannot be false. Hence, we have proved that for all integers x and y, if x and y are odd integers, then there does not exist an integer z such that $x^2+y^2 = z^2$. ■

Rational and Irrational Numbers

One of the most important ways to classify real numbers is as a rational number or an irrational number. (See the section on subsets of the real numbers in Investigation 3 of the textbook.)

Definition C.20. A real number x is defined to be a **rational number** provided that there exist integers m and n with $n \neq 0$ such that $x = \dfrac{m}{n}$. A real number that is not a rational number is called an **irrational number.**

We use the symbol $\mathbb{Q}$ to stand for the set of rational numbers.

Because the rational numbers are closed under the standard operations and the definition of an irrational number simply says that the number is not rational, we often use a proof by contradiction to prove that a number is irrational. This is illustrated in the next proposition.

Proposition C.21. *For all real numbers x and y, if x is rational and $x \neq 0$ and y is irrational, then $x \cdot y$ is irrational.*

Proof. We will use a proof by contradiction. So we assume that there exist real numbers x and y such that x is rational, $x \neq 0$, y is irrational, and $x \cdot y$ is rational. Since $x \neq 0$, we can divide by x, and since the rational numbers are closed under division by nonzero rational numbers, we know that $\dfrac{1}{x} \in \mathbb{Q}$. We now know that $x \cdot y$ and $\dfrac{1}{x}$ are rational numbers and since the rational numbers are closed under multiplication, we conclude that

$$\frac{1}{x} \cdot (xy) \in \mathbb{Q}.$$

However, $\dfrac{1}{x} \cdot (xy) = y$ and hence, y must be a rational number. Since a real number cannot be both rational and irrational, this is a contradiction to the assumption that y is irrational. We have therefore proved that for all real numbers x and y, if x is rational and $x \neq 0$ and y is irrational, then $x \cdot y$ is irrational. ■

Using Cases in Proofs

The method of using cases in a proof is often used when the hypothesis of a proposition is a disjunction. This is justified by the logical equivalency

$$[(P \vee Q) \to R] \equiv [(P \to R) \wedge (Q \to R)]\,.$$

This is one of the logical equivalencies in Theorem C.11 on page 89. In some other situations when we are trying to prove a proposition or a theorem about an element x in some set U, we often run into the problem that there does not seem to be enough information about x to proceed. For example, consider the following proposition:

Proposition. *If n is an integer, then $\left(n^2 + n\right)$ is an even integer.*

If we were trying to write a direct proof of this proposition, the only thing we could assume is that n is an integer. This is not much help. In a situation such as this, we will sometimes construct our own cases to provide additional assumptions for the forward process of the proof. Cases are usually based on some common properties that the given element may or may not possess. The cases must be chosen so that they exhaust all possibilities for the object in the hypothesis of the proposition. For this propostion, we know that an integer must be even or it must be odd. We can thus use the following two cases for the integer n:

- The integer n is an even integer; or
- The integer n is an odd integer.

Proposition C.22. *If n is an integer, then $\left(n^2 + n\right)$ is an even integer.*

Proof. We assume that n is an integer and will prove that $\left(n^2 + n\right)$. Since we know that any integer must be even or odd, we will use two cases. The first is that n is an even integer, and the second is that n is an odd integer.

In the case where n is an even integer, there exists an integer m such that

$$n = 2m.$$

Substituting this into the expression $n^2 + n$ yields

$$\begin{aligned} n^2 + n &= (2m)^2 + 2m \\ &= 4m^2 + 2m \\ &= 2(2m^2 + m) \end{aligned}$$

By the closure properties of the integers, $2m^2 + m$ is an integer, and hence $n^2 + n$ is even. So this proves that when n is an even integer, $n^2 + n$ is an even integer.

In the case where n is an odd integer, there exists an integer k such that

$$n = 2k + 1.$$

Substituting this into the expression $n^2 + n$ yields

$$\begin{aligned} n^2 + n &= (2k + 1)^2 + (2k + 1) \\ &= \left(4k^2 + 4k + 1\right) + 2k + 1 \\ &= \left(4k^2 + 6k + 2\right) \\ &= 2\left(2k^2 + 3k + 1\right) \end{aligned}$$

By the closure properties of the integers, $2k^2 + 3k + 1$ is an integer, and hence $n^2 + n$ is even. So this proves that when n is an odd integer, $n^2 + n$ is an even integer.

Since we have proved that $n^2 + n$ is even when n is even and when n is odd, we have proved that if n is an integer, then $\left(n^2 + n\right)$ is an even integer. ■

Some Common Situations to Use Cases

When using cases in a proof, the main rule is that the cases must be chosen so that they exhaust all possibilities for an object x in the hypothesis of the original proposition. Following are some common uses of cases in proofs.

When the hypothesis is, "n is an integer."	Case 1: n is an even integer. Case 2: n is an odd integer.		
When the hypothesis is, "m and n are integers."	Case 1: m and n are even. Case 2: m is even and n is odd. Case 3: m is odd and n is even. Case 4: m and n are both odd.		
When the hypothesis is, "x is a real number."	Case 1: x is rational. Case 2: x is irrational.		
When the hypothesis is, "x is a real number."	Case 1: $x = 0$. Case 2: $x \neq 0$.	OR	Case 1: $x > 0$. Case 2: $x = 0$. Case 3: $x < 0$.
When the hypothesis is, "a and b are real numbers."	Case 1: $a = b$. Case 2: $a \neq b$.	OR	Case 1: $a > b$. Case 2: $a = b$. Case 3: $a < b$.

Using Cases with the Division Algorithm

In Investigation 1 of the textbook, we introduced an important result for the set of integers is known as the Division Algorithm, which is stated below.

The Division Algorithm
Let a and b be integers with $a > 0$. Then there exist unique integers q and r such that

$$b = aq + r \text{ and } 0 \leq r < a.$$

When we speak of **the quotient** and **the remainder** when we "divide an integer b by the positive integer a," we will always mean the quotient (q) and the remainder (r) guaranteed by the Division Algorithm. So the remainder r is the least nonnegative integer such that there exists an integer (quotient) q with $b = aq + r$.

The Division Algorithm can sometimes be used to construct cases for a proof dealing with the integers. For example, if the hypothesis of a proposition is that "n is an integer," then we can use the Division Algorithm to claim that there are unique integers q and r such that

$$n = 3q + r \text{ and } 0 \leq r < 3.$$

We can then divide the proof into the following three cases: (1) $r = 0$; (2) $r = 1$; and (3) $r = 2$. This is done in Proposition C.23.

Proposition C.23. *If n is an integer, then 3 divides $n^3 - n$.*

Proof. Let n be an integer. We will show that 3 divides $n^3 - n$ by examining the three cases for the remainder when n is divided by 3. By the Division Algorithm, there exist unique integers q and r such that

$$n = 3q + r, \text{ and } 0 \leq r < 3.$$

This means that we can consider the following three cases: (1) $r = 0$; (2) $r = 1$; and (3) $r = 2$. In the case where $r = 0$, we have $n = 3q$. By substituting this into the expression $n^3 - n$, we get

$$\begin{aligned} n^3 - n &= (3q)^3 - (3q) \\ &= 27q^3 - 3q \\ &= 3\left(9q^3 - q\right). \end{aligned}$$

Since $\left(9q^3 - q\right)$ is an integer, the last equation proves that $3 \mid \left(n^3 - n\right)$. In the second case, $r = 1$ and $n = 3q + 1$. When we substitute this into $\left(n^3 - n\right)$, we obtain

$$\begin{aligned} n^3 - n &= (3q + 1)^3 - (3q + 1) \\ &= \left(27q^3 + 27q^2 + 9q + 1\right) - (3q + 1) \\ &= 27q^3 + 27q^2 + 6q \\ &= 3\left(9q^3 + 9q^2 + 2q\right). \end{aligned}$$

Since $\left(9q^3 + 9q^2 + 2q\right)$ is an integer, the last equation proves that $3 \mid \left(n^3 - n\right)$.

For the third case, $r = 2$ and $n = 3q + 2$. When we substitute this into $\left(n^3 - n\right)$, we obtain

$$\begin{aligned} n^3 - n &= (3q + 2)^3 - (3q + 2) \\ &= \left(27q^3 + 54q^2 + 36q + 8\right) - (3q + 2) \\ &= 27q^3 + 54q^2 + 33q + 6 \\ &= 3\left(9q^3 + 18q^2 + 11q + 2\right). \end{aligned}$$

Since $\left(9q^3 + 18q^2 + 11q + 2\right)$ is an integer, the last equation proves that $3 \mid \left(n^3 - n\right)$. We have now proved that 3 divides $(n^3 - n)$ in all 3 of the possible cases. Therefore, we have proved that for each integer n, 3 divides $(n^3 - n)$. ■

Exercises

(1) Construct a table of values for $\left(3m^2 + 4m + 6\right)$ using at least six different integers for m. Make one-half of the values for m even integers and the other half odd integers. Is the following proposition true or false?

If m is an odd integer, then $\left(3m^2 + 4m + 6\right)$ is an odd integer.

Justify your conclusion. (If the proposition is true, then write a proof of the proposition. If the proposition is false, provide an example of an odd integer for which $\left(3m^2 + 4m + 6\right)$ is an even integer.)

(2) The **Pythagorean Theorem** for right triangles states that if a and b are the lengths of the legs of a right triangle and c is the length of the hypotenuse, then $a^2 + b^2 = c^2$. For example, if $a = 5$ and $b = 12$ are the lengths of the two sides of a right triangle and if c is the length of the hypotenuse, then the $c^2 = 5^2 + 12^2$ and so $c^2 = 169$. Since c is a length and must be positive, we conclude that $c = 13$.

Construct and provide a well-written proof for the following proposition.

Proposition. If m is a real number and m, $m+1$, and $m+2$ are the lengths of the three sides of a right triangle, then $m = 3$.

(3) One way to prove that two sets are equal is to prove that each one is a subset of the other one. Consider the following proposition:

Proposition. *Let A and B be subsets of some universal set. Then $A - (A - B) = A \cap B$.*

Prove this proposition is true or give a counterexample to prove it is false.

(4) Are the following statements true or false? Justify your conclusions.

(a) For each $a \in \mathbb{Z}$, if $a \equiv 2 \pmod 5$, then $a^2 \equiv 4 \pmod 5$.

(b) For each $a \in \mathbb{Z}$, if $a^2 \equiv 4 \pmod 5$, then $a \equiv 2 \pmod 5$.

(c) For each $a \in \mathbb{Z}$, $a \equiv 2 \pmod 5$ if and only if $a^2 \equiv 4 \pmod 5$.

(5) A real number x is defined to be a **rational number** provided

$$\text{there exist integers } m \text{ and } n \text{ with } n \neq 0 \text{ such that } x = \frac{m}{n}.$$

A real number that is not a rational number is called an **irrational number.**

It is known that if x is a positive rational number, then there exist positive integers m and n with $n \neq 0$ such that $x = \dfrac{m}{n}$.

Is the following proposition true or false? Explain.

Proposition. For each positive real number x, if x is irrational, then $\sqrt{x}$ is irrational.

(6) (a) Determine at least five different integers that are congruent to 2 modulo 4. Are any of these integers congruent to 3 modulo 6?

(b) Is the following proposition true or false? Justify your conclusion with a counterexample (if it is false) or a proof (if it is true).

Propostion. For each integer n, if $n \equiv 2 \pmod 4$, then $n \not\equiv 3 \pmod 6$.

(7) For the following, it may be useful to use the facts that the set of rational numbers $\mathbb{Q}$ is closed under addition, subtraction, multiplication, and division by nonzero rational numbers.

Prove the following proposition:

Proposition. For all real numbers x and y, if x is rational and y is irrational, then $x + y$ is irrational.

(8) Consider the following proposition:

Proposition. For each integer a, if 3 divides a^2, then 3 divides a.

(a) Write the contrapositive of this proposition.

(b) Prove the proposition by proving its contrapositive. **Hint**: Consider using cases based on the Division Algorithm.

(9) Complete the details for the proof of Case 3 of Proposition C.23.

(10) Is the following proposition true or false? Justify your conclusion with a counterexample or a proof.

Proposition. For each integer n, if n is odd, then 8 divides $n^2 - 1$.

Appendix D

Proof that $R[x]$ is a Ring

In this appendix, we will give a formal proof that $R[x]$ is a commutative ring when R is a commutative ring. Before we give the proof, we will show how to write the sum and product of two polynomials using summation notation. Let $f(x), g(x) \in R[x]$ with

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 \text{ with } a_n \neq 0, \text{ and}$$
$$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_2 x^2 + b_1 x + b_0 \text{ with } b_m \neq 0.$$

Since it must be true that $m \leq n$ or $n \leq m$, we can assume that $m \leq n$ without loss of generality. We will then use the fact that $b_{m+1} = b_{m+2} = \cdots = b_n = 0$, and so we can write

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0, \text{ and}$$
$$g(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_2 x^2 + b_1 x + b_0.$$

Using summation notation, we can write the sum and product of these two polynomials as follows:

$$f(x) + g(x) = \sum_{i=0}^{n} a_i x^i + \sum_{i=0}^{n} b_i x^i = \sum_{i=0}^{n} (a_i + b_i)\, x^i, \quad \text{and}$$
$$f(x)g(x) = \left(\sum_{i=0}^{n} a_i x^i\right)\left(\sum_{i=0}^{n} b_i x^i\right) = \sum_{j=0}^{n+m}\left(\sum_{i=0}^{j}(a_{j-i}b_i)\right) x^j, \quad \text{or equivalently}$$
$$f(x)g(x) = \sum_{j=0}^{n+m}\left(\sum_{r+s=j}(a_r b_s)\right) x^j.$$

Theorem D.-1. *If R is a commutative ring, then $R[x]$ is a commutative ring. In addition, if the ring R has an identity, then the ring $R[x]$ has an identity.*

Proof. Let R be a commutative ring and let

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$
$$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0, \text{and}$$
$$h(x) = c_k x^k + c_{k-1}x^{k-1} + \cdots + c_1 x + c_0$$

be elements of $R[x]$. In proving the ring properties for $R[x]$, we will assume (without loss of generality) that $k \leq m \leq n$ and extend the polynomials $g(x)$ and $h(x)$ to have n coefficients. This means that for $m < i \leq n$, $b_i = 0$, and for $k < i \leq n$, $c_i = 0$.

The definitions of addition and multiplication of polynomials show that $f(x) + g(x)$ and $f(x)g(x)$ are polynomials in $R[x]$, so $R[x]$ is closed under addition and multiplication.

To show that addition is commutative in $R[x]$, we use the commutativity of addition in R and obtain

$$f(x) + g(x) = \sum_{i=0}^{n} (a_i + b_i)x^i = \sum_{i=0}^{n} (b_i + a_i)x^i = g(x) + f(x).$$

Thus, addition is commutative in $R[x]$.

For associativity of addition in $R[x]$, the associativity of addition in R gives us

$$\begin{aligned}
[f(x) + g(x)] + h(x) &= \left(\sum_{i=0}^{n} (a_i + b_i)x^i \right) + \sum_{i=0}^{n} c_i x^k \\
&= \sum_{i=0}^{n} [(a_i + b_i) + c_i]x^i \\
&= \sum_{i=0}^{n} [a_i + (b_i + c_i)]x^i \\
&= \sum_{i=0}^{n} a_i x^i + \sum_{i=0}^{n} (b_i + c_i)x^i \\
&= f(x) + [g(x) + h(x)].
\end{aligned}$$

Therefore, addition is associative in $R[x]$.

We can show that an additive identity in $R[x]$ is the polynomial $z(x) = 0_R$, the polynomial all of whose coefficients are 0_R, as follows:

$$f(x) + z(x) = \sum_{i=1}^{n} a_i x^i + z(x) = \sum_{i=1}^{n} (a_i + 0_R)\, x^i = \sum_{i=0}^{n} a_i x^i = f(x).$$

Therefore, $z(x)$ is an additive identity in $R[x]$.

We will now show that $R[x]$ contains an additive inverse for $f(x)$ (and hence for any element of $R[x]$). For each i with $0 \leq i \leq n$, we know that $a_i \in R$ and hence, a_i has an additive inverse, $-a_i \in R$. Let

$$q(x) = \sum_{i=0}^{n} (-a_i)x^i = (-a_n)x^n + (-a_{n-1})x^{n-1} + \cdots + (-a_1)x + (-a_0).$$

Then $q(x) \in R[x]$ and

$$f(x) + q(x) = \sum_{i=0}^{n} [a_i + (-a_i)]\, x^i = \sum_{i=0}^{n} 0_R x^i = z(x).$$

Therefore, $R[x]$ contains an additive inverse for each of its elements.

We now turn our attention to multiplication in $R[x]$. We will first show that multiplication is commutative. Using the definition of multiplication of polynomials, we see that

$$\begin{aligned}
f(x)g(x) &= \left(\sum_{i=0}^{n} a_i x^i \right) \left(\sum_{i=0}^{m} b_i x^i \right) \\
&= \left(\sum_{i=0}^{n+m} \left(\sum_{r+s=i} (a_r b_s) \right) x^i \right). \qquad \text{(D.1)}
\end{aligned}$$

Since $s + r = r + s = i$ and multiplication in R is commutative, we see that $a_r b_s = b_s a_r$ and hence, we can rewrite equation (D.1) as follows:

$$\begin{aligned} f(x)g(x) &= \left(\sum_{i=0}^{n+m} \left(\sum_{s+r=i} (b_s a_r) \right) x^i \right) \\ &= \left(\sum_{i=0}^{n} a_i x^i \right) \left(\sum_{i=0}^{n} b_i x^i \right) \\ &= g(x)f(x). \end{aligned}$$

This shows that multiplication in $R[x]$ is commutative.

We will now show that multiplication is an associative operation in $R[x]$. (Note that the notation in this part of the proof can be a bit overwhelming; this is why we explored a special case of the associative property of multiplication in Activity 8.16. It will help in understanding the notation in this proof to have the work from this activity to refer to while reading the proof.) We will start with the formula for $f(x)g(x)$ in equation (D.1). To simplify the notation, for $0 \leq i \leq n + m$, we let $u_i = \sum_{r+s=i} a_r b_s$. We then obtain

$$\begin{aligned} [f(x)g(x)]h(x) &= \left(\sum_{i=0}^{n+m} u_i x^i \right) \left(\sum_{i=0}^{k} c_i x^i \right) \\ &= \sum_{j=0}^{(n+m)+k} \left(\sum_{p+q=j} u_p c_q \right) x^j. \end{aligned}$$

We can now substitute for u_p and then use the fact that

$$\left(\sum_{r+s=p} a_r b_s \right) c_q = \sum_{r+s=p} (a_r b_s)\, c_q,$$

which is true by the distributive property in R. This gives

$$\begin{aligned} [f(x)g(x)]h(x) &= \sum_{j=0}^{(n+m)+k} \left[\sum_{p+q=j} \left(\sum_{r+s=p} a_r b_s \right) c_q \right] x^j \\ &= \sum_{j=0}^{(n+m)+k} \left[\sum_{p+q=j} \left(\sum_{r+s=p} (a_r b_s)\, c_q \right) \right] x^j. \end{aligned} \tag{D.2}$$

In equation (D.2), notice that $r+s+q = p+q = j$, and so equation (D.2) shows that the coefficient of x^j in $[f(x)g(x)]h(x)$ is the sum of all products of the form $(a_r b_s)\, c_q$ where $r + s + q = j$. This means that we can rewrite equation (D.2) as follows:

$$[f(x)g(x)]h(x) = \sum_{j=0}^{(n+m)+k} \left[\sum_{r+s+q=j} (a_r b_s)\, c_q \right] x^j. \tag{D.3}$$

Using a similar procedure for $f(x)[g(x)h(x)]$, we see that

$$\begin{aligned} f(x)[g(x)h(x)] &= \left(\sum_{i=0}^{n} a_i x^i\right)\left[\sum_{w=0}^{m+k}\left(\sum_{s+q=w} b_s c_q\right)x^w\right] \\ &= \sum_{j=0}^{n+(m+k)}\left[\sum_{r+w=j}\left(a_r \sum_{s+q=w} b_s c_q\right)\right]x^j \\ &= \sum_{j=0}^{n+(m+k)}\left[\sum_{r+w=j}\left(\sum_{s+q=w} a_r\left(b_s c_q\right)\right)\right]x^j. \end{aligned} \tag{D.4}$$

In equation (D.4), $r+s+q = r+w = j$, and so equation (D.4) shows that the coefficient of x^j in $f(x)[g(x)h(x)]$ is the sum of all products of the form $a_r\left(b_s c_q\right)$ where $r+s+q=j$. Therefore, we can write

$$f(x)[g(x)h(x)] = \sum_{j=0}^{n+(m+k)}\left[a_r\left(b_s c_q\right)\right]x^j. \tag{D.5}$$

Since multiplication in R is associative, we know that $a_r\left(b_s c_q\right) = \left(a_r b_s\right)c_q$, so using this and equation (D.5), we can conclude that

$$f(x)[g(x)h(x)] = \sum_{j=0}^{n+(m+k)}\left[\left(a_r b_s\right)c_q\right]x^j. \tag{D.6}$$

Comparing equations (D.3) and (D.6), we see that $[f(x)g(x)]h(x) = f(x)[g(x)h(x)]$, which proves that multiplication in $R[x]$ is associative.

We will now prove the distributive law. (Since we have proved that multiplication in $R[x]$ is commutative, we only have to prove one of the distributive laws.) For this, recall that we have assumed that $k \leq m \leq n$, and so

$$\begin{aligned} f(x)[g(x)+h(x)] &= \left(\sum_{i=0}^{n} a_i x^i\right)\left(\sum_{j=0}^{m}(b_j + c_j)x^j\right) \\ &= \sum_{j=0}^{n+m}\left[\sum_{r+s=j} a_r(b_s + c_s)\right]x^j \\ &= \sum_{j=0}^{n+m}\left[\sum_{r+s=j}\left(a_r b_s + a_r c_s\right)\right]x^j \\ &= \sum_{j=0}^{n+m}\left(\sum_{r+s=j} a_r b_s\right)x^j + \sum_{j=0}^{n+m}\left(\sum_{r+s=j} a_r c_s\right)x^j \\ &= f(x)g(x) + f(x)h(x). \end{aligned}$$

Therefore, multiplication distributes over addition in $R[x]$, and we conclude that $R[x]$ is a ring.

Finally, we will prove that if R has an identity, 1_R, then $R[x]$ also contains an identity. Let

$u(x) = 1_R$. Then $u(x) \in R[x]$ and

$$\begin{aligned} f(x)u(x) &= \left(\sum_{i=0}^{n} a_i x^i \right) (1_R) \\ &= f(x). \end{aligned}$$

Therefore, if R is a commutative ring with identity, then $R[x]$ is also a commutative ring with identity. ■

Appendix E

The Cubic Formula

The quadratic formula tells us how to find all solutions to quadratic equations in $\mathbb{C}[x]$. There is also a general formula for solving cubic equations, although it is much more complicated than the quadratic formula.

First note that when we want to find roots of polynomials, it suffices to work with only monic polynomials. To see why, consider the general polynomial $f(x) = \sum_{i=0}^{n} a_i x^i$ (with $a_n \neq 0$) in $F[x]$, where F is a field. Let $g(x) = x^n + \sum_{i=0}^{n+1} a_n^{-1} a_i x^i$, and assume $r \in R$ is a root of $f(x)$. Then $f(r) = 0$, and so

$$a_n r^n + a_{n-1} r^{n-1} + \cdots + a_1 r + a_0 = 0.$$

Multiplying both sides by a_n^{-1} gives us

$$r^n + a_n^{-1} a_{n-1} r^{n-1} + \cdots + a_n^{-1} a_1 r + a_n^{-1} a_0 = a_n^{-1} 0 = 0.$$

But the left hand side of the last equation is just $g(r)$. Thus, $g(r) = 0$.

Now assume $r \in R$ is a root of $g(x)$. Then $g(r) = 0$. So

$$r^n + a_n^{-1} a_{n-1} r^{n-1} + \cdots + a_n^{-1} a_1 r + a_n^{-1} a_0 = 0.$$

Multiplying both sides by a_n yields

$$a_n r^n + a_{n-1} r^{n-1} + \cdots + a_1 r + a_0 = a_n 0 = 0.$$

But the left hand side of the last equation is just $f(r)$. Thus, $f(r) = 0$. What we have shown is that r is a root of $f(x)$ if and only if r is a root of $g(x)$.

One conclusion we can draw from this is that when looking for roots of polynomials over a field, it is enough to consider roots of monic polynomials. So when finding roots of cubics, it suffices to consider only cubics of the form $x^3 + ax^2 + bx + c$. Let $a, b, c \in \mathbb{C}$, and consider the cubic equation

$$x^3 + ax^2 + bx + c = 0. \tag{E.1}$$

Our first step in solving this cubic equation is to reduce the cubic polynomial $x^3 + ax^2 + bx + c$ to what is often called a *depressed* cubic—that is, a cubic of the form $x^3 + px + q$. One way to do this is to make the change of variable $x = z - \frac{a}{3}$.

Activity E.1.

(a) Evaluate the cubic polynomial $x^3 + 6x^2 + x + 3$ at $x = z - \frac{a}{3}$ and show that the result is the depressed cubic $z^3 - 11z + 17$.

(b) Evaluate the general cubic $x^3 + ax^2 + bx + c$ at $x = z - \frac{a}{3}$ and show that

$$x^3 + ax^2 + bx + c = z^3 + \left(\frac{3b - a^2}{3}\right) z + \left(\frac{2a^3 - 9ab + 27c}{27}\right).$$

Activity E.1 shows that the substitution $x = z - \frac{a}{3}$ transforms our general cubic $x^3 + ax^2 + bx + c$ to the depressed cubic $z^3 + pz + q = 0$ with $p = \frac{3b - a^2}{3}$ and $q = \frac{2a^3 - 9ab + 27c}{27}$. Now we just need to solve the depressed cubic.

Theorem E.2. *Let $p \in \mathbb{C}$ be nonzero, and let $q \in \mathbb{C}$. The roots of*

$$z^3 + pz + q = 0 \tag{E.2}$$

are given by $z = \sqrt[3]{A} - \frac{p}{3\sqrt[3]{A}}$, where $A = -\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}$ and $\sqrt[3]{A}$ can be any one of the three cube roots of A in $\mathbb{C}$.

Proof. Our first step will be to reduce equation (E.2) to a quadratic. We do this by substituting $y - \frac{p}{3y}$ for z to obtain

$$\begin{aligned}
\left(y - \frac{p}{3y}\right)^3 + p\left(y - \frac{p}{3y}\right) + q &= 0 \\
y^3 - 3\left(\frac{p}{3y}\right)y^2 + 3\left(\frac{p}{3y}\right)^2 y - \left(\frac{p}{3y}\right)^3 + py - p\left(\frac{p}{3y}\right) + q &= 0 \\
y^3 - py + \left(\frac{p^2}{3y}\right) - \left(\frac{p^3}{27y^3}\right) + py - \left(\frac{p^2}{3y}\right) + q &= 0 \\
y^3 - \left(\frac{p^3}{27}\right)\left(\frac{1}{y^3}\right) + q &= 0 \\
\left(y^3\right)^2 + qy^3 - \left(\frac{p^3}{27}\right) &= 0.
\end{aligned} \tag{E.3}$$

Setting $v = y^3$ transforms the last equation into the quadratic equation

$$v^2 + qv - \left(\frac{p^3}{27}\right) = 0. \tag{E.4}$$

We can solve equation (E.4) with the quadratic formula, which yields the following two roots:

$$A = \frac{-q + \sqrt{q^2 + 4\left(\frac{p^3}{27}\right)}}{2} \qquad \text{and} \qquad B = \frac{-q - \sqrt{q^2 + 4\left(\frac{p^3}{27}\right)}}{2}.$$

After some simplifying, we see that

$$A = -\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}} \qquad \text{and} \qquad B = -\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}.$$

So our solutions to equation (E.3) are the cube roots of A and B. At first glance it might appear that there are then 6 solutions to equation (E.4) and therefore 6 solutions to equation (E.2). However, if r is a cube root of A, then

$$\left(-\frac{p}{3r}\right)^3 = -\frac{p^3}{27r^3} = \frac{p^3}{27A}.$$

Since

$$AB = \frac{q^2}{4} - \left(\frac{q^2}{4} + \frac{p^3}{27}\right) = \frac{p^3}{27},$$

we see that

$$\left(-\frac{p}{3r}\right)^3 = -\frac{p^3}{27r^3} = \frac{p^3}{27A} = B,$$

and so $-\frac{p}{3r}$ is a cube root of B. So if we let r_1, r_2, and r_3 be the cube roots of A, then $s_1 = -\frac{p}{3r_1}$, $s_2 = -\frac{p}{3r_2}$, and $s_3 = -\frac{p}{3r_3}$ are the cube roots of B. Since $z = y - \frac{p}{3y}$ is equal to $r_i + s_i$ whenever $y = r_i$ or $y = s_i$, it follows that the solutions to equation (E.2) are $r_1 + s_1$, $r_2 + s_2$, and $r_3 + s_3$. ■

To illustrate the cubic formula, we will find the solutions to the cubic equation $x^3 + x^2 + x + 1 = 0$. We begin by reducing this equation to the depressed cubic

$$z^3 + \frac{2}{3}z + \frac{20}{27} = 0$$

by substituting $x = z - \frac{1}{3}$, where $p = \frac{3(1)-(1)^2}{3} = \frac{2}{3}$ and $q = \frac{2(1)^3-9(1)(1)+27(1)}{27} = \frac{20}{27}$. To solve the depressed cubic, we need to find the cube roots of

$$A = -\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}} = -\frac{10}{27} + \frac{2\sqrt{3}}{9}.$$

Let $r = \sqrt[3]{A}$, and let ω be the primitive cube root* of 1 given by $\omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$. Then, using the same notation as in the proof of Theorem E.2, we obtain:

$$\begin{aligned} r_1 &= r\omega & s_1 &= -\frac{p}{3r_1} = -\frac{2}{9r\omega} = -\frac{2}{9r}\omega^2 \\ r_2 &= r\omega^2 & s_2 &= -\frac{p}{3r_2} = -\frac{2}{9r}\omega \\ r_3 &= r & s_3 &= -\frac{p}{3r_3} = -\frac{2}{9r}. \end{aligned}$$

So the roots of the depressed polynomial $z^3 + \frac{2}{3}z + \frac{20}{27}$ are

$$\begin{aligned} z_1 &= r\omega - \frac{2}{9r}\omega^2 \\ z_2 &= r\omega^2 - \frac{2}{9r}\omega \\ z_3 &= r - \frac{2}{9r}. \end{aligned}$$

We can simplify these roots a bit. First, note that $\left(-\frac{1}{3} + \frac{\sqrt{3}}{3}\right)^3 = -\frac{10}{27} + \frac{2\sqrt{3}}{9}$, so $r = -\frac{1}{3} + \frac{\sqrt{3}}{3}$.

*If A is a real number and $\omega = \cos(\frac{2\pi}{n}) + i\sin(\frac{2\pi}{n})$, then ω is called a *primitive* n^{th} root of 1, and the n complex n^{th} roots of A are given by $\sqrt[n]{A}, \omega\sqrt[n]{A}, \omega^2\sqrt[n]{A}, \ldots, \omega^{n-1}\sqrt[n]{A}$.

Then

$$\begin{aligned}
z_3 &= r - \frac{2}{9r} \\
&= -\frac{1}{3} + \frac{\sqrt{3}}{3} - \frac{2}{9\left(-\frac{1}{3} + \frac{\sqrt{3}}{3}\right)} \\
&= -\frac{1}{3} + \frac{\sqrt{3}}{3} - \frac{2}{9\left(-\frac{1}{3} + \frac{\sqrt{3}}{3}\right)} \left(\frac{-\frac{1}{3} - \frac{\sqrt{3}}{3}}{-\frac{1}{3} - \frac{\sqrt{3}}{3}}\right) \\
&= -\frac{1}{3} + \frac{\sqrt{3}}{3} - \frac{2\left(-\frac{1}{3} - \frac{\sqrt{3}}{3}\right)}{9\left(\frac{1}{9} - \frac{3}{9}\right)} \\
&= -\frac{1}{3} + \frac{\sqrt{3}}{3} + \left(-\frac{1}{3} - \frac{\sqrt{3}}{3}\right) \\
&= -\frac{2}{3}.
\end{aligned}$$

Similar simplification gives $z_1 = \frac{1}{3} + i$ and $z_2 = \frac{1}{3} - i$. The solutions to our original equation $x^3 + x^2 + x + 1 = 0$ have the form $x_i = z_i - \frac{a}{3} = z_i - \frac{1}{3}$. So the solutions to $x^3 + x^2 + x + 1 = 0$ are

$$x_1 = i, \quad x_2 = -i, \text{ and } \quad x_3 = -1.$$

Theorem E.2 shows us how to find the roots of cubic polynomials in $\mathbb{C}[x]$. There is a corresponding formula for finding roots of quartic (degree 4) polynomials as well, but we won't consider that formula here. When we use the quadratic and cubic formulas to solve equations, we are finding the solutions in a form that only depends on the sums, differences, products, or quotients of the coefficients of the polynomial along with roots (square, cube, etc.) of such combinations of the coefficients. When we do this, we say we are *solving an equation by radicals*. Some of the best mathematicians throughout history, including Euler and Lagrange, attempted to find solutions by radicals of general quintic (degree 5) polynomial equations over $\mathbb{C}$. It wasn't until 1826 that the first generally accepted proof of the insolvability of quintic polynomials was published by Abel. Galois later developed a theory of solvability of equations involving groups and fields, and he used this theory to show that polynomial equations of degree 5 or higher over $\mathbb{C}$ are not solvable by radicals.

Appendix F

The Fundamental Theorem of Algebra

In this appendix we provide a proof of the Fundamental Theorem of Algebra. This particular proof comes from the paper "The Fundamental Theorem of Algebra" by Frode Terkelsen in *The American Mathematical Monthly*, Vol. 83, No. 8, (Oct. 1976), p. 647. As is true with most proofs of this theorem, some complex analysis is required, and we will gloss over those points. While our proof will not be complete and rigorous, it is instructive to have some idea of how this important theorem is proved. For complete details, consult a text on complex analysis.

Theorem (The Fundamental Theorem of Algebra.)**.** *Every polynomial of degree 1 or greater in* $\mathbb{C}[x]$ *has a root in* $\mathbb{C}$*.*

Proof. Let $f(z) \in \mathbb{C}[z]$ be a non-constant polynomial. We know that $|f(z)| \to \infty$ as $|z| \to \infty$, and along with the continuity of the polynomial $f(z)$, this implies the existence of a complex number z_0 such that $|f(z_0)| \leq |f(z)|$ for all $z \in \mathbb{C}$. In other words, the function $|f(z)|$ attains a minimum value.* (Recall that the norm $|f(z)|$ of the complex number $f(z)$ is a nonnegative real number.) We can always translate f so that z_0 is at the origin (by considering the polynomial $f(z + z_0)$), so we will assume $z_0 = 0$ without loss of generality. Our job is then to show that $f(0) = 0$. We will proceed by contradiction and assume $f(0) \neq 0$.

First, we will rewrite $f(z)$ in a more useful form. By subtracting out the constant term of $f(z)$, we obtain a polynomial whose smallest degree term is n for some $n \geq 1$. Thus,

$$f(z) = a_0 + a_n z^n + z^{n+1} Q(z),$$

where $a_n \neq 0$ and $Q(z)$ is a polynomial. (As an illustration, consider the polynomial $f(z) = 2 + 3z^3 + 4z^4 + 5z^6$. Note that $n = 3$ and $Q(z) = 4 + 5z^2$ in this case.) For ease of notation, we let $a = a_0$ and $b = a_n$. Note that $f(0) \neq 0$ implies $a \neq 0$. We will now use the fact that every complex number has n^{th} roots. In particular, let ω be an n^{th} root of $-\frac{a}{b}$—that is $\omega^n = -\frac{a}{b}$. For each real number x, we have

$$\begin{aligned} f(x\omega) &= a + b(x\omega)^n + (x\omega)^{n+1} Q(x\omega) \\ &= a + bx^n \omega^n + (x\omega)^{n+1} Q(x\omega) \\ &= a + bx^n \left(-\frac{a}{b}\right) + (x\omega)^{n+1} Q(x\omega) \\ &= a(1 - x^n) + (x\omega)^{n+1} Q(x\omega). \end{aligned}$$

*The formal proof of this result requires some more sophisticated results from analysis and topology, but the idea is similar to why every even-degree polynomial $p(x)$ in $\mathbb{R}[x]$ attains a minimum value. For any such $p(x)$, the Extreme Value Theorem guarantees that $p(x)$ will attain a minimum value m_a on each closed interval of the form $[-a, a]$. But $p(x) \to \infty$ as $|x| \to \infty$, and so there must be some $a \in \mathbb{R}^+$ for which $p(x) > m_a$ for all $x < -a$ and all $x > a$. It follows that this m_a is the (global) minimum value of $p(x)$.

Suppose $Q(z) = q_0 + q_1 z + q_2 z^2 + \cdots + q_m z^m$. If $0 < t < 1$, then the triangle inequality shows that

$$|Q(t\omega)| \leq \sum_{k=0}^{m} |q_k (t\omega)^k| = \sum_{k=0}^{m} |t|^k |q_k \omega^k| < \sum_{k=0}^{m} |q_k \omega^k|.$$

Note that the quantity $\sum_{k=0}^{m} |q_k \omega^k|$ does not depend on t. Therefore, there exists $t \in \mathbb{R}$ with $0 < t < 1$ such that

$$t|\omega^{n+1} Q(t\,\omega)| < |a|.$$

Thus,

$$\begin{aligned} |f(t\omega)| &\leq |a|(1 - t^n) + t^n \left(t|\omega^{n+1} Q(t\omega)|\right) \\ &< |a|(1 - t^n) + t^n |a| \\ &= |a| \\ &= |f(0)|, \end{aligned}$$

which contradicts the fact that $|f(z)|$ attains its minimum value at 0. Therefore, it must be that $f(0) = 0$, and $f(z)$ has a root. ■

Appendix G

Complex Roots of Unity

Focus Questions

By the end of this investigation, you should be able to give precise and thorough answers to the questions listed below. You may want to keep these questions in mind to focus your thoughts as you complete the investigation.

- How do we represent complex numbers using a polar or trigonometric form?
- What is a complex root of unity?
- What is de Moivre's theorem and what does it tell us about powers of complex numbers?
- How do we find all complex nth roots of unity?
- How can we find complex nth roots of an arbitrary complex number?

Preview Activity G.1. Consider the problem of determining all solutions to the polynomial equation $x^3 - 1 = 0$. We know that $x = \sqrt[3]{1} = 1$ is one solution, but the Fundamental Theorem of Algebra tells us that there are two more solutions. That is, there are three complex numbers that satisfy the equation $x^3 - 1 = 0$. These solutions are called the *third roots of unity* (where 1 is the unity).

(a) Let $z_0 = \cos\left(\frac{0\pi}{3}\right) + i\sin\left(\frac{0\pi}{3}\right)$. Evaluate the trigonometric functions at $\frac{0\pi}{3}$ and calculate *exact* numeric values for the real and imaginary parts of z_0. Then calculate z_0^2 and z_0^3.

(b) Let $z_1 = \cos\left(\frac{2\pi}{3}\right) + i\sin\left(\frac{2\pi}{3}\right)$. Evaluate the trigonometric functions at $\frac{2\pi}{3}$ and calculate *exact* numeric values for the real and imaginary parts of z_1. Then calculate z_1^2 and z_1^3.

(c) Let $z_2 = \cos\left(\frac{4\pi}{3}\right) + i\sin\left(\frac{4\pi}{3}\right)$. Evaluate the trigonometric functions at $\frac{4\pi}{3}$ and calculate *exact* numeric values for the real and imaginary parts of z_2. Then calculate z_2^2 and z_2^3.

(d) How are z_0, z_1, and z_2 related to the polynomial $x^3 - 1$?

In this investigation we will expand on Preview Activity G.1 to explicitly determine all complex roots of unity and see how these roots of unity help us solve polynomial equations.

The Trigonometric form of a Complex Number

Multiplication of complex numbers is a pretty straightforward algebraic process, but the algebra doesn't provide much visual insight into a complex product. One idea that helps us visualize complex multiplication and makes the determination of complex roots of unity possible is the trigonometric (or polar) form of a complex number. This trigonometric form connects the algebra of complex numbers to trigonometry in a very useful way.

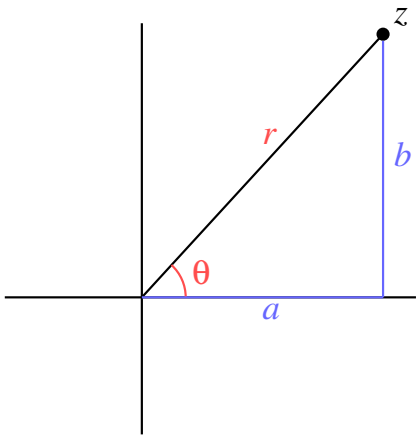

Figure G.1
Trigonometric form of a complex number.

If $z = a + bi$ is a complex number, then we can plot z in the plane as shown in Figure G.1. In this situation, we will let r be the magnitude of z (that is, the distance from z to the origin) and θ the angle z makes with the positive real axis as shown in Figure G.1.

We can use trigonometry to see that

$$a = r\cos(\theta) \text{ and } b = r\sin(\theta).$$

We can then write z in trigonometric form as

$$z = r(\cos(\theta) + i\sin(\theta)) \tag{G.1}$$

and call (G.1) the trigonometric (or polar) form (or representation) of the complex number z. (The word *polar* here comes from the fact that this process can be viewed as occurring with polar coordinates.) The angle θ is called the *argument* of the complex number z and the real number r is the *modulus* or *norm* of z. To find the polar representation of a complex number $z = a + bi$, we first notice that

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

To find θ, we have to consider cases.

- If $z = 0 = 0 + 0i$, then $r = 0$ and θ can have any real value.

- If $z \neq 0$, then $\tan(\theta) = \frac{b}{a}$ if $a \neq 0$ and θ is determined by the quadrant in which z lies.
- If $z \neq 0$ and $a = 0$ (so $b \neq 0$), then
  - $\theta = \frac{\pi}{2}$ if $b > 0$
  - $\theta = -\frac{\pi}{2}$ if $b < 0$.

Activity G.2.

(a) Find the polar form of the complex numbers $w = 4 + 4\sqrt{3}i$ and $z = 1 - i$.

(b) Find $a, b \in \mathbb{R}$ so that $a + bi = 3\left(\cos\left(\frac{\pi}{6}\right) + i\sin\left(\frac{\pi}{6}\right)\right)$.

Products of Complex Numbers in Polar Form

There is an important product formula for complex numbers that the polar form provides. We illustrate with an example.

Example G.3. Let $w = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$ and $z = \sqrt{3} + i$. Now $|w| = \sqrt{\left(-\frac{1}{2}\right)^2 + \left(\frac{\sqrt{3}}{2}\right)^2} = 1$ and the argument of w satisfies $\tan(\theta) = -\sqrt{3}$. Since w is in the second quadrant, we see that $\theta = \frac{2\pi}{3}$, so the polar form of w is

$$w = \cos\left(\frac{2\pi}{3}\right) + i\sin\left(\frac{2\pi}{3}\right).$$

Also, $|z| = \sqrt{\left(\sqrt{3}\right)^2 + 1^2} = 2$ and the argument of z satisfies $\tan(\theta) = \frac{1}{\sqrt{3}}$. Since z is in the first quadrant, we know that $\theta = \frac{\pi}{6}$ and the polar form of z is

$$z = 2\left[\cos\left(\frac{\pi}{6}\right) + i\sin\left(\frac{\pi}{6}\right)\right].$$

Computing wz directly gives

$$\begin{aligned} wz &= \left(\sqrt{3} + i\right)\left(-\frac{1}{2} + \frac{\sqrt{3}}{2}i\right) \\ &= -\sqrt{3} + i. \end{aligned}$$

Now $|wz| = 2$ and the argument of zw satisfies $\tan(\theta) = -\frac{1}{\sqrt{3}}$. Since wz is in quadrant II, we see that $\theta = \frac{5\pi}{6}$ and the polar form of wz is

$$wz = 2\left[\cos\left(\frac{5\pi}{6}\right) + i\sin\left(\frac{5\pi}{6}\right)\right].$$

Notice that $|wz| = |w|\,|z|$ and that the argument of zw is $\frac{2\pi}{3} + \frac{\pi}{6}$ or the sum of the arguments of w and z.

In general, we have the following important result about the product of two complex numbers.

Theorem G.4. *Let $w = r(\cos(\alpha) + i\sin(\alpha))$ and $z = s(\cos(\beta) + i\sin(\beta))$ be complex numbers in polar form. Then the polar form of the complex product wz is given by*

$$wz = rs\left(\cos(\alpha+\beta) + i\sin(\alpha+\beta)\right).$$

The verification is left for Exercise (2).

An illustration of this is given in Figure G.2. Theorem G.4 tells us that to multiply two complex numbers together, we add their arguments and multiply their magnitudes.

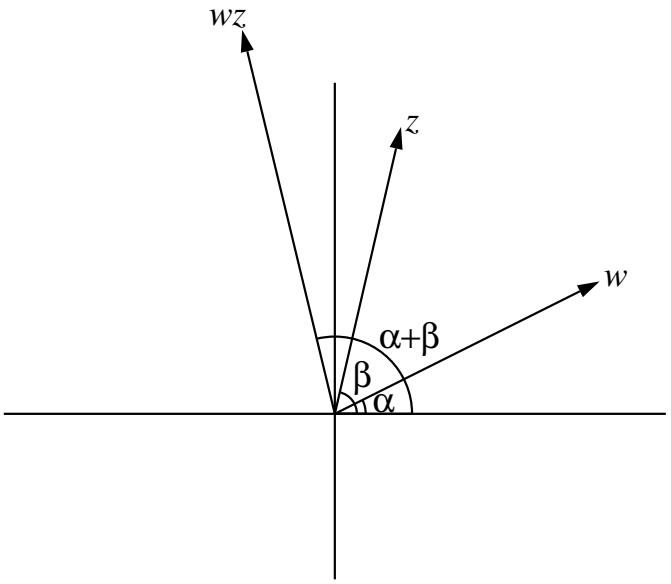

Figure G.2
A geometric interpretation of complex multiplication.

Activity G.5. Let $w = 3\left[\cos\left(\frac{5\pi}{3}\right) + \sin\left(\frac{5\pi}{3}\right)\right]$ and $z = 2\left[\cos\left(-\frac{\pi}{4}\right) + \sin\left(-\frac{\pi}{4}\right)\right]$.

(a) What is $|wz|$?

(b) What is the argument of wz?

(c) In which quadrant is wz? Explain.

(d) Find the polar form of wz.

An alternate representation for the polar form of a complex number is the *exponential form* of a complex number. If $w = r(\cos(\alpha) + i\sin(\alpha))$ and $z = s(\cos(\beta) + i\sin(\beta))$, then the polar product

$$wz = rs\left(\cos(\alpha+\beta) + i\sin(\alpha+\beta)\right)$$

allows us to multiply the magnitudes and add the arguments. Exponential functions have the same properties. That is, if $u = ae^x$ and $v = be^y$, then

$$uv = abe^{x+y}.$$

This similarity provides a shorthand way of representing the polar form of a complex number. We write $e^{i\theta}$ to mean

$$e^{i\theta} = \cos(\theta) + i\sin(\theta)$$

(this equation is called *Euler's formula*). Then the polar form $r(\cos(\theta) + i\sin(\theta))$ of a complex number can be written as $re^{i\theta}$:

$$re^{i\theta} = r(\cos(\theta) + i\sin(\theta)).$$

An interesting consequence is the elegant and unexpected relationship between four of the most famous mathematical constants (0, 1, e, and π). If we substitute π for θ in $e^{i\theta}$ we see that

$$e^{i\pi} = -1 \quad \text{or} \quad e^{i\pi} + 1 = 0.$$

This last equation is called *Euler's identity*.

Theorem G.4 gives us a quick way to compute powers of a complex number, and hence to find roots of complex numbers. If $z = r(\cos(\theta) + i\sin(\theta))$, then Theorem G.4 shows us that

$$\begin{aligned} z^2 &= (r)(r)\left(\cos(\theta+\theta) + i\sin(\theta+\theta)\right) \\ &= r^2\left(\cos(2\theta) + i\sin(2\theta)\right). \end{aligned}$$

We can continue this process to obtain the following theorem.

Theorem G.6 (de Moivre's Theorem)**.** *Let $z = r(\cos(\theta) + i\sin(\theta))$ be a complex number and n any integer. Then*

$$z^n = r^n(\cos(n\theta) + i\sin(n\theta)). \tag{G.2}$$

The proof is left for Exercise (5). In exponential form we have

$$\left(re^{i\theta}\right)^n = r^n e^{in\theta}.$$

Note that de Moivre's Theorem works for negative integer powers as well as positive integer powers.

Activity G.7. Use de Moivre's Theorem to find z^{10} where $z = 1 - i$.

Roots of Unity

Now we return to the problem of solving the polynomial equation $x^3 = 1$. If we draw the graph of $y = x^3 - 1$ we see that the graph intersects the x-axis at only one point, so there is only one real solution to $x^3 = 1$. That means the other two solutions must be complex. The key to finding the two complex third roots of unity is de Moivre's Theorem. Suppose

$$z = r\left[\cos(\theta) + i\sin(\theta)\right]$$

is a solution to $x^3 = 1$. Then

$$1 = z^3 = r^3(\cos(3\theta) + i\sin(3\theta)).$$

This implies that $r = 1$ (or $r = -1$, but we can incorporate the latter case into our choice of angle). We then have

$$1 = \cos(3\theta) + i\sin(3\theta). \tag{G.3}$$

Equation (G.3) has solutions when $\cos(3\theta) = 1$ and $\sin(3\theta) = 0$. This will occur when $3\theta = 2\pi k$, or $\theta = \frac{2\pi k}{3}$, where k is any integer. The distinct integer multiples of $\frac{2\pi k}{3}$ on the unit circle occur when $k = 0$ and $\theta = 0$, $k = 1$ and $\theta = \frac{2\pi}{3}$, and $k = 2$ with $\theta = \frac{4\pi}{3}$. In other words, the solutions to $x^3 = 1$ should be

$$\begin{aligned} z_0 &= \cos(0) + i\sin(0) = 1 \\ z_1 &= \cos\left(\frac{2\pi}{3}\right) + i\sin\left(\frac{2\pi}{3}\right) = -\frac{1}{2} + \frac{\sqrt{3}}{2}i \\ z_2 &= \cos\left(\frac{4\pi}{3}\right) + i\sin\left(\frac{4\pi}{3}\right) = -\frac{1}{2} - \frac{\sqrt{3}}{2}i. \end{aligned}$$

We already know that $z_0^3 = 1^3 = 1$, so z_0 actually is a solution to $x^3 = 1$. To check that z_1 and z_2 are also solutions to $x^3 = 1$, we apply de Moivre's Theorem:

$$\begin{aligned} z_1^3 &= \left[\cos\left(\frac{2\pi}{3}\right) + i\sin\left(\frac{2\pi}{3}\right)\right]^3 \\ &= \cos\left(3\frac{2\pi}{3}\right) + i\sin\left(3\frac{2\pi}{3}\right) \\ &= \cos(2\pi) + i\sin(2\pi) \\ &= 1, \end{aligned}$$

and

$$\begin{aligned} z_2^3 &= \left[\cos\left(\frac{4\pi}{3}\right) + i\sin\left(\frac{4\pi}{3}\right)\right]^3 \\ &= \cos\left(3\frac{4\pi}{3}\right) + i\sin\left(3\frac{4\pi}{3}\right) \\ &= \cos(4\pi) + i\sin(4\pi) \\ &= 1. \end{aligned}$$

Thus, $z_1^3 = 1$ and $z_2^3 = 1$ and we have found three solutions to the equation $x^3 = 1$. Since a cubic can have only three solutions, we have found them all.

In general, the solutions to $x^n = 1$ for a positive integer n are called the nth roots of unity.

Definition G.8. If n is a positive integer, the nth **roots of unity** are the n solutions to the equation $z^n = 1$.

The argument above can be extended to show that the nth roots of unity are given by

$$e^{\frac{2k\pi i}{n}} = \cos\left(\frac{2k\pi}{n}\right) + i\sin\left(\frac{2k\pi}{n}\right)$$

for k from 0 to $n - 1$. To verify this statement, note that

$$\left(e^{\frac{2k\pi i}{n}}\right)^n = e^{2k\pi i} = \cos(2k\pi) + i\sin(2k\pi) = 1.$$

The complex numbers $e^{\frac{2k\pi i}{n}}$ are distinct for k from 0 to $n - 1$, so we have the n different complex roots of 1.

Activity G.9.

(a) Find all solutions to $x^4 = 1$. That is, find the fourth roots of unity.

(b) Find all sixth roots of unity.

Activity G.9 illustrates something interesting. There are four complex roots of 1, given by 1, ω, ω^2, and ω^3, where

$$\omega = e^{\frac{2\pi i}{4}} = \cos\left(\frac{\pi}{2}\right) + i\sin\left(\frac{\pi}{2}\right) = i.$$

Now $\omega^2 = -1$, and $1^2 = 1$ and $(-1)^2 = 1$, while ω^2 and $\left(\omega^3\right)^2$ are not equal to 1. So 1 and -1 are also square roots of 1. We say that an nth root of unity z is a *primitive* nth root of unity if z is not a root of unity for some smaller power.

Definition G.10. Let n be a positive integer. The complex number z is a **primitive** nth root of unity if $z^n = 1$ but $z^k \neq 1$ for any k with $1 \leq k < n$.

The complex number $e^{\frac{2\pi i}{n}}$ is always a primitive nth root of unity, but this is not the only primitive nth root of unity. We determine all of the primitive nth roots of unity in the concluding activities.

Roots of Complex Numbers

The final topic we will consider in this investigation is extending this process for finding roots of unity to calculate roots of other complex numbers. To find the solutions to the equation $x^n = a + bi$, where n is a positive integer and $a + bi$ is a complex number with trigonometric form

$$a + bi = r\left[\cos(\theta) + i\sin(\theta)\right],$$

with $r > 0$, we assume that $z = s\left[\cos(\alpha) + i\sin(\alpha)\right]$ with $s > 0$ is a solution to $x^n = a + bi$. Then

$$\begin{aligned} a + bi &= z^n \\ &= \left(s\left[\cos(\alpha) + i\sin(\alpha)\right]\right)^n \\ &= s^n\left[\cos(n\alpha) + i\sin(n\alpha)\right] \end{aligned}$$

and so

$$s^n = r \quad \text{and} \quad \cos(\theta) + i\sin(\theta) = \cos(n\alpha) + i\sin(n\alpha).$$

Therefore,

$$s^n = r \quad \text{and} \quad n\alpha = \theta + 2\pi k$$

where k is any integer. This give us

$$s = \sqrt[n]{r} \quad \text{and} \quad \alpha = \frac{\theta + 2\pi k}{n}.$$

We will get n different solutions for $k = 0, 1, 2, \ldots, n-1$, and these will be all of the solutions. These solutions are called the nth roots of the complex number $a + bi$. We summarize the results in the next theorem.

Theorem G.11. *Let $n \in \mathbb{Z}^+$. The nth roots of the complex number $r\left[\cos(\theta) + i\sin(\theta)\right]$ are given by*

$$\sqrt[n]{r}\left[\cos\left(\frac{\theta + 2\pi k}{n}\right) + i\sin\left(\frac{\theta + 2\pi k}{n}\right)\right]$$

for $k = 0, 1, 2, \ldots, (n-1)$.

In other words, if r is an nth root of a complex number z, then all of the nth roots of z have the form $r\omega^k$ for $0 \leq k \leq n-1$, where ω is a primitive nth root of unity.

Example G.12. We solve the equation

$$x^4 = -8 + 8\sqrt{3}i.$$

Note that we can write the right hand side of this equation in trigonometric form as

$$-8 + 8\sqrt{3}i = 16\left(\cos\left(\frac{2\pi}{3}\right) + i\sin\left(\frac{2\pi}{3}\right)\right).$$

The fourth roots of $-8 + 8\sqrt{3}i$ are then

$$\begin{aligned} x_0 &= \sqrt[4]{16}\left[\cos\left(\frac{\frac{2\pi}{3} + 2\pi(0)}{4}\right) + i\sin\left(\frac{\frac{2\pi}{3} + 2\pi(0)}{4}\right)\right] \\ &= 2\left[\cos\left(\frac{\pi}{6}\right) + i\sin\left(\frac{\pi}{6}\right)\right] \\ &= 2\left(\frac{\sqrt{3}}{2} + i\frac{1}{2}\right) \\ &= \sqrt{3} + i, \end{aligned}$$

$$\begin{aligned} x_1 &= \sqrt[4]{16}\left[\cos\left(\frac{\frac{2\pi}{3} + 2\pi(1)}{4}\right) + i\sin\left(\frac{\frac{2\pi}{3} + 2\pi(1)}{4}\right)\right] \\ &= 2\left[\cos\left(\frac{2\pi}{3}\right) + i\sin\left(\frac{2\pi}{3}\right)\right] \\ &= 2\left(-\frac{1}{2} + i\frac{\sqrt{3}}{2}\right) \\ &= -1 + \sqrt{3}i, \end{aligned}$$

$$\begin{aligned} x_2 &= \sqrt[4]{16}\left[\cos\left(\frac{\frac{2\pi}{3} + 2\pi(2)}{4}\right) + i\sin\left(\frac{\frac{2\pi}{3} + 2\pi(2)}{4}\right)\right] \\ &= 2\left[\cos\left(\frac{7\pi}{6}\right) + i\sin\left(\frac{7\pi}{6}\right)\right] \\ &= 2\left(-\frac{\sqrt{3}}{2} - i\frac{1}{2}\right) \\ &= -\sqrt{3} - i, \end{aligned}$$

and

$$\begin{aligned} x_3 &= \sqrt[4]{16}\left[\cos\left(\frac{\frac{2\pi}{3}+2\pi(3)}{4}\right)+i\sin\left(\frac{\frac{2\pi}{3}+2\pi(3)}{4}\right)\right] \\ &= 2\left[\cos\left(\frac{5\pi}{3}\right)+i\sin\left(\frac{5\pi}{3}\right)\right] \\ &= 2\left(\frac{1}{2}-i\frac{\sqrt{3}}{2}\right) \\ &= 1-\sqrt{3}i. \end{aligned}$$

Activity G.13. Find all fourth roots of -256, that is find all solutions to $x^4 = -256$.

Concluding Activities

Activity G.14. Let ω be a primitive sixth root of unity.

(a) Determine the powers of ω^2. What does this say about ω^2?

(b) Determine the powers of ω^3. What does this say about ω^3?

(c) Parts (a) and (b) illustrate the following theorem.

Theorem G.15. *Let F be a field that contains a primitive nth root of unity. Then F contains a primitive kth root of unity for every positive divisor k of n.*

Complete the following steps to prove this theorem.

i. Let ω be a primitive nth root of unity in F, and let k be a positive divisor of n. Let $d = \frac{n}{k}$. Let $\rho = \omega^d$. Show that $\rho^k = 1$.

ii. We have shown that ρ is a kth root of unity. To prove that ρ is a primitive kth root of unity, we must demonstrate that $\rho^m \neq 1$ for $1 \leq m < k$. Let m be an integer with $1 \leq m < k$. Show that $\rho^m \neq 1$.

Activity G.16. Let n be a positive integer. The complex number $\omega = e^{\frac{2\pi i}{n}}$ is always a primitive nth root of unity, but this is not the only primitive nth root of unity. We know that the nth roots of unity are ω^k for $0 \leq k \leq n-1$. In this exercise we determine exactly which of these roots are primitive nth roots of unity.

(a) Consider the case with $n = 6$. Determine which powers ω^k are primitive sixth roots of unity. What is the relationship between k and 6 if ω^k is a primitive sixth root of unity?

(b) Repeat part (a) with $n = 8$.

(c) Now generalize parts (a) and (b) and determine in general which powers of ω are primitive nth roots of unity.

,

Exercises

(1) Let $w = 2 - 5i$ and $z = -4 + 7i$.

(a) Compute $w + z$

(b) Compute wz

(c) Find the polar forms of w and z.

(d) Use the polar forms to compute wz. Compare to your earlier result.

(2) Prove Theorem G.4 that if $w = r(\cos(\alpha) + i\sin(\alpha))$ and $z = s(\cos(\beta) + i\sin(\beta))$, then

$$wz = rs\left(\cos(\alpha + \beta) + i\sin(\alpha + \beta)\right).$$

(3) Find the cube roots of $a = -2 + 2i$.

(4) In this problem we will find the fifth roots of unity, or the solutions to the equation $z^5 - 1 = 0$. This is much more complicated than finding the fourth or sixth roots of unity due to the difficulty of finding $\cos\left(\frac{\pi}{5}\right)$ and $\sin\left(\frac{\pi}{5}\right)$. Assume the identity

$$\sin(5\alpha) = 5\sin(\alpha) - 20\sin^3(\alpha) + 16\sin^5(\alpha). \tag{G.4}$$

WARNING: This is a very messy problem!

(a) Use (G.4) to find an equation for $\sin\left(\frac{\pi}{5}\right)$.

(b) Let $w = \sin\left(\frac{\pi}{5}\right)$. Show that the value of $\sin\left(\frac{\pi}{5}\right)$ is a solution to

$$16w^4 - 20w^2 + 5 = 0.$$

(c) Show that $\sin\left(\frac{\pi}{5}\right) = \frac{\sqrt{10-2\sqrt{5}}}{4}$.

(d) Find the exact values of the fifth roots of unity. Hint: Let $\varphi = \frac{1+\sqrt{5}}{2}$. Show that $\sin\left(\frac{2\pi}{5}\right) = \frac{1}{2}\sqrt{2+\varphi}$ and $\cos\left(\frac{2\pi}{5}\right) = \left(\frac{1}{2}\right)(\varphi - 1)$. Note that $\varphi^2 = \varphi + 1$. Express all fifth roots of unity in terms of φ. Note also that computing inverses in $\mathbb{C}$ can be easier than computing products.

(e) Verify (G.4) for any angle α. You may use the following well-known trigonometric identities that are valid for any angles α and β:

$$\sin(\alpha + \beta) = \cos(\alpha)\sin(\beta) + \cos(\beta)\sin(\alpha) \tag{G.5}$$

$$\sin(2\alpha) = 2\cos(\alpha)\sin(\alpha) \tag{G.6}$$

$$\cos(2\alpha) = 1 - 2\sin^2(\alpha) \tag{G.7}$$

$$\cos^2(\alpha) + \sin^2(\alpha) = 1. \tag{G.8}$$

(5) Prove de Moivre's Theorem that if $z = r(\cos(\theta) + i\sin(\theta))$ is a complex number and n any integer, then

$$z^n = r^n(\cos(n\theta) + i\sin(n\theta)). \tag{G.9}$$

Index