An Introduction to Abstract Algebra via Applications

David R. Finston and Patrick J. Morandi
Department of Mathematical Sciences
New Mexico State University
Las Cruces NM 88003-8001

September 25, 2007
## Contents

1 Identification Numbers and Modular Arithmetic  
   1.1 Examples of Identification Numbers ................................. 1 
   1.2 Modular Arithmetic ................................................. 5 
   1.3 Error Detection with Identification Numbers ....................... 19 

2 Error Correcting Codes  
   2.1 Basic Notions ...................................................... 23 
   2.2 Gaussian Elimination ............................................... 28 
   2.3 The Hamming Code ................................................. 35 
   2.4 Coset Decoding ...................................................... 36 
   2.5 The Golay Code ..................................................... 40 

3 Rings and Fields  
   3.1 The Definition of a Ring .......................................... 43 
   3.2 First Properties of Rings .......................................... 47 
   3.3 Fields .............................................................. 55 

4 Linear Algebra and Linear Codes  
   4.1 Vector Spaces ...................................................... 59 
   4.2 Linear Independence, Spanning, and Bases ......................... 64 
   4.3 Linear Codes ....................................................... 69 

5 Field Extensions and Ruler and Compass Constructions  
   5.1 Field Extensions .................................................... 77 
   5.2 Ruler and Compass Constructions .................................. 79 
   5.3 Constructions and Field Extensions ................................ 80 
   5.4 Classical Construction Problems .................................. 84 
      5.4.1 Angle Trisection ................................................ 84 
      5.4.2 Duplicating a Cube ............................................. 85 
      5.4.3 Squaring the Circle ........................................... 86 
      5.4.4 Constructible Polygons ....................................... 86
6 Quotient Rings and Field Extensions 89
6.1 Arithmetic of Polynomial Rings .................................. 89
6.2 Ideals and Quotient Rings ........................................ 93

7 Cyclic Codes 103
7.1 Introduction to Cyclic Codes ..................................... 103
7.2 Finite Fields ....................................................... 106
7.3 Reed-Solomon Codes ............................................. 112
7.4 Error Correction for Reed-Solomon Codes ..................... 116

8 Cryptography and Group Theory 121
8.1 The RSA encryption system ................................... 121
8.2 Groups ............................................................. 123
8.3 Secure Signatures with RSA ................................... 130

9 Symmetry 133
9.1 Isometries ............................................................. 133
9.2 Symmetry Groups .................................................. 134
9.3 Examples of Symmetry Groups ................................. 137
9.4 Symmetry Groups of Bounded Figures ......................... 140

Bibliography 143

List of Symbols 145

Index 145
Chapter 1

Identification Numbers and Modular Arithmetic

The first topic we will investigate in this course is the mathematics of identification numbers. Many things are described by a code of digits; zip codes, items in a grocery store, and books, to name three. One feature to all of these codes is the inclusion of an extra numerical digit, called a check digit, designed to detect errors in reading the code. When a machine (or a human) reads information, there is always the possibility of the information being read incorrectly. For example, moisture or dirt on the scanner used by a grocery store clerk can prevent an item’s code from being read correctly. It would be unacceptable if, because of a scanning error, a customer is charged for caviar when they are buying tuna fish. The use of the check digit allows for the detection of some scanning errors. If an error is detected, the item is then rescanned until the correct code is read.

1.1 Examples of Identification Numbers

There are many different methods being used to produce identification numbers. We will discuss three of them; the United States Postal Service zip code, the Universal Product Code used for consumer products, and the International Standard Book Number.

The USPS Zip Code

The United States Postal Service uses a bar code to read zip codes on mail. The following bar code is that for the Mathematical Sciences Department of NMSU, whose zip code is 88003-8001.
The bar code represents a ten digit number. There are fifty two lines in the bar code. The first and last lines are just markers. The remaining fifty lines comprise ten groups of five, and each group of five represents a digit. The first nine digits form the nine digit zip code of the addressee. The tenth digit is a check digit. This digit is computed as follows: the digits forming the zip code are added, and the check digit is the smallest nonnegative integer needed to make the sum be divisible by 10. For example, given the zip code 88003-8001 for the Department of Mathematical Sciences at New Mexico State University, the nine digits sum to 28. Therefore, the check digit must be 2. Thus, the bar code this zip code represents the ten digit number 8800380012.

This scheme allows one to determine the check digit for any nine digit zip code. For example, if we only knew the nine digit zip code 88003-8001, the check digit $x$ would be the number between 0 and 9 such that the sum

$$8 + 8 + 0 + 0 + 3 + 8 + 0 + 0 + 1 + x$$

was evenly divisible by 10. Since this sum is $28 + x$, the only choice for $x$ is to be 2.

The purpose of the check digit is to detect errors in reading the code. For example, suppose that the zip code 8800380012 was incorrectly read as 8800880012 by reading the fifth digit as an 8 instead of as a 3. The sum of the digits would then be $8 + 8 + 8 + 8 + 1 + 2 = 35$, which is not divisible by 10. Therefore, the postal service’s scanners would detect an error, and the zip code would have to be read again.

**The Universal Product Code (UPC)**

The Universal Product Code, or UPC, appears on grocery items.

This is a twelve digit code consisting of two blocks of five digits preceded and followed with a single digit, as the example above indicates. The first six identify the country and the
1.1. EXAMPLES OF IDENTIFICATION NUMBERS

manufacturer of the product and the next five identify the product itself. The final digit is the check digit. A twelve digit code \((a_1, \ldots, a_{12})\) is valid provided that

\[
3a_1 + a_2 + 3a_3 + a_4 + \cdots + 3a_{11} + a_{12}
\]

is evenly divisible by 10. The UPC of the example above is 0 41390 30860 4. Therefore, the sum for this code is

\[
3 \cdot 0 + 1 \cdot 4 + 3 \cdot 1 + 1 \cdot 3 + 3 \cdot 9 + 1 \cdot 0 + 3 \cdot 3 + 1 \cdot 0 + 3 \cdot 8 + 1 \cdot 6 + 3 \cdot 0 + 1 \cdot 4 = 80.
\]

This sum is indeed evenly divisible by 10, so the number is valid.

As with the zip code, given the first eleven digits, there is enough information to uniquely determine the check digit. For example, given the partial UPC of 0 71142 00001, if the check digit is \(x\), then

\[
3 \cdot 0 + 7 + 3 \cdot 1 + 1 + 3 \cdot 4 + 2 + 3 \cdot 0 + 0 + 3 \cdot 0 + 0 + 3 \cdot 1 + x = 28 + x,
\]

which forces \(x = 2\).

As with the zip code scheme, the UPC adds the check digit to help detect errors. For example, if the code 0 71142 00001 2 was incorrectly read as 0 71342 00001 2 by reading the fourth digit as a 3 instead of as a 1, then the computation to check if this number is valid would give

\[
3 \cdot 0 + 7 + 3 \cdot 1 + 1 + 3 \cdot 4 + 2 + 3 \cdot 0 + 0 + 3 \cdot 0 + +0 + 3 \cdot 1 + 2 = 32,
\]

which is not divisible by 10. Therefore, a grocery store scanner would not recognize the code as valid, and the cashier would have to rescan the item.

The International Standard Book Number (ISBN)

Books are identified by a ten digit number, abbreviated by ISBN. For example, the book *Field and Galois Theory*, published by Springer, has ISBN 0-387-94753-1. The first digit identifies the language in which the book is written, the second block of digits identifies the publisher, the third block identifies the book itself, and the final digit is the check digit. In this scheme, each digit can be a numeral 0, \ldots, 9 or X, which represents 10. A ten digit number \((a_1, \ldots, a_{10})\) is a valid ISBN provided that

\[
10a_1 + 9a_2 + 8a_3 + \cdots + 2a_9 + a_{10}
\]

is evenly divisible by 11. In the number above, we have

\[
10 \cdot 0 + 9 \cdot 3 + 8 \cdot 8 + 7 \cdot 7 + 6 \cdot 9 + 5 \cdot 4 + 4 \cdot 7 + 3 \cdot 5 + 2 \cdot 3 + 1
\]

\[
= 11 \cdot 24,
\]
so the number is indeed valid. The digit X is only used, when appropriate, for the check digit.

As with the previous two examples, the check digit can be determined uniquely, given that it is between 0 and 10. For example, for the book *A Classical Introduction to Modern Number Theory*, published by Springer, whose number will start with 0-387-97329, the check digit \(x\) must result in

\[
10 \cdot 0 + 9 \cdot 3 + 8 \cdot 8 + 7 \cdot 7 + 6 \cdot 9 + 5 \cdot 7 + 4 \cdot 3 + 3 \cdot 2 + 2 \cdot 9 + x
= 265 + x \text{ divisible by 11.}
\]

Since \(11 \cdot 24 = 264\), if \(265 + x\) is to be divisible by 11 and \(0 \leq x \leq 10\), then \(x = 10\). Thus, the check digit for this book is X, and so the ISBN is 0-387-97329-X.

The ISBN scheme also allows for detection of some errors. When we discuss error detection in more detail below, we will see that all of these schemes will detect an error in a single digit. Unfortunately, errors in more than one digit are not always detected. However, the ISBN scheme does better, in some sense, than the other two schemes above because it detects transposition errors. For example, given the ISBN 0-387-97329-X, if the fifth and sixth digits are transposed, the resulting number is 0-387-79329-X. The check for validity of this number would result in the sum

\[
10 \cdot 0 + 9 \cdot 3 + 8 \cdot 8 + 7 \cdot 7 + 6 \cdot 9 + 5 \cdot 7 + 4 \cdot 3 + 3 \cdot 2 + 2 \cdot 9 + 10 = 273,
\]

which is not divisible by 11. Thus, this number is invalid. However, transposing digits in a valid zip code will always result in a number considered valid, since the sum of the digits is unchanged by this transposition. We will discuss transposition errors in more detail later.

**Exercises**

1. Check if the following numbers are valid ISBNs:
   
   (a) 0-8218-2169-5
   (b) 0-201-01361-9
   (c) 2-87647-089-6
   (d) 3-7643-3065-1

2. Suppose a UPC is read, but the third digit is left out, and the result is 0 7\(x\)172 38175 1, where \(x\) represents the missing digit. Calculate, in terms of \(x\), the sum needed to check if this is a valid number. Then write down the condition on \(x\) required for the number to be valid, and determine \(x\).

3. The number 0-8176-3165-1 is an invalid ISBN (check this!). It was created by taking the ISBN number of a book and changing one digit. Can you tell which digit was changed? Explain why not by giving two examples of a valid ISBN that differs from this one in exactly one digit.
4. Consider the following identification number scheme: If \( a = (a_1, a_2, a_3, a_4) \), where each \( a_i \) is between 0 and 4, then the number \( a \) is valid provided that \( 4a_1 + 3a_2 + a_3 + 2a_4 \) is divisible by 5. If \( (3, 2, 4, x) \) is a valid number, determine \( x \).

5. Consider the following identification number scheme: a valid number is a 5-tuple of integers \( a = (a_1, a_2, a_3, a_4, a_5) \) with \( 0 \leq a_i \leq 12 \) such that \( 2a_1 + 3a_2 + 5a_3 + a_4 + 6a_5 \) is divisible by 13. If \( (2, 3, 4, 11, x) \) is a valid number, determine \( x \).

6. Consider the scheme of the previous problem. If \( (a_1, a_2, a_3, a_4, a_5) \) is a valid number and if \( a_1 \neq a_2 \), prove that \( (a_2, a_1, a_3, a_4, a_5) \) is not valid.

7. Let \( n \) be a positive integer and let \( m \) be a positive divisor of \( n \). If \( a \) and \( b \) are integers with \( a \equiv b \mod n \), prove that \( a \equiv b \mod m \).

1.2 Modular Arithmetic

In order to investigate the error detection capabilities of the various identification number schemes we have discussed, and to work with the other applications in this course, we will look carefully at the computations involved in these schemes. In all three, a number is valid if some combination of its entries is divisible by some specific positive integer (10 or 11 in the examples). This actual result of the computation is not important in its own right. Rather what is important is only whether the result is divisible by the given integer. Phrased another way, what is important is not the combination but rather the remainder we would get if we divide our specific integer into the result. In some sense we are doing arithmetic with these remainders when we do calculations in these schemes.

Consider the following well known scenario. When we tell time in the U.S., the hour value is any whole number between 1 and 12. Three hours after 10 o’clock will be 1 o’clock. In general, to see what time it will be \( n \) hours after 10 o’clock, you add \( n \) to 10, and then remove enough multiples of 12 until you have a value between 1 and 12. For instance, in 37 hours past 10 o’clock, the time will be 11 o’clock since \( 47 = 36 + 11 \). In telling time, we then identify 13 o’clock with 1 o’clock, 14 o’clock with 2 o’clock, and so on. In this clock arithmetic, if we add 12 hours to any time, we get the same time (but changing AM to PM and vice-versa). Therefore, 12 acts in clock arithmetic like 0 acts in ordinary arithmetic.

There is nothing special about 12 with respect to obtaining a new type of arithmetic. As we will see in more detail below, in doing calculations in the various identification number schemes we talked about above, we are essentially doing clock arithmetic, but with 12 replaced by 10 for the zip code and UPC, and by 11 for the ISBN scheme. When we discuss coding theory, we will use clock arithmetic with 12 replaced by 2, and when we discuss cryptography, we will replace 12 by very large integers. We therefore need to discuss the general notion of clock arithmetic.

We begin with a very familiar concept.
Definition 1.1. Let \( a \) and \( n \) be integers. We say that \( n \) divides \( a \) (or \( a \) is divisible by \( n \)) if \( a = nb \) for some integer \( b \).

Definition 1.2. Let \( n \) be a positive integer. We say that two integers \( a \) and \( b \) are congruent modulo \( n \) if \( b - a \) is divisible by \( n \). When this occurs, we write \( a \equiv b \mod n \).

Since \( b - a \) is divisible by \( n \) exactly when \( b - a = qn \) for some integer \( q \), we see that \( a \equiv b \mod n \) if \( b = a + qn \) for some \( q \). This is a convenient way to express congruence modulo \( n \) in terms of an equation. If \( n = 12 \), then to say \( a \equiv b \mod 12 \) is equivalent to saying \( a \) o’clock is the same time as \( b \) o’clock, if we ignore AM and PM. Congruence modulo \( n \) is a relation on the set of integers. The first thing we point out is that this relation is an equivalence relation.

Proposition 1.3. The relation congruence modulo \( n \) is an equivalence relation for any positive integer \( n \).

Proof. Let \( n \) be a positive integer. We must prove that congruence modulo \( n \) is reflexive, symmetric, and transitive. For reflexivity, let \( a \) be any integer. Then \( a \equiv a \mod n \) since \( a - a = 0 \) is divisible by \( n \); for \( 0 = n \cdot 0 \). Next, for symmetry, suppose that \( a \) and \( b \) are integers with \( a \equiv b \mod n \). Then \( b - a \) is divisible by \( n \); say \( b - a = qn \) for some integer \( q \). Therefore, \( a \equiv b \mod n \), and so this relation is symmetric.

Finally, to prove transitivity, suppose that \( a, b, c \) are integers with \( a \equiv b \mod n \) and \( b \equiv c \mod n \). Then \( b - a \) and \( c - b \) are both divisible by \( n \). Then \( b - a = sn \) and \( c - b = tn \) for some integers \( s, t \). Adding these equations gives \( c - a = (s + t)n \), so \( c - a \) is divisible by \( n \), and so \( a \equiv c \mod n \). This proves transitivity. Since we have shown that congruence modulo \( n \) is reflexive, symmetric, and transitive, it is an equivalence relation.

Understanding the equivalence classes of this relation is of crucial importance. Recall that if \( \sim \) is an equivalence relation on a set \( X \), then the equivalence class of an element \( a \in X \) is the set . For ease of notation, in dealing with congruence modulo \( n \), we shall write the equivalence class of an integer \( a \) by \( \bar{a} \). Therefore, \( \{b \in X : b \sim a\} \)

\[
\bar{a} = \{b \in \mathbb{Z} : b \equiv a \mod n\}.
\]

Suppose \( n = 12 \). The equivalence class of 1 consists of all integers that are congruent to 1 modulo 12. That is, the equivalence class contains all integers \( c \) with \( c \) o’clock equal to 1 o’clock. We have \( \bar{1} = \{\ldots, -23, -11, 1, 13, 25, \ldots\} \). Similarly, \( \bar{2} = \{\ldots, -22, -10, 2, 14, 26, \ldots\} \). Note that \( \bar{12} = \{\ldots, -12, 0, 12, \ldots\} \) contains 0. An equivalence class can be represented in different ways. We have \( \bar{12} = \bar{0} = -\bar{24} \), and, more generally, \( \bar{12} = \bar{n}n \) for any integer \( n \). In other words, \( \bar{12} \) is the equivalence class of any element of the set \( \bar{12} = \{\ldots, -12, 0, 12, \ldots\} \).

If \( n \) is any positive integer, we denote by \( \mathbb{Z}_n \) the set of equivalence classes of integers for the equivalence relation of congruence modulo \( n \). For \( n = 2 \) we have

\[
\mathbb{Z}_2 = \{\bar{0}, \bar{1}\}.
\]
where \( \mathbb{T} \) represents the set of odd integers and \( \mathbb{U} \) represents the set of even integers. Anybody who can tell time will see that

\[
\mathbb{Z}_{12} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}.
\]

Notice that while \( \mathbb{Z} \) is an infinite set, \( \mathbb{Z}_{12} \) is finite; even though each equivalence class is infinite, there are only finitely many of them! This is not special to 12; we will prove that \( \mathbb{Z}_n \) has \( n \) elements for any \( n \). We first need an important result that will prove useful in many places in this course. This result, known as the division algorithm, can be viewed as a formal statement of the process writing a fraction as a proper fraction, i.e. the long division you learned in elementary school.

**Theorem 1.4 (Division Algorithm).** Let \( a \) and \( n \) be integers with \( n \) positive. Then there are unique integers \( q \) and \( r \) with \( a = qn + r \) and \( 0 \leq r < n \).

**Proof.** We use the well ordering property of the integers that says any nonempty subset of the nonnegative integers has a smallest element. To use this property, let us define

\[
S = \{ s \in \mathbb{Z} : s = a - qn \text{ for some } q \in \mathbb{Z} \text{ and } s \geq 0 \}.
\]

Note that \( S \) consists of all nonnegative remainders arising from divisions of \( a \) by \( n \). First, we show that \( S \) is indeed nonempty. If \( a \geq 0 \), then \( a - 0 \cdot n = a \in S \). If \( a < 0 \), then \( a - an = a(1 - n) \geq 0 \) since \( n \) is positive. Therefore, \( a - an \in S \). In either case, we see that \( S \) is nonempty. Therefore, by the well ordering property, \( S \) contains a smallest element, which we call \( r \). By definition of \( S \), there is a \( q \in \mathbb{Z} \) with \( r = a - qn \). Then \( a = qn + r \), proving one part of the theorem. To show that \( 0 \leq r < n \), we note that \( r \geq 0 \) since \( r \in S \) and by definition \( S \) consists of nonnegative integers. If \( r \geq n \), then \( r - n = a - (q + 1)n \in S \) since \( r - n \geq 0 \). This would be a contradiction since \( r - n \) is smaller than \( r \). Therefore, \( r < n \) as desired. Thus, we have produced integers \( q \) and \( r \) with \( a = qn + r \) and \( 0 \leq r < n \).

Next, we prove uniqueness of \( q \) and \( r \). Suppose that \( q' \) and \( r' \) are a second pair of integers with \( a = q'n + r' \) and \( 0 \leq r' < n \). Then \( q'n + r' = qn + r \), so \( (q' - q)n = r - r' \). Then \( |q' - q| n = |r - r'| \). Since \( r \) and \( r' \) are both between 0 and \( n - 1 \), the absolute value of their difference is less than \( n \). Since \( |q' - q| n \) is a multiple of \( n \), the only way the equation above can hold is if both sides are 0. Therefore, \( r' = r \) and \( q' = q \). This shows uniqueness of the integers \( q \) and \( r \).

We will use the common terminology of calling the \( r \) above the remainder after dividing \( n \) into \( a \). We can use the division algorithm to prove a simple but useful characterization of the relation congruence modulo \( n \).

**Lemma 1.5.** If \( n \) is a positive integer and \( a, b \) are integers, then \( a \equiv b \mod n \) if and only if \( a \) and \( b \) have the same remainder after division by \( n \). In other words, \( a \equiv b \mod n \) if and only if \( a = qn + r \) and \( b = q'n + r \) for some integers \( q \) and \( q' \).
Proof. Let $n$ be a positive integer and $a, b$ integers. Suppose that $a \equiv b \mod n$. Then $b - a$ is divisible by $n$; say $b - a = tn$. By the division algorithm, we may write $a = qn + r$ and $b = q'n + s$ with $q, q'$ integers, and $r, s$ integers with $0 \leq r, s < n$. Let's suppose that $r \leq s$. Then

$$tn = b - a = q'n + s - (qn + r)$$

$$= (q' - q)n + (s - r).$$

Observe that $s - r < n$, so that the uniqueness of the division algorithm applied to dividing $n$ into $b - a$ shows that $t = q' - q$ and $0 = s - r$. Therefore, $s - r = 0$, or $s = r$. The same conclusion is reached of course if we suppose that $s > r$. This shows that $a$ and $b$ have the same remainder after division by $n$. Conversely, suppose that $a = qn + r$ and $b = q'n + r$ for some integers $q, q', r$. Then

$$b - a = q'n + r - (qn + r) = (q' - q)n,$$

so $b - a$ is divisible by $n$. Therefore, $a \equiv b \mod n$. \qed

Corollary 1.6. Let $n$ be a positive integer. Every integer is congruent modulo $n$ to exactly one integer between 0 and $n - 1$.

Proof. Let $n$ be a positive integer. If $a$ is an integer, then the division algorithm gives us integers $q$ and $r$ with $a = qn + r$ and $0 \leq r < n$. Thus, $a - r = qn$ is divisible by $n$, so $a \equiv r \mod n$. If $s$ is between 0 and $n - 1$ and $a \equiv s \mod n$, then the previous lemma shows us that $a$ and $s$ have the same remainders after division by $n$. But, since $s = 0 \cdot n + s$ and $a = qn + r$, the lemma tells us that $s = r$. Thus, $a$ is congruent modulo $n$ to exactly one integer between 0 and $n - 1$. \qed

Corollary 1.7. If $n$ is a positive integer, then $|\mathbb{Z}_n| = n$ and $\mathbb{Z}_n = \{0, 1, \ldots, n - 1\}$.

Proof. The equivalence classes of integers modulo $n$ are in 1-1 correspondence with the remainders after division by $n$ according to the last corollary. These remainders are the integers in the set $\{0, 1, \ldots, n - 1\}$. This gives us the corollary. \qed

Arithmetic Operations in $\mathbb{Z}_n$

We now discuss a generalization of clock arithmetic for any modulus $n$. Recall that to determine what time it will be 7 hours after 9 o'clock, we add 7 + 9, getting 16, then subtract 12 to get 4, and conclude that the time will be 4 o'clock. In other words, we add the numbers and then subtract enough multiples of 12 to get a valid time. This is the idea behind addition in $\mathbb{Z}_n$. Similarly, we can define multiplication.

Definition 1.8. Let $n$ be a positive integer and let $\bar{a}$ and $\bar{b}$ be elements of $\mathbb{Z}_n$. Then $\bar{a} + \bar{b} = \overline{a + b}$ and $\bar{a} \cdot \bar{b} = \overline{ab}$. 

What this definition tells is that to add two elements of \( \mathbb{Z}_n \), we represent them as the equivalence class of some integers, then we add the integers, then take the equivalence class of the sum. The addition and multiplication tables for \( \mathbb{Z}_2 \) are as follows,

\[
\begin{array}{c|cc}
+ \mod 2 & 0 & 1 \\
0 & 0 & 1 \\
1 & 1 & 0 \\
\end{array}
\quad
\begin{array}{c|cc}
\cdot \mod 2 & 0 & 1 \\
0 & 0 & 0 \\
1 & 0 & 1 \\
\end{array}
\]

and the tables for \( \mathbb{Z}_6 \) are as follows

\[
\begin{array}{c|cccccc}
+ \mod 6 & 0 & 1 & 2 & 3 & 4 & 5 \\
0 & 0 & 1 & 2 & 3 & 4 & 5 \\
1 & 1 & 2 & 3 & 4 & 5 & 0 \\
2 & 2 & 3 & 4 & 5 & 0 & 1 \\
3 & 3 & 4 & 5 & 0 & 1 & 2 \\
4 & 4 & 5 & 0 & 1 & 2 & 3 \\
5 & 5 & 0 & 1 & 2 & 3 & 4 \\
\end{array}
\quad
\begin{array}{c|cccccc}
\cdot \mod 6 & 0 & 1 & 2 & 3 & 4 & 5 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 2 & 3 & 4 & 5 \\
2 & 0 & 2 & 4 & 0 & 2 & 4 \\
3 & 0 & 3 & 0 & 3 & 0 & 3 \\
4 & 0 & 4 & 2 & 0 & 4 & 2 \\
5 & 0 & 5 & 4 & 3 & 2 & 1 \\
\end{array}
\]

If we work in \( \mathbb{Z}_{12} \), we have, for example, that \( 7 + 9 = 16 = 4 \). For an alternative view of addition, the equivalence classes of 7 and of 9 are

\[
\bar{7} = \{ \ldots, -5, 7, 19, \ldots \},
\]
\[
\bar{9} = \{ \ldots, -3, 9, 21, \ldots \}.
\]

We can get a new set of integers by taking all possible sums of the integers in \( \bar{7} \) with the integers in \( \bar{9} \). This set is \( \{ \ldots, -8, 4, 16, 28, \ldots \} \), which is precisely \( \bar{4} \). We can view addition of equivalence classes as this method of adding sets of integers together. The definition we gave above allows us to do modular addition more simply than this set addition. However, there is one problem with the definition. When we write an equivalence class as \( \bar{a} \), this is describing the class by one particular member \( a \) of it. The choice of \( a \) is not unique. For example, \( \bar{7} = \bar{-5} = \bar{91} \), to give three choices. Similarly, \( \bar{9} = \bar{21} \). The problem is this: if we use different representations of two equivalence classes, do we get the same result when we add or multiply? If the answer is no, then we have a meaningless definition. Therefore, we need to verify that our definition is valid. For example, we have \( \bar{7} + \bar{9} = \bar{16} = \bar{4} \), and \( \bar{91} + \bar{21} = \bar{112} = \bar{4} \), since \( 112 = 9 \cdot 12 + 4 \). We show that this example holds in general in the following lemma.

**Lemma 1.9.** Let \( n \) be a positive integer. If \( a, b, c, d \) are integers with \( a \equiv c \mod n \) and \( b \equiv d \mod n \), then \( a + b \equiv c + d \mod n \) and \( ab \equiv cd \mod n \).

**Proof.** Let \( n \) be a positive integer, and suppose that \( a \equiv c \mod n \) and \( b \equiv d \mod n \). Then \( c - a \) and \( d - b \) are divisible by \( n \), so there are integers \( s, t \) with \( a - c = sn \) and \( b - d = tn \). Thus, \( a = c + sn \) and \( b = d + tn \). By adding both equations, we get

\[
a + b = (c + sn) + (d + tn) = (c + d) + (s + t)n,
\]
which shows that \((a + b) - (c + d)\) is a multiple of \(n\). Therefore, \(a + b \equiv c + d \mod n\). If we multiply both equations, we get
\[
ab = (c + sn)(d + tn) = cd + ctn + snd + sntn
\]
\[
= cd + (ct + sd + snt)n,
\]
so \(ab - cd\) is a multiple of \(n\). Thus, \(ab \equiv cd \mod n\). This proves the lemma.

This lemma tells us that Definition 1.8 is meaningful. Because of the simple formula for addition and multiplication in \(\mathbb{Z}_n\), we can obtain analogues of many of the common properties of integer arithmetic. In particular, the following properties hold in \(\mathbb{Z}_n\): if \(\bar{a}, \bar{b}, \bar{c}\) are elements of \(\mathbb{Z}_n\), then

- Commutativity of addition: \(\bar{a} + \bar{b} = \bar{b} + \bar{a}\);
- Associativity of addition: \((\bar{a} + \bar{b}) + \bar{c} = \bar{a} + (\bar{b} + \bar{c})\);
- Existence of an additive identity: \(\bar{a} + \bar{0} = \bar{a}\);
- Existence of additive inverses: \(\bar{a} + \overline{-a} = \bar{0}\);
- Commutativity of multiplication: \(\bar{a} \cdot \bar{b} = \bar{b} \cdot \bar{a}\);
- Associativity of multiplication: \((\bar{a} \cdot \bar{b}) \cdot \bar{c} = \bar{a} \cdot (\bar{b} \cdot \bar{c})\);
- Existence of a multiplicative identity: \(\bar{a} \cdot \overline{1} = \bar{a}\);
- Distributivity: \(\bar{a} \cdot (\bar{b} + \bar{c}) = \bar{a} \cdot \bar{b} + \bar{a} \cdot \bar{c}\).

While we will not write out the proof of all these properties, we give the idea of how to prove them with one example. For commutativity of addition, we have
\[
\bar{a} + \bar{b} = a + b = b + a = \bar{b} + \bar{a}.
\]
Note that we used the definition of addition in \(\mathbb{Z}_n\) twice, and the only other thing we used was the familiar commutative property of addition in \(\mathbb{Z}\). Every other property in the list above comes from a combination of the definition of addition and/or multiplication and the corresponding property of \(\mathbb{Z}\).

We can define subtraction on \(\mathbb{Z}_n\) by \(\bar{a} - \bar{b} = \bar{a} - \bar{b}\). Another way to write subtraction is by \(\bar{a} - \bar{b} = \bar{a} + \overline{-b}\). Because of the fourth property in the list above, \(\overline{-b}\) is the additive inverse of \(\bar{b}\), and the subtraction \(\bar{a} - \bar{b}\) is the same as the sum of \(\bar{a}\) and the additive inverse \(\overline{-b}\) of \(\bar{b}\), just as the case of the real numbers.

There are some differences between arithmetic in \(\mathbb{Z}_n\) and ordinary arithmetic. In \(\mathbb{Z}\), if two integers \(a\) and \(b\) satisfy \(ab = 0\), then either \(a = 0\) or \(b = 0\). However, this is not always true in \(\mathbb{Z}_n\). For example, in \(\mathbb{Z}_{10}\), we have \(\overline{2} \cdot \overline{5} = \overline{10} = \overline{0}\). In \(\mathbb{Z}_{12}\) we have \(\overline{8} \cdot \overline{3} = \overline{24} = \overline{0}\).
However, in $\mathbb{Z}_{11}$, one can show (and we leave it for a homework exercise) that if $a \cdot b = 0$, then $a = 0$ or $b = 0$. We will see shortly that this has consequences for detection of errors in the identification number schemes we discussed earlier. In particular, the ISBN scheme, which utilizes 11, can detect certain types of errors that neither the zip code or UPC scheme always detects, both of which utilize 10.

Another difference has to do with division. If we restrict ourselves to $\mathbb{Z}$, the only time we can solve the equation $ab = 1$ is with $a = b = 1$ or $a = b = -1$. In $\mathbb{Z}_n$, we have the corresponding solutions to the equation $a \cdot b = 1$. However, we may have more solutions. For example, in $\mathbb{Z}_{10}$, we have $3 \cdot 7 = 21 = 1$ and in $\mathbb{Z}_{11}$ we have $8 \cdot 7 = 56 = 1$. In fact, for every $a \in \mathbb{Z}_{11}$ with $a \neq 0$, there is a $b \in \mathbb{Z}_{11}$ with $a \cdot b = 1$. This is also left to a homework exercise. As we will see below, the fact that $3 \cdot x = 1$ in $\mathbb{Z}_{10}$ can be solved was crucial in the decision to use the vector $(3, 1, \ldots, 3, 1)$ in the UPC scheme.

**Greatest Common Divisors**

To facilitate our discussion of error detection, and for future applications, we need some facts about the greatest common divisor of two integers. When we discuss coding theory, we will see that everything we do here will have analogues for polynomials, and those facts will be crucial in constructing and working with error correcting codes.

**Definition 1.10.** Let $a$ and $b$ be integers, at least one of which is nonzero. Then the greatest common divisor of $a$ and $b$, denoted by $\gcd(a, b)$, is the largest integer that divides both $a$ and $b$.

The greatest common divisor of any pair $(a, b)$ of integers exists, provided at least one of them is nonzero. For, if $a \neq 0$, then any divisor of $a$ is no larger than $|a|$. Therefore, the set of common divisors of $a$ and $b$ is bounded above, so there is a largest one. If $a = b = 0$, then every integer divides both $a$ and $b$, so no largest common divisor exists.

For example, we have $\gcd(4, 6) = 2$ and $\gcd(-20, 24) = 4$. Note that the greatest common divisor of two integers is always positive, since if $d$ divides two integers, then both $d$ and $-d$ divides them. Thus, any pair always has a positive common divisor, so the greatest common divisor of the pair is positive.

In the next proposition we prove one of the most useful properties of greatest common divisors. This result is a consequence of the division algorithm, and it gives us a way of writing the gcd of two integers. To help with the proof, we prove the following lemma.

**Lemma 1.11.** Let $a$ and $b$ be integers, and suppose $c$ is an integer that divides both $a$ and $b$. Then $c$ divides $ax + by$ for any integers $x$ and $y$.

**Proof.** Suppose $c$ divides $a$ and $b$, and let $x$ and $y$ be arbitrary integers. Since $c$ divides $a$ and $c$ divides $b$, there are integers $\alpha$ and $\beta$ with $a = \alpha c$ and $b = \beta c$. Then

$$ax + by = \alpha cx + \beta cy = c(\alpha x + \beta y).$$

Since $\alpha x + \beta y$ is an integer, this equation shows that $c$ divides $ax + by$. \qed
For given integers $a, b$, expressions of the form $ax + by$ with $x$ and $y$ integers are called integer linear combinations of $a$ and $b$. More generally, given integers $a_1, a_2, \ldots, a_n$, an expression of the form $\sum_{i=1}^n a_ix_i$, with integers $x_1, x_2, \ldots, x_n$ is called an integer linear combination of $a_1, a_2, \ldots, a_n$.

**Proposition 1.12.** Let $a$ and $b$ be integers, not both zero, and set $d = \gcd(a, b)$. Then there are integers $x$ and $y$ with $d = ax + by$.

**Proof.** To prove this we use an argument reminiscent of that used to prove the division algorithm. Let

$$S = \{as + bt : s, t \in \mathbb{Z}, as + bt > 0\}.$$ 

This is the set of all positive “linear combinations” of $a$ and $b$. Our argument will be to take the least element of $S$ and then prove that this element is the gcd of $a$ and $b$. We first need to see that $S$ is nonempty. To see this, we note that $a^2 + b^2$ is positive and is a linear combination of $a$ and $b$. Thus, this element is an element of $S$. Let $e$ be the least element of $S$. By definition of $S$, there are integers $x$ and $y$ with $e = ax + by$. Set $d = \gcd(a, b)$. We wish to show that $e = d$. First note that since $d$ divides $a$ and $b$, then $d$ divides $e = ax + by$ by the lemma. Since $e$ and $d$ are both positive, this forces $d \leq e$. We show the reverse inequality by first showing that $e$ is a common divisor of $a$ and $b$. By the division algorithm, we may write $a = qe + r$ with $q$ and $r$ integers and $0 \leq r < e$. Then

$$r = a - qe = a - q(ax + by) = a(1 - qx) + b(-qy).$$

If $r > 0$, this equation would show that $r \in S$. This is impossible since $r < e$ and $e$ is the least element of $S$. Therefore, $r = 0$, which means that $a = qe$, so $e$ divides $a$. By a similar argument, $e$ divides $b$. This forces $e \leq d$, since $d$ is the greatest of the common divisors of $a$ and $b$. Since $e \leq d$ and $d \leq e$, we get $e = d$. Therefore, we have written $d = e = ax + by$, as desired.

**Corollary 1.13.** Let $a$ and $b$ be integers, not both zero, and let $d = \gcd(a, b)$. If $c$ is any common divisor of $a$ and $b$, then $c$ divides $d$.

**Proof.** By the proposition, we may write $d = ax + by$ for some integers $x, y$. If $c$ is a common divisor of $a$ and $b$, then $c$ divides $d = ax + by$ by the lemma.

Recall that a positive integer $p$ is prime if the only positive divisors of $p$ are 1 and $p$. It follows immediately that if $a$ is any integer, then $\gcd(a, p) = 1$ or $\gcd(a, p) = p$. Since the latter equality holds only if $p$ divides $a$, we see that $\gcd(a, p) = 1$ for any integer $a$ not divisible by $p$. It is quite common for the condition $\gcd(a, b) = 1$ without neither $a$ or $b$ prime. For example, $\gcd(9, 16) = 1$ and $\gcd(10, 21) = 1$. This condition is important enough for us to study it.

**Definition 1.14.** Two integers $a$ and $b$ are said to be relatively prime if $\gcd(a, b) = 1$. 
The previous proposition has the following consequence.

**Corollary 1.15.** If $a$ and $b$ are integers, not both zero, then $\gcd(a, b) = 1$ if and only if there are integers $x$ and $y$ with $1 = ax + by$.

**Proof.** One direction follows immediately from the proposition: if $\gcd(a, b) = 1$, then $1 = ax + by$ for some integers $x$ and $y$. For the converse, suppose there are integers $x$ and $y$ with $1 = ax + by$. We wish to show that $\gcd(a, b) = 1$. Let $d = \gcd(a, b)$. By the lemma, $d$ divides $1 = ax + by$. However, $d$ is a positive integer, and the only positive integer that divides $1$ is $1$ itself. Therefore, $d = 1$. \hfill \square

A nice consequence of this corollary, in the terminology of modular arithmetic, is the following result.

**Corollary 1.16.** Let $n$ be a positive integer. The equation $\overline{a} \cdot x = \overline{1}$ has a solution in $\mathbb{Z}_n$ if and only if $\gcd(a, n) = 1$.

**Proof.** Let $n$ be a positive integer, and $a$ an integer with $\gcd(a, n) = 1$. Then there are integers $s$ and $t$ with $as + nt = 1$. Rewriting this, we have $as = 1 + n(-t)$, which says that $as \equiv 1 \mod n$, or $\overline{as} = \overline{1}$ in $\mathbb{Z}_n$. Thus, we have produced a solution to the equation $\overline{a}x = \overline{1}$ in $\mathbb{Z}_n$.

Conversely, if $\overline{as} = \overline{1}$ in $\mathbb{Z}_n$, then $as = 1 + n(-t)$ for some integer $t$. Thus $as + nt = 1$ and 1.15 yields that $\gcd(a, n) = 1$. \hfill \square

Caution Corollary 1.15 does not generalize to larger greatest common divisors, i.e. the fact that some (small) integer $c > 1$ may show up as a linear combination of $a$ and $b$ does not guarantee $c = \gcd(a, b)$. To be more specific, let $a = 10$ and $b = 7$. Then we can write $2$ as $2 = 10 \cdot (3) + 7 \cdot (-4)$. However, this does not say that $2$ is the gcd of $10$ and $7$; we know that $\gcd(10, 7) = 1$. Likewise $5 = 10 \cdot 4 + 7 \cdot (-5)$, even though $\gcd(10, 7) \neq 5$. All we can say in general is that the gcd of two integers divides any linear combination of the integers, which clearly happens in this example since $1$ divides $2$ and $5$.

We will not use the following important result to determine the error detection capability of identification number schemes. However, we will use it in other applications.

**Proposition 1.17.** Let $a$, $b$, and $c$ be integers such that $a$ divides $bc$. If $\gcd(a, b) = 1$, then $a$ divides $c$.

**Proof.** Suppose $a$ divides $bc$ and $\gcd(a, b) = 1$. The first condition implies that there is an integer $\alpha$ with $bc = \alpha a$. The second condition and the previous proposition shows that there are integers $x, y$ with $1 = ax + by$. Multiplying this equation by $c$ yields $c = axc + bcy$, and so $c = axc + \alpha y$ by substituting the equation $bc = \alpha a$. Therefore, $c = a(xc + \alpha y)$. Since $xc + \alpha y$ is an integer, $a$ divides $c$, as desired. \hfill \square
The Euclidean Algorithm

The most common method for computing the greatest common divisor of two integers is the Euclidean algorithm. While prime factorization can be used to compute gcds, if the numbers are relatively large, then it is hard to factor them. The Euclidean algorithm has the benefit of being computationally easy to perform. When you have Maple or some other computer program compute gcds, this algorithm is used.

The following steps performs the Euclidean algorithm to calculate \( \gcd(a, b) \) for two positive integers \( a \) and \( b \).

1. Set \( a_0 = \max(a, b) \) and \( a_1 = \min(a, b) \). Set \( i = 1 \).
2. if \( a_i = 0 \), then \( a_{i-1} = \gcd(a, b) \).
3. if \( a_i \neq 0 \), then divide \( a_i \) into \( a_{i-1} \), getting \( a_{i-1} = qa_i + r \).
4. Replace \( i \) by \( i + 1 \).
5. Define \( a_i = r \).
6. Go to step 2

Essentially what you do is this: suppose that \( a < b \). Divide \( a \) into \( b \) and find the remainder. Then divide the remainder into \( a \), finding the second remainder. Then divide the second remainder into the first, getting the third remainder. Continuing this process, the final nonzero remainder is the gcd.

Example 1.18. To perform this algorithm to calculate \( \gcd(10, 14) \), we list the steps one would take.

- Set \( a_0 = 14 \) and \( a_1 = 10 \). We also start with \( i = 1 \).
  
  \[ 14 = 1 \cdot 10 + 4 \]
  
  \[ i = 2 \]
  
  \[ a_2 = 4 \]

- \( a_1 = 10 \) and \( a_2 = 4 \)
  
  \[ 10 = 2 \cdot 4 + 2 \]
  
  \[ i = 3 \]
  
  \[ a_3 = 2 \]

- \( a_2 = 4 \) and \( a_3 = 2 \)
  
  \[ 4 = 2 \cdot 2 + 0 \]
  
  \[ i = 4 \]
  
  \[ a_4 = 0 \]
• $a_3 = 2$ and $a_4 = 0$
  
  since $a_4 = 0$, $a_3 = \gcd(10, 14)$.

Example 1.19. To calculate $\gcd(12342, 2738470)$, we list, without all the notation, the divisions we need to perform to calculate the gcd.

\[
2738470 = 221 \cdot 12342 + 10888 \\
12342 = 1 \cdot 10888 + 1454 \\
10888 = 7 \cdot 1454 + 710 \\
1454 = 2 \cdot 710 + 34 \\
710 = 20 \cdot 34 + 2 \\
34 = 1 \cdot 30 + 4 \\
30 = 7 \cdot 4 + 2 \\
4 = 2 \cdot 2 + 0.
\]

The last nonzero remainder is 2, so $2 = \gcd(12342, 2738470)$.

Example 1.20. To calculate $\gcd(849149, 9889)$, we do the following arithmetic.

\[
849149 = 85 \cdot 9889 + 8584 \\
9889 = 1 \cdot 8584 + 1305 \\
8584 = 6 \cdot 1305 + 754 \\
1305 = 1 \cdot 754 + 551 \\
754 = 1 \cdot 551 + 203 \\
551 = 2 \cdot 203 + 145 \\
203 = 1 \cdot 145 + 58 \\
145 = 2 \cdot 58 + 29 \\
58 = 2 \cdot 29 + 0
\]

Since the final nonzero remainder is 29, we have $29 = \gcd(849149, 9889)$.

Let us now discuss why the Euclidean algorithm works. It is based on two facts. First, if $a$ is any nonzero positive integer, then $\gcd(a, 0) = a$. This is clear from the definition since $a$ is clearly the largest integer that divides $a$, and any nonzero integer divides 0. Second, and more important, is the following result.

Lemma 1.21. Suppose that $a$ and $b$ are integers, not both zero, with $b = qa + r$ for some integers $q$ and $r$. Then $\gcd(a, b) = \gcd(a, r)$. 

CHAPTER 1. IDENTIFICATION NUMBERS AND MODULAR ARITHMETIC

What this lemma says is that to calculate \( \gcd(a, b) \), we can divide \( a \) into \( b \), and then replace \( b \) by the remainder \( r \). The Euclidean algorithm uses this fact repeatedly until a zero remainder is obtained. Then the previous fact is used. To make this more explicit, we work one more example.

**Example 1.22.** To find \( \gcd(24, 112) \), we list both the calculations and what the calculation yields.

\[
egin{align*}
112 &= 4 \cdot 24 + 16 & \gcd(24, 112) &= \gcd(24, 16) \\
24 &= 1 \cdot 16 + 8 & \gcd(24, 16) &= \gcd(16, 8) \\
16 &= 2 \cdot 8 + 0 & \gcd(16, 8) &= \gcd(8, 0) = 8
\end{align*}
\]

The algorithm actually does more than just compute greatest common divisors. It also produces a method for expressing the gcd as a linear combination of the two integers. To illustrate this, we revisit the previous example. We saw that \( \gcd(24, 112) = 8 \). The second to last equation is \( 24 = 1 \cdot 16 + 8 \), or

\[
8 = 1 \cdot 24 + (-1) \cdot 16.
\]

Thus, we have written the gcd 8 as a linear combination of 24 and 16. The equation before that (the first) is \( 112 = 4 \cdot 24 + 16 \), or \( 16 = 1 \cdot 112 + (-4) \cdot 24 \). Replacing 16 in the previous displayed equation by this expression yields

\[
\begin{align*}
8 &= 1 \cdot 24 + (-1) \cdot (1 \cdot 112 + (-4) \cdot 24) \\
&= 1 \cdot 24 - 1 \cdot 112 + 4 \cdot 24 \\
&= 5 \cdot 24 + (-1) \cdot 112,
\end{align*}
\]

which is the desired representation.

**Exercises**

1. For which \( \bar{a} \) in \( \mathbb{Z}_{10} \) does the equation \( \bar{a} \cdot \bar{x} = \bar{0} \) have only the trivial solution \( \bar{x} = \bar{0} \)? When there is a nontrivial solution, give the full solution.

2. For which \( \bar{a} \) in \( \mathbb{Z}_{12} \) does the equation \( \bar{a} \cdot \bar{x} = \bar{0} \) have only the trivial solution \( \bar{x} = \bar{0} \)? When there is a nontrivial solution, give the full solution.

3. For which \( \bar{a} \) in \( \mathbb{Z}_{11} \) does the equation \( \bar{a} \cdot \bar{x} = \bar{0} \) have only the trivial solution \( \bar{x} = \bar{0} \)? When there is a nontrivial solution, give the full solution.

4. Let \( n \) be a positive integer. Based on the previous problems, come up with a precise conjecture for which \( \bar{a} \) in \( \mathbb{Z}_n \) does the equation \( \bar{a} \cdot \bar{x} = \bar{0} \) have only the trivial solution \( \bar{x} = \bar{0} \). State your reasoning for coming up with your conjecture. For extra credit, prove your conjecture.
5. For which $\bar{a}$ in $\mathbb{Z}_{10}$ does the equation $\bar{a} \cdot \bar{x} = \bar{1}$ have a solution? For each $\bar{a}$ for which the equation has a solution, state the solution.

6. Solve the equation $\bar{4} \cdot \bar{x} = \bar{2}$ in $\mathbb{Z}_6$. Can you solve the equation $\bar{4} \cdot \bar{x} = \bar{3}$ in $\mathbb{Z}_6$? Why or why not?

7. Solve the equation $\bar{e} \cdot \bar{x} = \bar{1}$ in $\mathbb{Z}_n$, where $n = 7325494815531218239807$ and $e = 1977326753$. If you do not wish to do this by hand, the values of $n$ and $e$ are input in the Maple worksheet Assignment2.mws on the class drive. Read fully that worksheet and follow the instructions in it.

(We will perform this type of calculation, except with much larger values of $n$, when we talk about cryptography.)

8. Prove the distributive law in $\mathbb{Z}_n$: if $\bar{a}$, $\bar{b}$, and $\bar{c}$ are arbitrary elements of $\mathbb{Z}_n$, then $\bar{a} \cdot (\bar{b} + \bar{c}) = \bar{a} \cdot \bar{b} + \bar{a} \cdot \bar{c}$. State explicitly what facts and/or definitions you use in your proof.

9. If $p$ and $q$ are distinct prime numbers, prove that $p$ and $q$ are relatively prime.

10. Let $a$ and $b$ be integers. If $b = qa + r$ with $q$ and $r$ integers, show that $\gcd(a, b) = \gcd(a, r)$.

11. Explain, with the help of the previous problem, why the Euclidean algorithm produces the gcd of two integers.

12. Apply the Euclidean algorithm, by hand, to find $\gcd(132, 50)$ and to express it as a linear combination of the two given integers. Write out all the steps you take.

13. Use your calculation of the previous problem to find the solution to the equation $50x \equiv 4 \pmod{132}$. Again, do this by hand.

14. Let $n = 7325494815531218239807$ and $e = 1977326753$. Calculate $\gcd(e, n)$, write this gcd as a linear combination of $e$ and $n$, and use this data to solve the equation $ex \equiv 1 \pmod{n}$. Feel free to use Maple; the file Assignment3.mws has the values of $e$ and $n$ along with a reminder of what Maple command will help you do this.

15. Suppose we have an unknown integer $x$ such that $46 + 3x$ is a multiple of 10. Give the justification for each step of the following calculation. Note that some steps require
16. Let $n$ be a positive integer. If $a$ and $b$ are integers with $a \equiv b \mod n$, prove that $a^t \equiv b^t \mod n$ for all positive integers $t$.

17. Suppose that $a, b, n, s$ are integers with $as \equiv b \mod n$. Prove that $\gcd(a, n)$ divides $b$.

18. Suppose that $a, b, n$ are integers such that $\gcd(a, n)$ divides $b$. Prove that $ax \equiv b \mod n$ has a solution. Use the linear combination representation of the gcd to find a solution to the equation $10x \equiv 4 \mod 22$.

(Hint: write $d = \gcd(a, n)$ as a linear combination of $a$ and $n$ and write $b = qd$ for some integer $q$, and manipulate these two equations.)

19. Let $n$ be a positive integer. If $a$ is an integer with $\gcd(a, n) > 1$, prove that there is a nonzero $\bar{b}$ in $\mathbb{Z}_n$ with $\bar{a} \cdot \bar{b} = \bar{0}$ in $\mathbb{Z}_n$. Conclude that the equation $\bar{a} \cdot x = \bar{1}$ cannot be solved in $\mathbb{Z}_n$.

20. Prove the divisibility test for 9: a number, written in normal base 10 form, is divisible by 9 if and only if the sum of its digits is divisible by 9.

21. Prove the divisibility test for 11: a number, written in normal base 10 form, is divisible by 11 if and only if the alternating sum of its digits is divisible by 11. To get the alternating sum, add the first digit, subtract the second, add the third, and so on.

22. Let $a$ be a number written in base 9. Show that $a$ is divisible by 8 if and only if the sum of the base 9 digits is divisible by 8.

(Recall that $a = (a_na_{n-1} \cdots a_1a_0)_9$ is the base 9 representation of $a$ if $0 \leq a_i \leq 8$ for each $i$ and if $a = a_n \cdot 9^n + a_{n-1} \cdot 9^{n-1} + \cdots + a_1 \cdot 9 + a_0$. The sum of the base 9 digits is then $a_n + a_{n-1} + \cdots + a_1 + a_0$.)

23. The greatest common divisor of three integers is the largest integer dividing all three. If $a, b, c$ are nonzero integers, Prove that $\gcd(a, b, c) = \gcd(\gcd(a, b), c)$.

24. Let $n$ be a positive integer and let $\bar{a} \in \mathbb{Z}_n$. 
(a) Prove that the additive inverse of $a$ is unique. In other words, prove that if $\bar{a} + \bar{x} = \bar{0} = \bar{a} + \bar{y}$, then $\bar{x} = \bar{y}$.

(b) Prove that the additive inverse of $a$ is $-a$. In other words, prove that $-(a) = -a$.

25. Solve the equation $14 \cdot x = 3$ in $\mathbb{Z}_{17}$ by first writing $1 = \gcd(14, 17)$ as a linear combination of $14$ and $17$.

26. An open question is whether or not there are infinitely many twin primes; a pair of twin primes is a pair of primes whose difference is 2. Prove that $(3, 5, 7)$ is the only set of triplet primes. That is, if $a, a + 2, a + 4$ are all prime numbers, prove that $a = 3$.

(Hint: show that $a, a + 2, a + 4$ are distinct modulo 3. Conclude that one of them is equivalent to 0 modulo 3.

1.3 Error Detection with Identification Numbers

The purpose of including the check digit in an identification number is to detect errors in reading the number. We now discuss this idea in more detail. First, let us describe all three examples above in a unified way. First of all, we use a generalized notion of the dot product of vectors in $\mathbb{R}^3$; if $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ are $n$-tuples of numbers, then we set

$$a \cdot b = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + \cdots + a_n b_n.$$  

We can represent the test for validity in all three schemes in terms of the dot product. A ten digit number, or more accurately, a 10-tuple $a$ of digits is a valid zip code if $a \cdot (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$ is divisible by 10. Likewise, a UPC is a 12-tuple $a$ of digits such that $a \cdot (3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1)$ is divisible by 10. Finally, a 10-tuple $a$ is a valid ISBN provided that $a \cdot (10, 9, 8, 7, 6, 5, 4, 3, 2, 1)$ is divisible by 11. There are many more examples of identification numbers, most constructed in the following manner: a valid number is an $n$-tuple $a = (a_1, \ldots, a_n)$ satisfying $a \cdot w$ is divisible by some integer $m$ for some fixed $n$-tuple $w$ consisting of integers. For example, $w = (10, 9, 8, 7, 6, 5, 4, 3, 2, 1)$ and $m = 11$ for the ISBN scheme.

Suppose we have an $n$-tuple $w$ consisting of integers, and let $m$ be a positive integer. We can make an identification number scheme as follows. Consider $n$-tuples $a = (a_1, \ldots, a_n)$ where $0 \leq a_i \leq m - 1$. Then $a$ is a valid identification number if $a \cdot w \equiv 0 \text{ mod } m$. We will refer to this scheme as the identification number scheme associated to the vector $w$ and to the integer $m$. The reason for the restriction on the $a_i$ is that $a_i$ and $a_i + m$ are the same modulo $m$. Therefore, the scheme cannot detect the difference between these two numbers. To eliminate this problem, we need to restrict the choices of the $a_i$ so that two different possibilities for $a_i$ are distinguished by the scheme; that is, any two possibilities must be distinct modulo $m$. 
The errors which occur with the most often in identification number schemes are single digit errors and transposition errors, in which adjacent digits are switches. The next proposition shows how to design a scheme which will detect every single digit error.

**Proposition 1.23.** Given \( w \) and \( m \) as above. Then any error in reading the \( i \)-th entry of \( w \) can be detected provided that \( w_i \) is relatively prime to \( m \).

**Proof.** Let \( a = (a_1, \ldots, a_n) \). Suppose that the \( i \)-th entry \( a_i \) is changed to another number, say \( b_i \), with \( 0 \leq b_i \leq m - 1 \). If \( w = (w_1, \ldots, w_n) \), then in \( \mathbb{Z}_m \) we have

\[
\overline{0} = \sum_{i=1}^{n} a_i w_i = \sum_{i=1}^{n} a_i w_i
\]

since \( a \) is a valid number, and by definition of the operations of \( \mathbb{Z}_m \). If \( b = (a_1, \ldots, a_{i-1}, b_i, a_{i+1}, \ldots, a_n) \) is the number obtained by replacing the \( i \)-th entry of \( a \) to \( b_i \), set \( \epsilon = b_i - a_i \) so that \( b_i = a_i + \epsilon \).

Then we have

\[
\overline{b} \cdot w = \sum_{j=1}^{i-1} a_j w_j + b_i w_i + \sum_{j=i+1}^{n} a_j w_j
\]

\[
= \sum_{j=1}^{i-1} a_j w_j + (a_i + \epsilon) w_i + \sum_{j=i+1}^{n} a_j w_j
\]

\[
= a \cdot w + \epsilon w_i.
\]

For \( b \) to be a valid number, \( \overline{b} \cdot w = \overline{0} \). We know that \( a \cdot w = \overline{0} \) since \( a \) is a valid word. Therefore, \( \overline{b} \cdot w - a \cdot w = \epsilon w_i \). But if \( w_i \) is relatively prime to \( m \), then there is an \( \overline{x} \) with \( \overline{w} \overline{x} = 1 \) in \( \mathbb{Z}_m \); this is by Corollary 1.16. Multiplying both sides of the equation \( \overline{0} = \overline{w} \overline{x} \) by \( \overline{x} \) and simplifying gives \( \overline{0} = \overline{\epsilon} = \overline{b_i - a_i} \), or \( \overline{b_i} = \overline{a_i} \). However, this forces \( b_i = a_i \), since both \( a_i \) and \( b_i \) are between 0 and \( m - 1 \). This is a contradiction, so \( b \cdot w \) cannot be zero, and hence the error \( a_i \mapsto b_i \) in reading the \( i \)-th entry of \( a \) will be detected by this identification number scheme. \( \square \)

**Corollary 1.24.** If each entry of \( w \) is relatively prime to \( m \), then the identification number scheme associated to \( w \) can detect a single error in any digit.

We now return to the three schemes we discussed earlier. The zip code scheme uses the weight vector \( w = (1, 1, 1, 1, 1, 1, 1, 1, 1) \) and \( m = 10 \). Each entry of \( w \) is clearly relatively prime to 10, so the zip code scheme will detect single errors. The UPC scheme uses \( w = (3, 1, 3, 1, 3, 1, 3, 1, 3, 1) \) and \( m = 10 \). Both 1 and 3 are relatively prime to 10, so the UPC scheme also detects single errors. In fact, if an identification number scheme works modulo 10, then as long as the vector \( w \) has entries coming from the set \( \{1, 3, 7, 9\} \), then the scheme will detect single errors. The ISBN scheme works modulo 11, and its weight vector is \((10, 9, 8, 7, 6, 5, 4, 3, 2, 1)\). Each entry is relatively prime to 11, so the ISBN scheme also detects single errors.
The corollary above tells us when an identification number scheme detects a single error. Unfortunately, no scheme of the type we have discussed will always detect errors in more than one digit. For example, if the zip code 8800380012 is replaced by 7900380012 by changing the first two digits, then this number satisfies the test for validity. Similarly, if the UPC 0 49000 01134 0 is changed to 0 49000 01163 0 by changing the tenth and eleventh digits, as shown, the number is still valid.

Now let’s consider transposition errors since a common problem with reading of numbers is the tendency to transpose digits. For example, the zip code 8800380012 might be read as 8800830012, by interchanging the 3 and an 8. Some coding schemes detect this type of error, and some do not. For example, the zip code scheme does not detect interchanging of digits. This is because the test for a valid number is that the sum of the ten digits must be divisible by 10; interchanging two digits does not change the sum, so will not change whether or not the test passes. UPC can detect some transposition errors, e.g. interchanging an adjacent 12, but not all, e.g. interchanging an adjacent 16. However, as is to be seen in a homework exercise, the ISBN scheme detects any interchanging of digits. For example, with the ISBN 0387947531, if we interchange the second and third digits, we get 0837947531. The test for validity has us calculate

\[(0, 8, 3, 7, 9, 4, 7, 5, 3, 1) \cdot (10, 9, 8, 7, 6, 5, 4, 3, 2, 1) = 269,\]

which is not divisible by 11. Therefore, the number 0837947531 is not valid. Similarly, if we interchange the final two digits, we get 0387947513, and

\[(0, 3, 8, 7, 9, 4, 7, 5, 1, 3) \cdot (10, 9, 8, 7, 6, 5, 4, 3, 2, 1) = 262,\]

which is also not divisible by 11. The ISBN scheme detects any transposition error. We prove a special case of this here and leave the general case for a homework problem.

**Proposition 1.25.** Suppose that \((a_1, a_2, \ldots, a_{10})\) is a valid ISBN. If \(a_1 \neq a_2\), then \((a_2, a_1, a_3, \ldots, a_{10})\) is not a valid ISBN. Thus, the ISBN scheme detects transposition of the first two digits.

**Proof.** Let \(a = (a_1, a_2, \ldots, a_{10})\) be a valid ISBN. If \(w = (10, 9, 8, 7, 6, 5, 4, 3, 2, 1)\), then \(\overline{a \cdot w} = \overline{a}\) in \(\mathbb{Z}_{11}\). Suppose that \(a_1 \neq a_2\), and set \(b = (a_2, a_1, \ldots, a_{10})\). Then

\[
\overline{a \cdot w - b \cdot w} = 10a_1 + 9a_2 - 10a_2 - 9a_1 = a_1 - a_2.
\]

Since \(a_1 \neq a_2\) and \(0 \leq a_1, a_2 \leq 10\), we see that \(\overline{a_1} \neq \overline{a_2}\). Consequently, the equation above shows that \(\overline{a \cdot w - b \cdot w} \neq \overline{0}\), or \(\overline{b \cdot w} \neq \overline{a \cdot w} = \overline{0}\). Thus, \(b\) is not a valid ISBN.

**Exercises**

1. Let \(a = (a_1, \ldots, a_{10})\) be an invalid ISBN. Show that any single digit of \(a\) can be changed appropriately to give a valid ISBN.
2. Let \((a_1, \ldots, a_{12})\) be a valid UPC. Show that the error in transposing the first two digits of this number is detected by the UPC scheme if and only if \(a_2 - a_1\) is not divisible by 5. Use this to give an example of a UPC \((a_1, \ldots, a_{12})\) with \(a_1 \neq a_2\) but such that \((a_2, a_1, a_3, \ldots, a_{12})\) is also a valid UPC.

3. Consider the following identification number scheme: if \(w = (3, 5, 2, 7)\), then a 4-tuple \(a = (a_1, a_2, a_3, a_4)\), with each \(a_i\) an integer with \(0 \leq a_i \leq 7\), is valid if and only if \(\bar{a} \cdot w = 0\) in \(\mathbb{Z}_8\). Show that this scheme detects any error in the first, second, and fourth digit. Give an example of a valid codeword and an error in the third digit of the codeword that is not detected.

4. Define an identification number scheme as follows: set \(w = (2, 5, 6, 4, 7)\), and a 5-tuple \(a = (a_1, a_2, a_3, a_4, a_5)\) is a valid number if \(0 \leq a_i \leq 8\) for each \(i\), and if \(a \cdot w\) is divisible by 9. Determine which single errors are detected by this scheme. That is, determine for which \(i\) an error in reading the \(i\)-th digit is always detected. Describe how you can change \(w\) in order to guarantee that an error in any digit is always detected.

5. Prove that transposition of any two digits can be detected with the ISBN scheme. That is, if \((a_1, \ldots, a_{10})\) is a valid ISBN, and if \(i < j\) with \(a_i \neq a_j\), show that \((a_1, \ldots, a_j, \ldots, a_i, \ldots, a_{10})\) is not a valid ISBN.

6. The number 0-387-79847-X was obtained from a valid ISBN by transposing two digits. Can you tell which two digits were transposed? Either explain how you can tell which digits were transposed or give an example of two valid ISBNS obtained by transposing two digits of this number.
Chapter 2

Error Correcting Codes

The identification number schemes we discussed in the previous chapter give us the ability to determine if an error has been made in recording or transmitting information. However, they are limited in two ways. First, each allows detection of an error in just one digit, except for some special types of errors, such as interchanging digits. Second, they provide no way to recover the intended information. By making use of more sophisticated ideas and mathematical concepts, we will study methods of encoding and transmitting information that allow us to both detect and correct errors. There are many places that use these so-called error correcting codes, from transmitting photographs from planetary probes to playing of compact discs and DVD movies.

2.1 Basic Notions

To discuss error correcting codes, we need first to set the context and define some terms. We work throughout in binary; that is, we will work over \( \mathbb{Z}_2 \). To simplify notation, we will write the two elements of \( \mathbb{Z}_2 \) as 0 and 1 instead of as \( \overline{0} \) and \( \overline{1} \). If \( n \) is a positive integer, then the set \( \mathbb{Z}_2^n \) is the set of all \( n \)-tuples of \( \mathbb{Z}_2 \)-entries. Elements of \( \mathbb{Z}_2^n \) are called words, or words of length \( n \). For convenience we will write elements of \( \mathbb{Z}_2^n \) either with the usual notation, or as a concatenation of digits. For instance, we will write \((0,1,0,1)\) and 0101 for the same 4-tuple. We can equip \( \mathbb{Z}_2^n \) with an operation of addition by using point-wise addition. That is, we define

\[
(a_1, \ldots, a_n) + (b_1, \ldots, b_n) = (a_1 + b_1, \ldots, a_n + b_n).
\]

Note that, as a consequence of the facts that \( 0 + 0 = 0 = 1 + 1 \) in \( \mathbb{Z}_2 \), we have \( a + a = 0 \) for any \( a \in \mathbb{Z}_2^n \), where \( 0 \) is the vector \((0, \ldots, 0)\) consisting of all zeros.

A linear code of length \( n \) is a nonempty subset of \( \mathbb{Z}_2^n \) that is closed under the addition in \( \mathbb{Z}_2^n \). Much of the discussion simplifies greatly for linear codes and, although nonlinear codes exist, linear codes are used most frequently in applications. Because of their importance we will consider only linear codes and drop the adjective “linear” from now on. We will refer to elements of a code as codewords.
Example 2.1. The set \{00, 01, 10, 11\} is a code of length 2, and \{0000, 1010, 0101, 1111\} is a code of length 4.

Let \( w = a_1 \cdots a_n \) be a word of length \( n \). Then the weight of \( w \) is the number of digits of \( w \) equal to 1. We denote the weight of \( w \) by \( \text{wt}(w) \). An equivalent and useful way to think about the weight of the word \( w = a_1 \cdots a_n \) is to treat the \( a_i \) as the integers 0 or 1 rather than as residue classes for the moment and note that

\[
\text{wt}(w) = \sum_{i=1}^{n} a_i.
\]

There are some obvious consequences of this definition. First of all, \( \text{wt}(w) = 0 \) if and only if \( w = 0 \). Second, \( \text{wt}(w) \) is a nonnegative integer. A more sophisticated fact about weight is its relation with addition. If \( v, w \in \mathbb{Z}_2^n \), then \( \text{wt}(v + w) \leq \text{wt}(v) + \text{wt}(w) \). To see why this is true, we write \( x_i \) for the \( i \)-th component of a word \( x \). The weight of \( x \) is then given by the equation

\[
\text{wt}(x) = |\{i : 1 \leq i \leq n, x_i = 1\}|.
\]

Using this description of weight, we note that \( (v + w)_i = v_i + w_i \). Therefore, if \( (v + w)_i = 1 \), then either \( v_i = 1 \) or \( w_i = 1 \) (but not both). Therefore,

\[
\{i : 1 \leq i \leq n, (v + w)_i = 1\} \subseteq \{i : v_i = 1\} \cup \{i : w_i = 1\}.
\]

Since \( |A \cup B| \leq |A| + |B| \) for any two finite sets \( A, B \), the inclusion above yields \( \text{wt}(v + w) \leq \text{wt}(v) + \text{wt}(w) \), as desired.

From the idea of weight we can define the notion of distance on \( \mathbb{Z}_2^n \). If \( v, w \) are words, then we set the distance \( D(v, w) \) between \( v \) and \( w \) to be

\[
D(v, w) = \text{wt}(v + w).
\]

Alternatively, \( D(v, w) \) is equal to the number of positions in which \( v \) and \( w \) differ. The function \( D \) shares the basic properties of distance in Euclidean space \( \mathbb{R}^3 \). More precisely, it satisfies the properties of the following lemma.

Lemma 2.2. The distance function \( D \) defined on \( \mathbb{Z}_2^n \times \mathbb{Z}_2^n \) satisfies

1. \( D(v, v) = 0 \) for all \( v \in \mathbb{Z}_2^n \);

2. for any \( v, w \in \mathbb{Z}_2^n \), if \( D(v, w) = 0 \), then \( v = w \);

3. \( D(v, w) = D(w, v) \) for any \( v, w \in \mathbb{Z}_2^n \);

4. triangle inequality: \( D(v, w) \leq D(v, u) + D(u, w) \) for any \( u, v, w \in \mathbb{Z}_2^n \).

Proof. Since \( v + v = 0 \), we have \( D(v, v) = \text{wt}(v + v) = \text{wt}(0) = 0 \). This proves (1). We note that \( 0 \) is the only word of weight 0. Thus, if \( D(v, w) = 0 \), then \( \text{wt}(v + w) = 0 \), which forces \( v + w = 0 \). However, adding \( w \) to both sides yields \( v = w \), and this proves (2). The equality
2.1. BASIC NOTIONS

\[ D(v, w) = D(w, v) \] is obvious since \( v+w = w+v \). Finally, we prove (4), the only non-obvious statement, with a cute argument. Given \( u, v, w \in \mathbb{Z}_2^n \), we have, from the definition and the fact about weight given above,

\[
D(v, w) = \text{wt}(v + w) = \text{wt}((v + u) + (u + w)) \\
\leq \text{wt}(v + u) + \text{wt}(u + w) \\
= D(v, u) + D(u, w).
\]

\[ \square \]

To discuss error correction we must first formalize the notion. Let \( C \) be a code. If \( w \) is a word, to correct, or decode, \( w \) means to select the codeword \( v \in C \) such that

\[
D(v, w) = \min \{ D(u, w) : u \in C \}.
\]

In other words, we decode \( w \) by choosing the closest codeword to \( w \), under our notion of distance. There may not be a unique closest codeword, however. When this happens we can either randomly select a closest codeword, or do nothing. We refer to this notion of decoding as maximum likelihood detection, or MLD.

**Example 2.3.** Let \( C = \{00000, 10000, 011000, 11100\} \). If \( w = 10001 \), then \( w \) is distance 1 from 10000 and distance more than 1 from the other two codewords. Thus, we would decode \( w \) as 10000. However, if \( u = 11000 \), then \( u \) is distance 1 from both 10000 and from 11100. Thus, either is an appropriate choice to decode \( u \).

We now define what it means for a code to be an error correcting code.

**Definition 2.4.** Let \( C \) be a code and let \( t \) be a positive integer. Then \( C \) is a \( t \)-error correcting code if whenever a word \( w \) differs from the nearest codeword \( v \) by a distance of at most \( t \), then \( v \) is the unique closest codeword to \( w \).

If a codeword \( v \) is transmitted and received as \( w \), we can express \( w \) as \( v + u \), and we say that \( u = v + w \) is the error in transmission. As a word, the error \( u \) has a certain weight. So \( C \) is \( t \)-error correcting if for every codeword \( v \) and every word \( u \) whose weight is at most \( t \), then \( v \) is the unique closest codeword to \( v + u \).

If \( C \) is a \( t \)-error correcting code, then we say that \( C \) corrects \( t \) errors. Thus one way of interpreting the definition is that if \( v \) is a codeword, and if \( w \) is obtained from \( v \) by changing at most \( t \) entries of \( v \), then \( v \) is the unique closest codeword to \( w \). Therefore, by MLD decoding, \( w \) will be decoded as \( v \).

**Example 2.5.** The code \( C = \{00000, 10000, 011000, 11100\} \) in the previous example corrects no errors. Note that the word \( u = 11000 \) given in that example is a distance 1 from a codeword, but that codeword is not the unique closest codeword to \( u \).
To determine for what \( t \) a code corrects \( t \) errors, we relate error correction to the distance of a code.

**Definition 2.6.** The distance \( d \) of a code is defined by \( d = \min \{ D(u, v) : u, v \in C, u \neq v \} \).

We denote by \( \lfloor a \rfloor \) the greatest integer less than or equal to the number \( a \).

**Proposition 2.7.** Let \( C \) be a code of distance \( d \) and set \( t = \lfloor (d - 1)/2 \rfloor \). Then \( C \) is a \( t \)-error correcting code but not a \( (t + 1) \)-error correcting code.

**Proof.** Let \( w \) be a word, and suppose that \( v \) is a codeword with \( D(v, w) \leq t \). We need to prove that \( v \) is the unique closest codeword to \( w \). We do this by proving that \( D(u, w) > t \) for any codeword \( u \neq v \). If not, suppose that \( u \) is a codeword with \( u \neq v \) and \( D(u, w) \leq t \). Then, by the triangle inequality,

\[
D(u, v) \leq D(u, w) + D(w, v) \leq t + t = 2t < d.
\]

This is a contradiction to the definition of \( d \). Thus, \( v \) is indeed the unique closest codeword to \( w \).

To finish the proof, we need to prove that \( C \) does not correct \( t + 1 \) errors. Since the code has distance \( d \), there are codewords \( u_1, u_2 \) with \( d = D(u_1, u_2) \), in other words, \( u_1 \) and \( u_2 \) differ in exactly \( d \) positions. Let \( w \) be the word obtained from \( u_1 \) by changing exactly \( t + 1 \) of those \( d \) positions. Then \( D(u_1, w) = t + 1 \) and \( D(u_2, w) = d - (t + 1) \). Since \( t = \lfloor (d - 1)/2 \rfloor \) by our assumption, \( (d - 2)/2 \leq t \leq (d - 1)/2 \). In particular, \( d - 2 \leq 2t \) so that \( D(u_2, w) = d - (t + 1) \leq t + 1 \). Thus \( u_1 \) is not the unique closest codeword to \( w \), since \( u_2 \) is either equally close or closer to \( w \). Therefore, \( C \) is not a \((t + 1)\)-error correcting code.
Example 2.8. Let $C = \{00000, 00111, 11100, 11011\}$. The distance of $C$ is 3, and so $C$ is a 1-error correcting code.

Example 2.9. Let $n$ be an odd positive integer, and let $C = \{0 \cdot \cdot 0, 1 \cdot \cdot 1\}$ be a code of length $n$. If $n = 2t + 1$, then $C$ is a $t$-error correcting code since the distance of $C$ is $n$. Thus, by making the length of $C$ long enough, we can correct any number of errors that we wish. However, note that the fraction of components of a word that can be corrected is $t/n$, and this is always less than 1/2.

Exercises

1. Find distance and error correction capability of the following codes:
   
   (a) $\{000000, 1010101, 0101010, 1111111\}$,
   
   (b) $\{0000000, 11111111, 11100000, 00011111\}$,
   
   (c) $\{00000000, 11110000, 00001111, 10101010, 1111111, 01011010, 10100101, 01010101\}$.

2. Construct a linear code of length 5 with more than 2 codewords that corrects one error. Can you construct a linear code of length 4 with more than 2 words that corrects one error?

3. Let $C$ be the code consisting of the solutions to the matrix equation $Ax = 0$, where

   $$A = \begin{pmatrix}
   1 & 0 & 1 & 1 & 1 & 0 \\
   0 & 1 & 1 & 1 & 0 & 1 \\
   1 & 1 & 0 & 0 & 1 & 1
   \end{pmatrix}.$$ 

   Determine the codewords of $C$, and determine the distance and error correction capability of $C$.

4. Let $A$ be a matrix, and let $C$ be the code consisting of all solution to $Ax = 0$. If $A$ has neither a column of zeros nor two equal columns, prove that the distance of $C$ is at least 3.

   (Hint: If $v$ has weight 1 or weight 2, look at how $Av$ can be written in terms of the columns of $A$.)

5. Let $C$ be a code such that if $u, v \in C$, then $u + v \in C$. Prove that the distance of $C$ is equal to the smallest weight of a nonzero codeword.

6. Let $C$ be the code consisting of all solutions to a matrix equation $Ax = 0$. Let $d$ be the largest integer such that any sum of fewer than $d$ columns of $A$ is nonzero. Prove that $C$ has distance $d$. 
2.2 Gaussian Elimination

In this section we discuss the idea of Gaussian elimination for matrices with entries in \( \mathbb{Z}_2 \). We do this now precisely because we need to work with matrices with entries in \( \mathbb{Z}_2 \) in order to discuss the Hamming code, our first example of an error correcting code.

In linear algebra, if you are given a system of linear equations, then you can write this system as a single matrix equation \( AX = b \), where \( A \) is the matrix of coefficients, and \( X \) is the column matrix of variables. For example, the system

\[
2x + 3y - z = 1 \\
x - y + 5z = 2
\]

is equivalent to the matrix equation

\[
\begin{pmatrix}
2 & 3 & -1 \\
1 & -1 & 5
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
=
\begin{pmatrix}
1 \\
2
\end{pmatrix}.
\]

The primary matrix-theoretic method for solving such a system is to perform Gaussian elimination on the augmented matrix, that matrix that adds to the coefficient matrix one column at the right equal to the column on the right side of the equation. Recall that Gaussian elimination performs operations on the rows of a matrix in order to replace the matrix by one in which the solution to the system can be found easily. There are three such row operations:

- multiply or divide a row by a nonzero scalar,
- interchange two rows,
- add a multiple of one row to another row.

It is likely that in all your work with matrices, the entries of the matrices were real numbers. However, to perform the row operations, all you need is to be able to add, subtract, multiply, and divide the entries. In many situations, matrices arise whose entries are not real numbers. For coding theory we need to work with matrices whose entries lie in \( \mathbb{Z}_2 \). Within this set we can add, subtract, multiply, and divide just as if we had real numbers, with the usual warning prohibiting division by 0. Furthermore, all the theorems of linear algebra have analogues to the setting where entries lie in \( \mathbb{Z}_2 \). In fact, we will generalize the idea of linear algebra later on to include many more sets of scalars. Again, all we need is to be able to perform the four arithmetic operations on the scalars, and we need properties analogous to those that hold for real number arithmetic.

Recall that the only symbolic difference between \( \mathbb{Z}_2 \) arithmetic and ordinary arithmetic of these symbols is that \( 1 + 1 = 0 \) in \( \mathbb{Z}_2 \). Note that the first of the three row operations...
listed above is not useful; multiplying a row by 1 does not affect the row, so is an operation that is not needed. Also, the third operation in the case of $\mathbb{Z}_2$ reduces to adding one row to another.

Before working some examples, we recall what it means for a matrix to be in row reduced echelon form.

**Definition 2.10.** A matrix $A$ is in row reduced echelon form if

1. the first nonzero entry of any row is 1. This entry is called a leading 1;

2. If a column contains a leading 1, then all other entries of the column are 0;

3. If $i > j$, and if row $i$ and row $j$ each contain a leading 1, then the column containing the leading 1 of row $i$ is further to the right than the column containing the leading 1 of row $j$.

To help understand Condition 3 of the definition, the leading 1’s go to the right as you go from top to bottom in the matrix, so that the matrix is in some sense triangular.

We now give several examples of reducing matrices with $\mathbb{Z}_2$ entries to echelon form. In each example once we have the matrix in row reduced echelon form, the leading 1’s are marked in boldface. In understanding the computations below, recall that since $-1 = 1$ in $\mathbb{Z}_2$, subtraction and addition are the same operation.

**Example 2.11.** Consider the matrix

$$
\begin{pmatrix}
1 & 0 & 0 & 1 \\
1 & 1 & 0 & 1 \\
0 & 1 & 1 & 1
\end{pmatrix}.
$$

We reduce the matrix with the following steps. You should determine which row operation was done in each step.

$$
\begin{pmatrix}
1 & 0 & 0 & 1 \\
1 & 1 & 0 & 1 \\
0 & 1 & 1 & 1
\end{pmatrix} \Rightarrow \begin{pmatrix}
1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 \\
0 & 1 & 1 & 1
\end{pmatrix} \Rightarrow \begin{pmatrix}
1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 1
\end{pmatrix}.
$$

**Example 2.12.** Consider the matrix

$$
\begin{pmatrix}
1 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1
\end{pmatrix}.
$$
To reduce this matrix, we can do the following steps.

\[
\begin{pmatrix}
1 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 \\
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
1 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 \\
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
1 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 1 & 0 & 1 & 1 \\
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
1 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{pmatrix}.
\]

**Example 2.13.** To reduce the matrix

\[
\begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 \\
\end{pmatrix}
\]

we can apply the following single row operation.

\[
\begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 \\
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
\end{pmatrix}.
\]

We now recall why having a matrix in row reduced echelon form will give us the solution to the corresponding system of equations. The row operations on the augmented matrix correspond to performing various algebraic manipulations to the equations, such as interchanging equations. So, the system of equations corresponding to the reduced matrix is equivalent to the original system; that is, the two systems have exactly the same solutions.

**Example 2.14.** Consider the system of equations

\[
\begin{align*}
x &= 1 \\
x + y &= 1 \\
y + z &= 1.
\end{align*}
\]

This system has augmented matrix

\[
\begin{pmatrix}
1 & 0 & 0 & 1 \\
1 & 1 & 0 & 1 \\
0 & 1 & 1 & 1 \\
\end{pmatrix},
\]

and reducing this matrix yields

\[
\begin{pmatrix}
1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
\end{pmatrix}.
\]
This new matrix corresponds to the system of equations

\[
\begin{align*}
    x &= 1, \\
    y &= 0, \\
    z &= 1.
\end{align*}
\]

Thus, we have already the solution to the original system.

**Example 2.15.** The augmented matrix

\[
\begin{pmatrix}
    1 & 1 & 0 & 0 & 1 & 0 \\
    1 & 0 & 1 & 0 & 0 & 1 \\
    0 & 1 & 1 & 1 & 1 & 0 \\
    0 & 1 & 1 & 0 & 1 & 1
\end{pmatrix}
\]

corresponds to the system of equations

\[
\begin{align*}
    x_1 + x_2 + x_5 &= 0, \\
    x_1 + x_3 &= 1, \\
    x_2 + x_3 + x_4 + x_5 &= 0, \\
    x_2 + x_3 + x_5 &= 1.
\end{align*}
\]

Reducing the matrix yields

\[
\begin{pmatrix}
    1 & 0 & 1 & 0 & 0 & 1 \\
    0 & 1 & 1 & 0 & 1 & 1 \\
    0 & 0 & 0 & 1 & 0 & 1 \\
    0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

which corresponds to the system of equations

\[
\begin{align*}
    x_1 + x_3 &= 1, \\
    x_2 + x_3 + x_5 &= 1, \\
    x_4 &= 1.
\end{align*}
\]

We have left the leading ones in boldface in the echelon matrix. These correspond to the variables \(x_1, x_2, \) and \(x_4\). These variables can be solved in terms of the other variables. Thus, we have the full solution

\[
\begin{align*}
    x_1 &= 1 + x_3, \\
    x_2 &= 1 + x_3 + x_5, \\
    x_4 &= 1, \\
    x_3 \text{ and } x_5 \text{ are arbitrary}.
\end{align*}
\]
We can write out all solutions to this system of equations, since each of \(x_3\) and \(x_5\) can take on the two values 0 and 1. This gives us four solutions, which we write as row vectors:

\[
(x_1, x_2, x_3, x_4, x_5) = (1 + x_3, 1 + x_3 + x_5, x_3, 1, x_5),
\]

where \(x_3 = 0, 1\) and \(x_5 = 0, 1\).

Note that \((1 + x_3, 1 + x_3 + x_5, x_5, 1, x_5)\) = \((1, 1, 0, 1, 0) + x_3(1, 1, 1, 0, 0) + x_5(0, 1, 0, 0, 1)\). Since \((1, 1, 0, 1, 0)\) corresponds to the values \(x_3 = x_5 = 0\) in *, this yields a particular solution to the linear system. On the other hand, the vectors \((1, 1, 1, 0, 0), (0, 1, 0, 0, 1)\) solve the homogeneous system

\[
\begin{align*}
x_1 + x_2 + x_5 &= 0, \\
x_1 + x_3 &= 0, \\
x_2 + x_3 + x_4 + x_5 &= 0, \\
x_2 + x_3 + x_5 &= 0
\end{align*}
\]

(check this!), thus any solution to the inhomogeneous system is obtained as the sum of a particular solution and a solution to the associated homogenous system.

**Example 2.16.** Let \(H\) be the Hamming matrix (named for Richard Hamming, mathematician, pioneer computer scientist, and inventor of the Hamming error correcting codes):

\[
H = \begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1
\end{pmatrix},
\]

and consider the homogeneous system of equations \(HX = 0\), where \(0\) refers to the 3 \(\times\) 1 zero matrix. Also, \(X\) is a 7 \(\times\) 1 matrix of seven variables \(x_1, \ldots, x_7\). To solve this system we reduce the augmented matrix

\[
\begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 1
\end{pmatrix},
\]

yielding

\[
\begin{pmatrix}
1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 & 0
\end{pmatrix}.
\]

This matrix corresponds to the system of equations

\[
\begin{align*}
x_1 + x_3 + x_5 + x_7 &= 0, \\
x_2 + x_3 + x_6 + x_7 &= 0, \\
x_4 + x_5 + x_6 + x_7 &= 0.
\end{align*}
\]
Again, we have marked the leading 1’s in boldface, and the corresponding variables can be solved in terms of the others, which can be arbitrary. So, the solution to this system is

\[
\begin{align*}
x_1 &= x_3 + x_5 + x_7, \\
x_2 &= x_3 + x_6 + x_7, \\
x_4 &= x_5 + x_6 + x_7, \\
\end{align*}
\]

\(x_3, x_5, x_6, x_7\) are arbitrary.

Since we have four variables, \(x_3, x_5, x_6,\) and \(x_7,\) that are arbitrary, and since there are two scalars in \(\mathbb{Z}_2,\) each variable can take two values. Therefore, we have \(2^4 = 16\) solutions to this system of equations.

To finish this chapter, we recall a theorem that will help us determine numeric data about error correcting codes. To state the theorem we need to recall some terminology of linear algebra. We will not bother to define all of the terms here; you should review them in a linear algebra textbook. The definitions will be given later when we discuss vector spaces. The row space of a matrix is the vector space spanned by the rows of the matrix. If the matrix is \(m \times n,\) then the rows are \(n\)-tuples, so the row space is a subspace of the space of all \(n\)-tuples. The column space of a matrix is the space spanned by the columns of the matrix. Again, if the matrix is \(m \times n,\) then the columns are \(m\)-tuples, so the column space is a subspace of the space of all \(m\)-tuples. The dimension of the row space and the dimension of the column space are always equal. This common positive integer is called the rank of a matrix \(A.\) One benefit to reducing a matrix \(A\) to echelon form is that the rows of the reduced matrix that contain a leading 1 form a basis for the row space of \(A.\) Consequently, the dimension of the row space is the number of leading 1’s. Thus, an alternative definition of the rank of a matrix is that it is equal to the number of leading 1’s in the row reduced echelon form obtained from the matrix.

The kernel, or nullspace, of a matrix \(A\) is the set of all solutions to the homogeneous equation \(AX = 0.\) To help understand this example, consider the Hamming matrix \(H\) of the previous example.

**Example 2.17.** The solution to the homogeneous equation \(HX = 0\) from the previous example is

\[
\begin{align*}
x_1 &= x_3 + x_5 + x_7, \\
x_2 &= x_3 + x_6 + x_7, \\
x_4 &= x_5 + x_6 + x_7, \\
\end{align*}
\]

\(x_3, x_5, x_6, x_7\) are arbitrary.

In order to describe systematically the set of solutions, we find a basis for the nullspace in the following manner. Successively set one of the variables described as "arbitrary" equal to 1 and all other arbitrary variables equal to 0. The resulting vector will be a solution to
HX = 0. If we do this for each "arbitrary" variable, we will have a basis for the nullspace. Doing this, we get the four vectors

\[
\begin{pmatrix}
1 \\
1 \\
1 \\
0 \\
0 \\
0 \\
0 \\
1
\end{pmatrix} = \begin{pmatrix}
0 \\
1 \\
1 \\
1 \\
1 \\
0 \\
0 \\
1
\end{pmatrix} = \begin{pmatrix}
0 \\
0 \\
0 \\
1 \\
1 \\
0 \\
0 \\
1
\end{pmatrix}.
\]

These vectors do form a basis for the nullspace of \(H\) since the general solution of \(HX = 0\) is

\[
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4 \\
x_5 \\
x_6 \\
x_7
\end{pmatrix} = \begin{pmatrix}
x_3 + x_5 + x_7 \\
x_3 + x_6 + x_7 \\
x_3 \\
x_5 + x_6 + x_7 \\
x_5 \\
x_6 \\
x_7
\end{pmatrix} = \begin{pmatrix}
1 \\
1 \\
1 \\
0 \\
0 \\
0 \\
0
\end{pmatrix} + \begin{pmatrix}
1 \\
0 \\
0 \\
1 \\
0 \\
0 \\
1
\end{pmatrix} + \begin{pmatrix}
0 \\
1 \\
0 \\
1 \\
0 \\
0 \\
1
\end{pmatrix}.
\]

From this equation, we see that every solution is a linear combination of the four specific solutions written above, and a little work will show that every solution can be written in a unique way as a linear combination of these vectors. For example, one can check that \((0,1,1,1,1,0,0)\) is a solution to the system \(HX = 0\), and that to write this vector as a linear combination of the four given vectors, we must have \(x_3 = x_5 = 1\) and \(x_6 = x_7 = 0\), and so

\[
\begin{pmatrix}
0 \\
1 \\
1 \\
1 \\
0 \\
0 \\
0 \\
0
\end{pmatrix} = \begin{pmatrix}
1 \\
1 \\
0 \\
1 \\
0 \\
0 \\
0 \\
0
\end{pmatrix} + \begin{pmatrix}
1 \\
1 \\
1 \\
0 \\
0 \\
0 \\
0 \\
0
\end{pmatrix}.
\]

is a sum of two of the four given vectors, and can be written in no other way in terms of the four.

This example indicates the following general fact that for a homogeneous system \(AX = 0\), the number of variables not corresponding to leading 1’s is equal to the dimension of the nullspace of \(A\). Let us call these variables basic variables. If we reduce \(A\), the basic variables can be solved in terms of the other variables, and since these other variables can take on arbitrary values, we call them free variables. By mimicking the example above, any solution can be written uniquely in terms of a set of solutions, one for each free variable. This set
of solutions is a basis for the nullspace of $A$; therefore, the number of free variables is equal to the dimension of the nullspace. Every variable is then either a basic variable or a free variable. The number of variables is the number of columns of the matrix. This observation leads to the rank-nullity theorem. The *nullity* of a matrix $A$ is the dimension of the nullspace of $A$.

Let $A$ be an $n \times m$ matrix. Then $m$ is equal to the sum of the rank of $A$ and the nullity of $A$.

The point of this theorem is that once you know the rank of $A$, the nullity of $A$ can be immediately calculated. The number of solutions to $AX = 0$ can then be found. In coding theory this will allow us to determine the number of codewords in a given code.

### 2.3 The Hamming Code

The Hamming code, discovered independently by Hamming and Golay, was the first example of an error correcting code. Let $$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$ be the Hamming matrix, described in Example 2.16 above. Note that the columns of this matrix give the base 2 representation of the integers 1-7. The Hamming code $C$ of length 7 is the nullspace of $H$. More precisely, $$C = \{ v \in \mathbb{Z}_2^7 : Hv^T = 0 \}.$$

Also by Gaussian elimination, we can solve the linear equation $Hx = 0$, and we get the solution

$$x_1 = x_3 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + x_5 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} + x_6 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} + x_7 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Therefore, $C$ has dimension 4, and the set $\{1110000, 1001100, 0101010, 1101001\}$ forms a basis for $C$. If we write out all 16 codewords in $C$, we will see that the distance of $C$ is 3. Thus, $C$ is a $(7, 4, 3)$-code, the 7 referring to the length of the code, the 4 to its dimension, and the 3 to its minimum distance. It will then correct 1 error.

The code $C$ has a particularly elegant decoding algorithm, which we now describe. Let $e_1, \ldots, e_7$ be the standard basis for $\mathbb{Z}_2^7$. We point out a simple fact of matrix multiplication:
$He_i^T$ is equal to the $i$-th column of $H$. Moreover, we note that the 7 nonzero vectors in $K^3$ are exactly the 7 columns of $H$.

Suppose that $v$ is a codeword that is transmitted as a word $w \neq v$. Suppose that exactly one error has been made in transmission. Then $w = v + e_i$ for some $i$. However, we do not yet know $i$, so we cannot yet determine $v$ from $w$. However,

$$Hw^T = H(v + e_i)^T = Hv^T + He_i^T = He_i^T,$$

and $He_i^T$ is the $i$-th column of $H$, as we pointed out above. Therefore, we can determine $i$ by computing $Hw^T$ and determining which column of $H$ this. Then $w$ is decoded to $w + e_i$, which must be equal to $v$ since we assumed that only one error was made in transmission.

The Hamming code $C$ has an additional property: every word is within 1 of a codeword. To see this, suppose that $w$ is a word. If $Hw^T = 0$, then $w$ is a codeword. If not, then $Hw^T$ is a nonzero 3-tuple. Therefore, it is equal to a column of $H$; say that $Hw^T$ is equal to the $i$-th column of $H$. Then $Hw^T = He_i^T$, so $H(w^T + e_i^T) = 0$. Therefore, $w + e_i := v \in C$.

The word $v$ is then a codeword a distance of 1 from $w$. A code that corrects $t$ errors and for which every word is within $t$ of some codeword is called perfect. Such codes are particularly nice, in part because a decoding procedure will always return a word. Later we will see some important codes that are not perfect.

**Exercises**

1. Let $C$ be the code (of length $n$) of solutions to a matrix equation $Ax = 0$. Define a relation on the set $\mathbb{Z}_2^n$ of words of length $n$ by $u \equiv v \pmod{C}$ if $u + v \in C$. Prove that this is an equivalence relation, and that for any word $w$, the equivalence class of $w$ is the coset $C + w$.

**2.4 Coset Decoding**

To apply MLD (Maximum Likelihood Decoding, section 2.1) what we must do, given a received word $w$, is search through all the codewords to find the codeword $c$ closest to $w$. This can be a slow and tedious process. There are more efficient methods, assuming the code is built in a manner similar to that of the Hamming code. We will assume that we have a code $C$ of length $n$ such that there is an $m \times n$ matrix $H$ with $C = \{v \in \mathbb{Z}_2^n : Hv^T = 0\}$. We will fix the symbols $C$ and $H$ in this section.

**Definition 2.18.** Let $w$ be a word. Then the coset $C + w$ of $w$ is the set $\{c + w : c \in C\}$.

Recall two facts about $C$. First, by the definition of $C$, the zero vector $\mathbf{0}$ is an element of the code, since $H\mathbf{0} = \mathbf{0}$. From this we see that $w \in C + w$, since $w = \mathbf{0} + w$. Second, if $u, v \in C$, our definitions require that $u + v \in C$ ($H(u + v) = Hu + Hv = \mathbf{0} + \mathbf{0} = \mathbf{0}$).
We now discuss an important property of cosets, namely that any two cosets are either equal or are disjoint. In fact cosets are the equivalence classes for an the following equivalence relation defined on $C$:

Two words $x$ and $y$ are related if $x + y \in C$.

We write $x \sim y$ when this occurs. To see that this is an equivalence relation, we must verify the three properties. To show reflexivity, we have $x \sim x$ since $x + x = 0$, which is an element of $C$. Next, suppose that $x \sim y$. We must show that $y \sim x$. The assumption that $x \sim y$ means $x + y \in C$. However, $x + y = y + x$; therefore, since $y + x \in C$, we have $y \sim x$. Finally, to see transitivity, suppose that $x \sim y$ and $y \sim z$. Then $x + y \in C$ and $y + z \in C$. If we add these codewords, we will get a codeword, by the previous paragraph. However,

$$(x + y) + (y + z) = x + (y + y) + z = x + 0 + z = x + z,$$

by the properties of vector addition. since the result $x + z$ is an element of $C$, we have $x \sim z$, as desired. So, we have an equivalence relation. The equivalence class of a word $x$ is

$$\{y : y \sim x\} = \{y : x + y \in C\} = \{y : y = c + x \text{ for some } c \in C\} = C + x.$$

The third equality follows since if $x + y = c$, then $y = c + x$.

**Proposition 2.19.** If $x$ and $y$ are words, then $C + x = C + y$ if and only if $Hx = Hy$.

**Proof.** Suppose first that $C + x = C + y$. Then $x \sim y$, so $x + y \in C$. By definition of $C$, we have $H(x + y) = 0$. Expanding the left hand side, we get $Hx + Hy = 0$, so $Hx = Hy$. Conversely, suppose that $Hx = Hy$. Then $Hx + Hy = 0$, or $H(x + y) = 0$. This last equation says $x + y \in C$, and so $x \sim y$. From this relation between $x$ and $y$, we obtain $C + x = C + y$, since these are the equivalence classes of $x$ and $y$, and these classes are equal since $x$ and $y$ are related.

**Example 2.20.** Let

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}. $$

A short calculation shows that $C = \{0000, 1100, 0011, 1111\}$. The cosets of $C$ are then seen to be

$$C + 0000 = \{0000, 1100, 0011, 1111\},$$

$$C + 1000 = \{1000, 0100, 1011, 0111\},$$

$$C + 0010 = \{0010, 1110, 0001, 1101\},$$

$$C + 1010 = \{1010, 0110, 1001, 0101\}. $$

We also point out that $C = C + 0000 = C + 1100 = C + 0011 = C + 1111$; in other words, $C = C + v$ for any $v \in C$. Each coset in this example is equal to the coset of four vectors, namely the four vectors in the coset.
For some coding theory terminology, we call $Hx$ the *syndrome* of $x$. We can make use of syndromes to more quickly decode. Suppose that a word $w$ is received. If $c$ is the closest codeword to $w$, let $e = c + w$. Then $e$ is the *error word*, in that $e$ has a digit equal to 1 exactly when that digit was transmitted incorrectly in $c$. Note that $e$ is the word of smallest possible weight of the form $v + w$ with $v \in C$ since $\text{wt}(e) = D(c, w)$. If we can determine $e$, then we can determine $c$ by $c = e + w$. To see how to do this, if we take the equation $e = c + w$ and multiply both sides by $H$, we get

$$He = H(c + w) = Hc + Hw = 0 + Hw = Hw.$$  

We then can compute $He$ by computing $Hw$. The proposition says that $C + e = C + w$; in other words, $e \in C + w$. Since $c$ is the closest codeword to $w$, the word $e$ is then the word of least weight in the coset $C + w$. We then find $e$ by searching the words in $C + w$ and finding the word of least weight; such an element is called a *coset leader*. To do decoding with cosets, we compute, for each coset, the coset leader.

**Example 2.21.** Let

$$H = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$  

Then $C = \{00000, 11100, 00111, 11011\}$. We see that the distance of $C$ is 3, so $C$ is 1-error correcting. The cosets of $C$ are

- $\{00000, 00111, 11011, 11100\}$,
- $\{01110, 10010, 01001, 10101\}$,
- $\{00010, 00101, 11001, 11110\}$,
- $\{11111, 11000, 00011, 00100\}$,
- $\{01111, 01000, 10100, 10011\}$,
- $\{01101, 10110, 01010, 10001\}$,
- $\{01100, 10000, 10111, 01011\}$,
- $\{11010, 00001, 11101, 00110\}$.

By searching through each of the eight cosets, we can then build the following syndrome table:
Syndrome  |  Coset Leader
---|---
000  | 00000  
101  | 10010  
010  | 00010  
011  | 00100  
100  | 01000  
110  | 01010  
111  | 10000  
001  | 00001  

To see how we use the syndrome table to decode, we give an example. Suppose that \( w = 10010 \) is received. If we calculate \( Hw \), we get 101. First of all, since \( Hw \neq \mathbf{0} \), the vector \( w \) is not a codeword. By looking at the syndrome table, 101 is the second syndrome listed. The corresponding coset leader is \( e = 10010 \). We then decode \( w \) as \( c = w + e = 00000 \).

Similarly, if we receive the word \( w = 11111 \), we calculate \( Hw = 011 \). The corresponding coset leader is \( e = 00100 \), so the correct codeword is \( e + w = 11011 \). Using the syndrome table required much less computation than checking the distance between \( w \) and all 16 codewords to find that the closest codeword is 00000.

**Exercises**

1. Let \( C \) be the code consisting of all solutions of the matrix equation \( Ax = \mathbf{0} \), where

\[
A = \begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}.
\]

Calculate \( C \), determine its distance and error correcting capability. Construct the syndrome table for \( C \), and use the table to decode the vectors 10101101, 01011011 and 11000000.

2. List all of the cosets of the code \( C = \{00000, 11100, 00111, 11011\} \).

3. Find the cosets of the Hamming code.

4. Let \( C \) be the code consisting of solutions to \( Ax = \mathbf{0} \), where

\[
A = \begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 1
\end{pmatrix}.
\]

Build the syndrome table for \( C \). Determine the distance of \( C \). Use it to decode, if possible, 111110 and 100000. Feel free to use the Maple worksheet Cosets.mws.
2.5 The Golay Code

In this section we discuss an example of a code which is called the extended Golay code. This code is the set of solutions to the matrix equation $Hx = 0$, where $H$ is the $12 \times 24$ matrix whose left half is the $12 \times 12$ identity matrix and whose right half is

$$B = \begin{pmatrix}
1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\
1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0
\end{pmatrix}.$$ 

This length 24 code was used by the Voyager spacecraft to photograph Jupiter and Saturn between 1979 and 1981. The photographs were made using 4096 colors. Each color was encoded with a codeword from the Golay code. By solving the matrix equation $Hx = 0$, we can see that there are indeed 4096 codewords. Furthermore, a tedious check of all codewords shows that the distance of the Golay code is $d = 8$. Thus, the code can correct $[(8 - 1)/3] = 3$ errors. Thus, up to three out of the 24 digits of a codeword can be corrupted and still the original information will be retrievable.

Because this code can correct more than 1 error, any decoding procedure is bound to be more complicated than that for the Hamming code. We give here a decoding procedure. It is based on some simple facts about the matrix $B$. Its validity is left to a series of homework problems.

To make it more convenient to work with this code, we write a word $u = (u_1, u_2)$, where $u_1$ consists of the first 12 digits and $u_2$ the remaining 12. Since $H = [I : B]$, we see that $u \in C$ if and only if $Hu = 0$, which is true if and only if $u_1 + Bu_2 = 0$. If we have a received word $w$, we perform the following steps to decode $w$. We write $v$ for the codeword to be determined from $w$. Finally, we write $e_i$ for the 12-tuple with $i$-entry 1 and all other entries 0, and $b_i$ for the $i$-th row of $b_i$.

1. Compute $s = Hw$. If $s = 0$, then $w$ is a codeword.
2. If $1 \leq \text{wt}(s) \leq 3$, then $v = w + (s, 0)$.
3. If $\text{wt}(s) > 3$ and $\text{wt}(s + b_i) \leq 2$ for some $i$, then $v = w + (s + b_i, e_i)$.
4. If we haven’t yet determined $v$, then compute $Bs$. 

5. If $1 \leq \text{wt}(Bs) \leq 3$, then $v = w + (0, Bs)$.

6. If $\text{wt}(Bs) > 3$ and $\text{wt}(Bs + b_i) \leq 2$ for some $i$, then $v = w + (e_i, Bs + b_i)$.

7. If we haven’t determined $v$, then $w$ cannot be decoded.

**Example 2.22.** Suppose that $w = 0010010011011010010$ is received. We calculate $s = Hw$, and we get $s = 11000001001$. We see that $\text{wt}(s) = 5$. We see that $\text{wt}(s + b_5) = 2$. Therefore, by Step 3, we decode $w$ as $v = w + (s + b_5, e_5) = w + (0000001010010, 00001000000) = 001001011111010101010$.

**Exercises**

For these problems, we verify some of the theoretical facts needed to see that the decoding procedure for the Golay code is valid. We use the following setup: $C$ is the Golay code, $H$ is the $12 \times 24$ matrix $[I \mid B]$ mentioned in the text, $w$ is a received word, $s = Hw$. We write some 24-tuples as $(u_1, u_2)$ with $u_i$ a 12-tuple. We denote the $i$-th row (and column) of $B$ by $b_i$, and $e_i$ denotes the 12-tuple whose $i$-th component is 1 and whose other components are 0. Let $v$ be the closest codeword to $w$, write $v = w + e$, and suppose that $\text{wt}(e) \leq 3$.

1. Suppose that $e = (u, 0)$; with $\text{wt}(u) \leq 3$. Show that $s = u$, and conclude that $v = w + (s, 0)$.

2. Suppose that $e = (u, e_i)$ with $\text{wt}(u) \leq 2$. Show that $s = u + b_i$. Conclude that $\text{wt}(s) > 3$ and $\text{wt}(s + b_i) \leq 2$, and that $v = w + (s + b_i, e_i)$.

3. Suppose that $e = (0, u)$ with $\text{wt}(u) \leq 3$. Show that $s$ is the sum of at most three of the $b_i$ and that $u = Bs$. Conclude that $\text{wt}(s) > 3$ but $\text{wt}(Bs) \leq 3$, and that $v = w + (0, Bs)$.

4. Suppose that $e = (e_i, u)$ with $\text{wt}(u) \leq 2$. Show that $s = e_i + Bu$, and that $Bs = b_i + u$. Conclude that $\text{wt}(s) > 3$, $\text{wt}(s + b_i) > 2$ for any $i$, and that $e = (e_i, Bs + b_i)$, so $v = w + (e_i, Bs + b_i)$.

These four problems show how, given any possibility of an error vector $e$ having weight at most 3, how we can determine it in terms of the syndrome $s$. Reading these four problems backwards yields the decoding procedure we gave in class.
Chapter 3

Rings and Fields

We are all long familiar with the natural, rational, real, and complex number systems. Moreover, in the previous chapter we introduced the set of integers modulo \( n \), and we defined addition, subtraction, and multiplication on it. In high school algebra you worked with polynomials, and saw how to add, subtract, and multiply them. In a linear algebra course you saw that arithmetic operations can be defined on matrices, and that there are structures called vector spaces in which you can add and subtract vectors and perform scalar multiplication. In calculus, if not earlier, you saw that functions can be combined by adding, subtracting, multiplying, and dividing, along with composing them. There are many similarities between these different systems. The idea of abstract algebra is to distill down the ideas and properties common to these systems, and then study any system with the same properties. There are several different algebraic systems that we will study. The first system we will look at generalizes the integers and the integers modulo \( n \). If you look back at the list of properties that the modular operations satisfy, you will notice that analogues of these properties hold in all the ordinary number systems. Also, these properties, except for commutativity of multiplication, hold for matrix operations. These properties occur so often that it is worth studying systems that satisfy them.

3.1 The Definition of a Ring

In order to discuss algebraic systems, we first need to formalize the meaning of an operation. When you add integers you start with two integers and the process of adding them returns to you a single integer. Therefore, addition is a function that takes as input a pair of integers and gives as output a single integer. Multiplication can be viewed similarly. These are special examples of binary operations. Recall that the Cartesian product \( A \times B \) of two sets \( A \) and \( B \) is the set of all pairs \((a, b)\) with \( a \in A \) and \( b \in B \). In other words,

\[
A \times B = \{(a, b) : a \in A, b \in B\}.
\]

Definition 3.1. If \( S \) is a set, then a binary operation on \( S \) is a function from \( S \times S \) to \( S \).
Example 3.2. The operations of addition, subtraction, and multiplication on the various number systems are all examples of binary operations. Division is not an example. Division on the real numbers is not a function from \(\mathbb{R} \times \mathbb{R}\) to \(\mathbb{R}\) but a function from \(\mathbb{R} \times (\mathbb{R} - \{0\})\) to \(\mathbb{R}\). In other words, since we cannot divide by 0, division is not defined on pairs of the form \((a, 0)\).

Example 3.3. If \(T\) is any set, then union \(\cup\) and intersection \(\cap\) are examples of binary operations on the set of all subsets of \(T\). That is, given any two subsets of \(T\), the union and intersection of the sets is again a subset of \(T\).

Example 3.4. Here is an example from multivariable calculus. Consider the set \(\mathbb{R}^3\) of 3-tuples of real numbers. The cross product is an operation on \(\mathbb{R}^3\). Recall that the cross product is given by the formula

\[
(a_1, a_2, a_3) \times (b_1, b_2, b_3) = \det \begin{pmatrix}
i & j & k \\
a_1 & a_2 & a_3 \\
b_1 & b_2 & b_3
\end{pmatrix} = (a_2b_3 - a_3b_2)i + (a_3b_1 - a_1b_3)j + (a_1b_2 + a_2b_1)k;
\]

Here \(i\), \(j\), and \(k\) are the usual unit vectors \((1, 0, 0)\), \((0, 1, 0)\), and \((0, 0, 1)\). The reason for giving this example here is that this is an example of a non-associative operation. For instance, \((i \times i) \times j = 0 \times j = 0\) while \(i \times (i \times j) = i \times k = -j\). Thus, the associative property is not something that holds in every reasonable example.

Merely having an operation on a set is not, in general, a very useful thing. The operation needs to satisfy some appropriate properties in order to be useful. We know various examples of useful properties, such as the commutative and associative laws. Now that we have the notion of a binary operation on a set, we can give the definition of a ring. This is the structure that generalizes the examples mentioned above.

Definition 3.5. A **ring** is a nonempty set \(R\) together with two binary operations \(+\) and \(\cdot\) such that, for all \(a, b, c \in R\), the following properties hold

- **commutativity of addition**: \(a + b = b + a\);
- **associativity of addition**: \(a + (b + c) = (a + b) + c\);
- **existence of an additive identity**: there is an element \(0 \in R\) with \(a + 0 = a\) for every \(a \in R\);
- **existence of additive inverses**: for each \(a \in R\) there is an element \(s \in R\) with \(a + s = 0\);
- **associativity of multiplication**: \(a \cdot (b \cdot c) = (a \cdot b) \cdot c\);
- **left distributivity**: \(a \cdot (b + c) = a \cdot b + a \cdot c\);
3.1. **THE DEFINITION OF A RING**

- **right distributivity:** \((b + c) \cdot a = b \cdot a + c \cdot a;\)

- **existence of a multiplicative identity:** there is an element \(1 \in R\) with \(1 \cdot a = a \cdot 1 = a\) for all \(a \in R\).

Before we give examples, we point out a couple things about this definition. First of all, we did not include commutativity of multiplication in this list. If a ring \(R\) satisfies \(a \cdot b = b \cdot a\) for all \(a, b \in R\), then we call \(R\) a **commutative ring**. If a ring is commutative, then the two distributivity laws reduce down to the same thing, and the requirement that \(1 \cdot a = a \cdot 1\) in the last property is redundant. Second, we do not say anything about the existence of multiplicative inverses. In our example of \(\mathbb{Z}_n\), we have seen that elements need not have multiplicative inverses. The existence of multiplicative inverses is an important question that we will address in several examples.

**Example 3.6.** The set \(\mathbb{Z}\) of integers forms a ring under the usual addition and multiplication operations. The defining properties of a ring are well known to hold for \(\mathbb{Z}\). The set \(\mathbb{Q}\) of rational numbers also forms a ring under the usual operations. So does the set \(\mathbb{R}\) of real numbers, and the set \(\mathbb{C}\) of complex numbers. All four are commutative rings.

**Example 3.7.** The set \(M_n(\mathbb{R})\) of \(n \times n\) matrices with real number entries under matrix addition and multiplication forms a ring. It is proved in any linear algebra course that the matrix operations satisfy the properties above. For instance, the zero matrix in \(M_2(\mathbb{R})\)

\[
\begin{pmatrix}
0 & 0 \\
0 & 0
\end{pmatrix}
\]

is the additive identity, and the identity matrix

\[
\begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix}
\]

is the multiplicative identity. This ring is not commutative. For example,

\[
\begin{pmatrix}
1 & 0 \\
1 & 1
\end{pmatrix}
\begin{pmatrix}
0 & 1 \\
2 & 1
\end{pmatrix}
= \begin{pmatrix}
0 & 1 \\
2 & 2
\end{pmatrix}
\]

while

\[
\begin{pmatrix}
0 & 1 \\
2 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 0 \\
1 & 1
\end{pmatrix}
= \begin{pmatrix}
1 & 1 \\
3 & 1
\end{pmatrix},
\]

so the order of multiplication matters.

**Example 3.8.** If \(n\) is a positive integer, then the set \(\mathbb{Z}_n\) of integers modulo \(n\) forms a ring. We saw that the ring properties hold for \(\mathbb{Z}_n\) in the previous chapter. We also saw that multiplication is commutative, so \(\mathbb{Z}_n\) is a commutative ring.
Example 3.9. Let $R$ be the set of all continuous (real valued) functions defined on the interval $[0, 1]$. Recall that these function operations are defined pointwise. That is, if $f$ and $g$ are functions, then $f + g$, $f - g$, and $fg$ are defined by

\[
(f + g)(x) = f(x) + g(x),
\]
\[
(f - g)(x) = f(x) - g(x),
\]
\[
(fg)(x) = f(x)g(x).
\]

In calculus one shows that the sum, difference, and product of continuous functions is again continuous. Thus, we have operations of addition and multiplication on the set $R$. It is possible that the ring properties are verified in a calculus class. We do not verify them all here, but discuss some. The additive identity of $R$ is the zero function $z$, which is defined by $z(x) = 0$ for all $x \in [0, 1]$. The multiplicative identity of $R$ is the constant function $1$ defined by $1(x) = 1$ for all $x \in [0, 1]$. Commutativity of addition holds because if $f, g \in R$ and $x \in [0, 1]$, then

\[
(f + g)(x) = f(x) + g(x) \\
= g(x) + f(x) \\
= (g + f)(x).
\]

Thus, the functions $f + g$ and $g + f$ agree at every value in the domain, which means, by definition, that $f + g = g + f$. To prove this we used the definition of function addition and commutativity of addition of real numbers. All of the other ring properties follow from the definition of the operations and appropriate ring properties of $\mathbb{R}$. Furthermore, multiplication of function multiplication is commutative, so the ring of continuous functions on $[0, 1]$ is commutative.

Example 3.10. Let $\mathbb{R}[x]$ be the set of all polynomials with real number coefficients. We have seen in high school how to add and multiply polynomials; we recall how these operations are defined primarily to give a simple notation for denoting them. First note that given two polynomials $f = \sum_{i=1}^{n} a_i x^i$ and $g = \sum_{i=1}^{m} b_i x^i$ we can add on terms with zero coefficient and assume that $m = n$. With this in mind, we have

\[
\sum_{i=1}^{n} a_i x^i + \sum_{i=1}^{n} b_i x^i = \sum_{i=1}^{n} (a_i + b_i) x^i,
\]
\[
\sum_{i=0}^{n} a_i x^i \cdot \sum_{i=0}^{m} b_i x^i = \sum_{j=0}^{n} \left( \sum_{i=0}^{m} a_i b_j \right) x^{i+j} \\
= \sum_{j=0}^{n+m} \left( \sum_{i=0}^{j} a_i b_{j-i} \right) x^j.
\]

It is likely that in some high school algebra course you saw that all of the ring properties hold for $\mathbb{R}[x]$. Multiplication of polynomials is commutative, so $\mathbb{R}[x]$ is a commutative ring.
Example 3.11. Let \( T \) be a set, and let \( R \) be the set of all subsets of \( T \). We have two binary operations on \( R \), namely union and intersection. However, \( R \) does not form a ring under these two operations. The identity element for union is the empty set \( \emptyset \) since \( A \cup \emptyset = \emptyset \cup A = A \) for any \( A \in T \). Similarly \( R \) serves as the identity element for intersection. On the other hand one can check that neither operation allows for inverses: given a nonempty subset \( A \) of \( R \), what subset \( B \) can satisfy \( A \cup B = \emptyset \)? And if \( A \) is a proper subset of \( T \) what subset \( B \) can satisfy \( A \cap B = R \)? Neither operation can be considered as addition. However, we can introduce a new operations to come up with a ring. Define addition and multiplication on \( T \) by
\[
A + B = (A - B) \cup (B - A) = (A \cup B) - (A \cap B),
\]
\[
A \cdot B = A \cap B.
\]
Then \( R \), together with these operations, forms a commutative ring. We leave the details of the proof to an exercise. We do point out some interesting properties of this ring. First of all, \( 0 = \emptyset \) for this ring since \( A + \emptyset = (A \cup \emptyset) - (A \cap \emptyset) = A - \emptyset = A \) for any subset \( A \) of \( T \). Also, \( 1 = T \) since \( A \cdot T = A \cap T = A \) for any subset \( A \). Next, for any \( A \) we have \( A + A = (A \cup A) - (A \cap A) = A - A = \emptyset \). Therefore, \( -A = A \) for any \( A \). This would seem to be a very unusual property, although we will see it all the time when we deal with coding theory. Finally, note that \( A \cdot A = A \cap A = A \). This example is important in the study of set theory and logic.

Example 3.12. In this example we describe a method of constructing a new ring from two existing rings. Let \( R \) and \( S \) be rings. Define operations \( + \) and \( \cdot \) on the Cartesian product \( R \times S \) by
\[
(r, s) + (r', s') = (r + r', s + s'),
\]
\[
(r, s) \cdot (r', s') = (rr', ss')
\]
for all \( r, r' \in R \) and \( s, s' \in S \), where \( r + r' \) and \( s + s' \) are calculated in the rings \( R \) and \( S \) respectively, and similarly for \( rr' \) and \( ss' \). Then one can verify that \( R \times S \), together with these operations, is indeed a ring. If \( 0_R \) and \( 0_S \) are the additive identities of \( R \) and \( S \) respectively, then \( (0_R, 0_S) \) is the additive identity for \( R \times S \). Similarly, if \( 1_R \) and \( 1_S \) are the multiplicative identities of \( R \) and \( S \) respectively, then \( (1_R, 1_S) \) is the multiplicative identity of \( R \times S \).

3.2 First Properties of Rings

There are some properties with which you are very familiar for the ordinary number systems that hold for any ring. First of all, the additive and multiplicative identities are unique. We leave one of these facts for homework, and prove one now.

Lemma 3.13. Let \( R \) be a ring. Then the additive identity is unique.
Proof. Suppose there are elements 0 and 0' in R that satisfy \( a + 0 = a \) and \( a + 0' = a \) for all \( a \in R \). Recalling that addition is commutative, we have \( a + 0 = 0 + a = a \) and \( a + 0' = 0' + a = a \) for any \( a \). If we use the first pair of equations for \( a = 0' \), we get \( 0' + 0 = 0' \). On the other hand, if we use the second pair with \( a = 0 \), we get \( 0' + 0 = 0 \). Thus, \( 0 = 0' \). This shows that there can be only one additive identity for \( R \).

We next remark that the additive inverse of any element is unique, which is left to a homework exercise. Since the additive inverse of an element is unique, we give a notation for it, using the familiar notation. Thus, \(-a\) represents the additive inverse of \( a \). We can use additive inverses to define subtraction. We do this in the usual manner: \( a - b = a + (-b) \).

Lemma 3.14. Let \( R \) be a ring, and let \( a, b \in R \).

1. \(-(-a) = a\);
2. \(- (a + b) = -a + (-b)\);
3. \(- (a - b) = -a + b\).

Proof. Let \( a \in R \). By definition of \(-a\), we have \( a + (-a) = (-a) + a = 0 \). This equation tells us that \( a \) is the additive inverse of \(-a\). That is, \( a = -(-a) \), which proves the first property.

For the second property, let \( a, b \in R \). Then \(- (a + b)\) is the additive inverse of \( a + b \). We show \(-a + (-b)\) is also the additive inverse of \( a + b \). To do this, we calculate

\[
(a + b) + (-a + (-b)) = ((a + b) + (-a)) + (-b) \\
= (a + (b + (-a)) + (-b) \\
= (a + ((-a) + b)) + (-b) \\
= ((a + (-a)) + b) + (-b) \\
= (0 + b) + (-b) \\
= b + (-b) \\
= 0.
\]

Each of these steps followed from one of the properties in the definition of a ring. Because we have shown that \( (a + b) + (-a + (-b)) = 0 \), the element \(-a + (-b)\) is the additive inverse of \( a + b \). This tells us that \(-a + (-b) = -(a + b)\). Finally, for the third property, again let \( a, b \in R \). We use a similar argument as in part 2, leaving out a few steps which you are encouraged to fill in. We have

\[
(a - b) + (-a + b) = (a + (b)) + (-a + b) \\
= (a + (-b) + (-a)) + b \\
= (a + (-a) + (-b)) + b \\
= (0 + (-b)) + b \\
= (-b) + b \\
= 0.
\]
3.2. FIRST PROPERTIES OF RINGS

Therefore, \(-a + b = -(a - b)\).

**Proposition 3.15** (Cancellation Law of Addition). Let \(a, b, c \in R\). If \(a + b = a + c\), then \(b = c\).

*Proof.* Let \(a, b, c \in R\), and suppose that \(a + b = a + c\). By adding \(-a\) to both sides, we get the following string of equalities.

\[
\begin{align*}
    a + b &= a + c \\
    -a + (a + b) &= -a + (a + c) \\
    (-a + a) + b &= (-a + a) + c \\
    0 + b &= 0 + c \\
    b &= c
\end{align*}
\]

Again, notice that in each line we used properties from the definition of a ring.

The element 0 is defined with respect to addition, but we want to know how it behaves with respect to multiplication. The distributive properties are the only properties in the definition of a ring that involve both operations. They will then be necessary to use to prove anything that relates the two operations. Such an example is the following lemma, that proves for any ring the familiar property of how 0 multiplies in the usual number systems.

**Lemma 3.16.** Let \(R\) be a ring. If \(a \in R\), then \(a \cdot 0 = 0 \cdot a = 0\).

*Proof.* Let \(a \in R\). To prove that \(a \cdot 0 = 0\), we use cancellation together with the definition of an additive identity. We have

\[
\begin{align*}
    a \cdot 0 + a &= a \cdot 0 + a \cdot 1 = a \cdot (0 + 1) \\
    &= a \cdot 1 \\
    &= a \\
    &= 0 + a.
\end{align*}
\]

Canceling \(a\) yields \(a \cdot 0 = 0\). A similar argument will show that \(0 \cdot a = 0\).

Let us now discuss the question of existence of multiplicative inverses. There is a concept related to this that is a part of high school mathematics that is important but often gets too little attention. We bring it up with an example. Suppose you wish to solve the equation \(x^2 - 5x - 6 = 0\) by factoring. Since \(x^2 - 5x - 6 = (x + 1)(x - 6)\), our equation is equivalent to \((x + 1)(x - 6) = 0\). Now, assuming we are dealing with real numbers, the product of two nonzero real numbers is again nonzero. Therefore, we must have \(x + 1 = 0\) or \(x - 6 = 0\), thus getting the two solutions \(x = -1\) and \(x = 6\). What makes factoring a useful technique in solving equations is this property of the real numbers that we just used. However, we
have seen that \( \mathbb{Z}_n \) does not always share this property. For example, in \( \mathbb{Z}_6 \) we have \( \overline{2} \cdot \overline{3} = \overline{0} \). Likewise, this bad behavior occurs with matrices. For example,

\[
\begin{pmatrix}
1 & 0 \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
0 & 0 \\
0 & 1
\end{pmatrix} =
\begin{pmatrix}
0 & 0 \\
0 & 0
\end{pmatrix}.
\]

If we are trying to solve equations in a ring, then it may be crucial to know whether or not we can have the product of nonzero elements equal to zero.

**Definition 3.17.** If \( R \) is a ring, then a nonzero element \( a \in R \) is said to be a zero divisor if there is a nonzero element \( b \in R \) with \( a \cdot b = 0 \).

A useful alternative phrasing of the definition is the contrapositive. A nonzero element \( a \in R \) is not a zero divisor if whenever \( b \in R \) with \( a \cdot b = 0 \), then \( b = 0 \). Terminology about zero divisors varies from textbook to textbook. Some books only define zero divisors for commutative rings. Some books consider 0 to be a zero divisor and others do not. Others talk about left and right zero divisors. If \( a \cdot b = 0 \) with both \( a \) and \( b \) nonzero, one could call \( a \) a left zero divisor and \( b \) a right zero divisor, but we will not worry about such things. The name zero divisor comes from the usual meaning of divisor in \( \mathbb{Z} \). If \( c \) and \( d \) are integers, then \( c \) is called a divisor of \( d \) if there is an integer \( e \) with \( ce = d \). If \( ce = 0 \), then this terminology would lead us to say that \( c \) is a divisor of 0. However, since \( c \cdot 0 = 0 \) for all \( c \), this would seem to lead us to call every integer a divisor of zero. This is not a useful statement. The restriction in the definition above to require \( b \neq 0 \) in order to call \( a \) a zero divisor if \( ab = 0 \) eliminates this worry.

**Example 3.18.** The rings \( \mathbb{Z} \), \( \mathbb{Q} \), \( \mathbb{R} \), and \( \mathbb{C} \) have no zero divisors. Each of these rings has the familiar property that \( ab = 0 \) implies \( a = 0 \) or \( b = 0 \).

**Example 3.19.** The ring of \( 2 \times 2 \) matrices with real number entries has zero divisors, as the example above shows. In fact, if

\[
A = \begin{pmatrix}
a & b \\
c & d
\end{pmatrix},
\]

then the adjoint of \( A \) is the matrix

\[
\text{adj}(A) = \begin{pmatrix}
d & -b \\
-c & a
\end{pmatrix},
\]

and from a simple calculation or a recollection from linear algebra,

\[
A \cdot \text{adj}(A) = \det(A) \begin{pmatrix}
1 & 0 \\
0 & 1
\end{pmatrix} = \begin{pmatrix}
ad - bc & 0 \\
0 & ad - bc
\end{pmatrix}.
\]

Thus, if \( \det(A) = 0 \), then \( A \) is a zero divisor.
Example 3.20. We have seen that \( \mathbb{Z}_6 \) has zero divisors; for example, \( 2 \cdot 3 = 0 \). Similarly, \( \mathbb{Z}_9 \) has zero divisors since \( 3 \cdot 3 = 0 \). Also, \( \mathbb{Z}_{12} \) has zero divisors since \( 6 \cdot 4 = 24 = 0 \). However, \( \mathbb{Z}_5 \) has no zero divisors; if we view the multiplication table for \( \mathbb{Z}_5 \), we see that the product of two nonzero elements is always nonzero.

\[
\begin{array}{c|cccc}
\cdot \mod 5 & 0 & 1 & 2 & 3 \\
0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 2 & 3 \\
2 & 0 & 2 & 4 & 1 \\
3 & 0 & 3 & 1 & 4 \\
4 & 0 & 4 & 3 & 2 \\
\end{array}
\]

Example 3.21. Let \( R \) and \( S \) be rings, and consider the ring \( R \times S \) with operations defined earlier. Then this ring has zero divisors, as long as both \( R \) and \( S \) contain nonzero elements. For, if \( r \in R \) and \( s \in S \), then \( (r, 0) \cdot (0, s) = (0, 0) \). For example, if \( R = S = \mathbb{Z} \), then elements of the form \( (n, 0) \) or \( (0, m) \) with \( n, m \) nonzero are zero divisors in \( \mathbb{Z} \times \mathbb{Z} \).

There is a cancellation law of multiplication for the ordinary number systems. Because of the existence of zero divisors, the generalization to arbitrary rings is more complicated.

Proposition 3.22 (Cancellation Law of Multiplication). Let \( R \) be a ring. Suppose that \( a, b, c \in R \) with \( ab = ac \) and \( a \neq 0 \). If \( a \) is not a zero divisor, then \( b = c \).

Proof. Let \( a, b, c \in R \) with \( ab = ac \). Suppose that \( a \) is not a zero divisor. From the equation we get \( ab - ac = 0 \), so \( a(b - c) = 0 \). Since \( a \) is not a zero divisor, \( b - c = 0 \), or \( b = c \). \(\square\)

The cancellation law fails if \( a \) is a zero divisor. For example, in \( \mathbb{Z}_6 \) we have \( \overline{2} \cdot \overline{4} = \overline{2} \cdot \overline{1} \) even though \( \overline{4} \neq \overline{1} \). Therefore, we cannot cancel \( \overline{2} \) in such an equation. Similarly, with matrices we have the equation

\[
\begin{pmatrix}
1 & 0 \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
2 & 0 \\
-3 & 4
\end{pmatrix}
= \begin{pmatrix}
1 & 0 \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
2 & 0 \\
-3 & 6
\end{pmatrix}
\]

while

\[
\begin{pmatrix}
2 & 0 \\
-3 & 4
\end{pmatrix}
\neq \begin{pmatrix}
2 & 0 \\
-3 & 6
\end{pmatrix}
\]

More generally, in any ring, if \( ab = 0 \) with \( a, b \neq 0 \), then \( a \cdot b = a \cdot 0 \), so cancellation of the zero divisor \( a \) is not valid.

We now return to the idea of multiplicative inverses. Consider \( \mathbb{Z}_6 \). The multiplication table for \( \mathbb{Z}_6 \) is
From this table we see that \(2, 3, \) and \(4\) are zero divisors. The remaining nonzero elements \(1\) and \(5\) are not. In fact, \(1\) and \(5\) each have a multiplicative inverse, namely each itself, since \(1 \cdot 1 = 1\) and \(5 \cdot 5 = 1\).

**Definition 3.23.** If \(R\) is a ring, then an element \(a \in R\) is said to be a unit if there is an element \(b \in R\) with \(a \cdot b = b \cdot a = 1\). In this case \(a\) and \(b\) are said to be multiplicative inverses.

A unit is never a zero divisor, as we now prove.

**Proposition 3.24.** Let \(R\) be a ring. If \(a \in R\) is a unit, then \(a\) is not a zero divisor.

**Proof.** Let \(a \in R\) be a unit with multiplicative inverse \(c \in R\). Then \(ac = ca = 1\). Suppose that \(b \in R\) with \(ab = 0\). Then

\[
0 = c \cdot 0 = c(ab) = (ca)b = 1 \cdot b = b.
\]

We have shown that if \(ab = 0\), then \(b = 0\). Thus, \(a\) is not a zero divisor. An obvious modification of the argument just given shows that if \(ba = 0\), then \(b = 0\) as well. \(\square\)

**Exercises**

1. Find all solutions in \(\mathbb{Z}_8\) to the equation \(x^2 - 1 = 0\).

2. Find all solutions in \(\mathbb{Z}_{60}\) to the equation \(x^2 - 1 = 0\). How many did you find? Do you recall any fact about how many roots a polynomial of degree \(n\) can have?

3. Let

\[
A = \begin{pmatrix}
0 & -1 \\
1 & -1
\end{pmatrix}.
\]

Show that \(A\) satisfies the equation \(X^3 = I\). Moreover, if \(B\) is any invertible \(2 \times 2\) matrix, prove that \(BAB^{-1}\) also satisfies the equation \(X^3 - I = 0\). Finally, verify that

\[
\begin{pmatrix}
1 & b \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
0 & -1 \\
1 & -1
\end{pmatrix}
\begin{pmatrix}
1 & b \\
0 & 1
\end{pmatrix}^{-1} = \begin{pmatrix}
b & -b^2 - 1 - b \\
1 & -1 - b
\end{pmatrix},
\]

and conclude that there are infinitely many real-valued \(2 \times 2\) matrices satisfying the equation \(X^3 - I = 0\).
4. Let $R$ be a commutative ring in which there are no zero divisors. Prove that the only solutions to $x^2 - 1 = 0$ in $R$ are 1 and $-1$.
   (Hint: factor the left hand side of the equation.)

5. Let $R$ and $S$ be rings. Referring to Example 3.12, prove that $R \times S$ satisfies the distributive properties.

6. If $R$ is a ring, prove that $a(-b) = (-a)b = -(ab)$ for all $a, b \in R$.

7. If $R$ is a ring, prove that $a(b - c) = ab - ac$ for all $a, b, c \in R$.

8. Prove that the set $R$ of Example 3.11 with the operations
   
   
   $A + B = (A - B) \cup (B - A) = (A \cup B) - (A \cap B),$
   
   $A \cdot B = A \cap B.$

   is a ring.

9. A unit of a ring $R$ is an element $a$ such that the multiplicative inverse $a^{-1}$ of $a$ exists. Prove that if $a$ and $b$ are units, then $ab$ is a unit.
   (Hint: a fact from linear algebra about invertible matrices may stimulate your thinking.)

10. Prove that $\pi \in \mathbb{Z}_n$ is a zero divisor if and only if $\gcd(a, n) > 1$.

11. Prove that a matrix $^1 A \in M_n(\mathbb{R})$ is a zero divisor if and only if $\det(A) = 0$. You may wish to recall something about the formula for the inverse of a matrix to help you do this.

12. Prove that the zero divisors of $\mathbb{Z} \times \mathbb{Z}$ are the elements of the form $(a, 0)$ or $(0, a)$ for any $a \in \mathbb{Z}$.

13. Give an example of a ring $R$ and an element $a \in R$ that is neither a zero divisor nor a unit.

14. Let $R$ be a ring in which $0 = 1$. Prove that $R = \{0\}$.

15. Prove that the multiplicative identity of a ring is unique.

16. If $a$ and $b$ are elements of a ring $R$, prove that $(-a) \cdot (-b) = ab$.

17. Prove that if $a$ and $b$ are elements of a ring, then $(a + b)^2 = a^2 + ab + ba + b^2$.

18. Let $R$ be a ring in which $a^2 = a$ for all $a \in R$. Prove that $-a = a$. Then prove that $R$ is commutative.

$^1$define notation in section
19. Let $R$ be a ring. If $a \in R$, define powers of $a$ inductively by $a^0 = 1$ and $a^{n+1} = a^n \cdot a$.

(a) If $n$ and $m$ are positive integers, prove that $a^n \cdot a^m = a^{n+m}$.

(b) Prove or disprove: if $a, b \in R$, then $(ab)^n = a^n \cdot b^n$. If this is not always true, determine under what circumstances it is true.

20. Define a relation between elements of a commutative ring $R$ by $a \sim b$ if there is a unit $u \in R$ with $b = au$. Prove that this relation is an equivalence relation. Determine the equivalence classes of this relation if $R = \mathbb{Z}$ and if $R = \mathbb{Z}_{12}$.

21. Is subtraction on $\mathbb{Z}$ an associative operation? Is it commutative? Why or why not?

22. Let $R$ be a ring. Suppose that $S$ is a subset of $R$ containing 1 satisfying (i) if $a, b \in S$, then $a - b \in S$, and (ii) if $a, b \in S$, then $ab \in S$.

(a) Show that $0 \in S$. Then show that if $a \in S$, then $-a \in S$.

(b) Show that if $a, b \in S$, then $a + b \in S$. From this and (ii), we can view $+$ and $\cdot$ as operations on $S$.

(c) Conclude that all of the ring axioms hold for $S$ from the ring axioms for $R$, from the hypotheses, and from (a). The ring $S$ is said to be a subring of $R$.

23. Let $\mathbb{H}$ be the set of all $2 \times 2$ matrices with complex number entries of the form \[
\begin{pmatrix}
a & -b \\
b & a
\end{pmatrix}.
\] Prove that $\mathbb{H}$ satisfies the hypotheses of the previous problem, with $R = M_2(\mathbb{C})$, and so that we can consider $\mathbb{H}$ to be a subring of $M_2(\mathbb{C})$. Moreover, prove that $\mathbb{H}$ is noncommutative, and that every nonzero element of $\mathbb{H}$ has a multiplicative inverse in $\mathbb{H}$.

(The ring $\mathbb{H}$, discovered in 1843 by Hamilton, was the first example discovered of a noncommutative ring in which every nonzero element has a multiplicative inverse. This matrix theoretic description is not Hamilton’s original description. The ring $\mathbb{H}$ is now called the ring of Hamilton’s quaternions.)

24. Let $\mathbb{H}$ be the ring of the previous problem. Let
\[
I = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, K = IJ = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix}.
\]

We write 1 for the identity matrix. Prove that $I^2 = J^2 = K^2 = -1$ and $K = -JI$. Also, note that if $a = \alpha + \beta i$ and $b = \gamma + \delta i$, then
\[
\begin{pmatrix} a & -b \\ b & a \end{pmatrix} = \alpha \cdot 1 + \beta I + \gamma J + \delta K,
\]
if we view $\alpha$ as $\alpha$ times the identity matrix. Finally, if $\Delta = \alpha^2 + \beta^2 + \gamma^2 + \delta^2$, show that

$$(\alpha + \beta I + \gamma J + \delta K)^{-1} = \left(\frac{\alpha}{\Delta}\right) \cdot 1 - \left(\frac{\beta}{\Delta}\right) I - \left(\frac{\gamma}{\Delta}\right) J - \left(\frac{\delta}{\Delta}\right) K.$$ 

(Hamilton defined the quaternions as all symbols of the form $a + bi + cj + dk$ with $a, b, c, d \in \mathbb{R}$ with addition given by

- $(a + bi + cj + dk) + (a' + b'i + c'j + d'k) = (a + a') + (b + b')i + (c + c')j + (d + d')k + bi + cj + dk$

and multiplication obtain by using the distributive law along with the identities $i^2 = j^2 = k^2 = -1$ and $k = ij = -ji$.)

3.3 Fields

In the previous section we discussed the notion of a zero divisor and saw that if an element has a multiplicative inverse, then it is not a zero divisor. The ring $\mathbb{Z}$ has no zero divisors. However, few integers have multiplicative inverses in $\mathbb{Z}$; in fact, only 1 and $-1$ have multiplicative inverses in $\mathbb{Z}$. However, every nonzero integer has a multiplicative inverse in the ring $\mathbb{Q}$. Even more than that, every nonzero element of $\mathbb{Q}$ has a multiplicative inverse in $\mathbb{Q}$. A similar statement holds for $\mathbb{R}$ and for $\mathbb{C}$. Therefore, in $\mathbb{Q}$ or $\mathbb{R}$, we may divide, as long as we do not try to divide by 0. However, if we restrict ourselves to $\mathbb{Z}$, we cannot always divide; we must work in $\mathbb{Q}$ in order to do general division of integers. Being able to divide is naturally useful. Rings in which we can divide are important enough and common enough for us to investigate.

**Definition 3.25.** A field is a commutative ring $F$ containing at least two elements such that every nonzero element of $F$ has a multiplicative inverse in $F$.

The hypothesis that a field contains at least two elements ensures that $0 \neq 1$; necessarily $0 = 1$ if the field has only one element. If it has more than one, a short argument will prove that $0 \neq 1$. This assumption is not important, but it is convenient. It is similar to the assumption that 1 is not a prime number, even though it cannot be factored in a nontrivial way.

**Example 3.26.** The number systems $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all examples of a field.

The following proposition gives us infinitely many examples of fields.

**Proposition 3.27.** Let $n$ be a positive integer. Then $\mathbb{Z}_n$ is a field if and only if $n$ is prime.
Proof. Let \( n \) be a positive integer. We know that \( \mathbb{Z}_n \) is a commutative ring, so it suffices to prove that every nonzero element of \( \mathbb{Z}_n \) has a multiplicative inverse if and only if \( n \) is prime. First, suppose that \( n \) is prime. Let \( \overline{a} \in \mathbb{Z}_n \) be nonzero. Then \( a \not\equiv 0 \mod n \). Consequently, \( n \) does not divide \( a \). Since \( n \) is prime, this forces \( \gcd(a, n) = 1 \). Thus, by Proposition 1.12, there are integers \( s \) and \( t \) with \( 1 = as + nt \). This means \( as \equiv 1 \mod n \), or \( \overline{a} \cdot \overline{s} = \overline{1} \) in \( \mathbb{Z}_n \). We have thus produced a multiplicative inverse for \( \overline{a} \). Since \( \overline{a} \) was an arbitrary nonzero element, this proves that \( \mathbb{Z}_n \) is a field.

For the converse, consider its contrapositive: if \( n \) is not prime, then \( \mathbb{Z}_n \) is not a field. we prove this statement. Assume that \( n \) is composite and factors as \( n = ab \) with \( 1 < a, b < n \). Then \( \overline{a} \cdot \overline{b} = \overline{ab} = \overline{a} = \overline{0} \). Since \( \overline{a} \not= 0 \), and \( \overline{b} \not= 0 \) we find that both are zero divisors. But a field has no zero divisors, so \( \mathbb{Z}_n \) is not a field.

Example 3.28. Let \( F = \{0, 1, a, b\} \) be a set with four elements, and define operations by the following tables

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>b</td>
<td>a</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>b</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>a</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>a</td>
<td>0</td>
<td>a</td>
<td>b</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>b</td>
<td>1</td>
<td>a</td>
</tr>
</tbody>
</table>

A painful and tedious computation will show that \( F \), together with these operations, is a field, which we will not bother with. This example looks very ad-hoc. However, there is a systematic way to obtain this example from the field \( \mathbb{Z}_2 \); we will need to use this method of building fields from others to work with Reed-Solomon codes, an important class of error correcting codes.

Example 3.29. We can construct fields as subsets of \( \mathbb{R} \) or \( \mathbb{C} \). For example, let

\[
\mathbb{Q}(\sqrt{2}) = \left\{ a + b\sqrt{2} : a, b \in \mathbb{Q} \right\}
\]

and

\[
\mathbb{Q}(i) = \left\{ a + bi : a, b \in \mathbb{Q} \right\}
\]

The set \( \mathbb{Q}(\sqrt{2}) \) is a subset of \( \mathbb{R} \) and \( \mathbb{Q}(i) \) is a subset of \( \mathbb{C} \). We (partially) verify that \( \mathbb{Q}(\sqrt{2}) \) is a field and leave the other example for an exercise. We first note that ordinary addition and multiplication yield operations on \( \mathbb{Q}(\sqrt{2}) \). To see this, let \( a + b\sqrt{2} \) and \( c + d\sqrt{2} \in \mathbb{Q}(\sqrt{2}) \). Inherent in this statement is that \( a, b, c, d \in \mathbb{Q} \). Then

\[
(a + b\sqrt{2}) + (c + d\sqrt{2}) = (a + c) + (b + d)\sqrt{2}.
\]

\[\text{2 don't do BCH anymore}\]
Since \( a + c \) and \( b + d \) are rational numbers, this sum lies in \( \mathbb{Q}(\sqrt{2}) \). Thus, the sum of two elements of \( \mathbb{Q}(\sqrt{2}) \) is an element of \( \mathbb{Q}(\sqrt{2}) \). For multiplication, we have

\[
(a + b\sqrt{2}) \cdot (c + d\sqrt{2}) = ac + ad\sqrt{2} + bc\sqrt{2} + 2bd
\]

\[
= (ac + 2bd) + (ad + bc)\sqrt{2}.
\]

Since \( ac + 2bd \) and \( ad + bc \) are rational numbers, this product lies in \( \mathbb{Q}(\sqrt{2}) \). Therefore, we have operations of addition and multiplication on \( \mathbb{Q}(\sqrt{2}) \). Almost all of the properties to be a field are automatic. For example, commutativity of addition in \( \mathbb{Q}(\sqrt{2}) \) is a special case of commutativity of addition in \( \mathbb{R} \). The properties that are not automatic are existence of identities and existence of inverses. For the first, we have \( 0 = 0 + 0\sqrt{2} \) and \( 1 = 1 + 0\sqrt{2} \), which shows that \( \mathbb{Q}(\sqrt{2}) \) does contain identities. Also, if \( a, b \in \mathbb{Q} \), then the additive inverse of \( a + b\sqrt{2} \) is \( -(a + b\sqrt{2}) = (-a) + (-b)\sqrt{2} \), which lies in \( \mathbb{Q}(\sqrt{2}) \). Finally, we check for multiplicative inverses. This is the most involved part of verifying that \( \mathbb{Q}(\sqrt{2}) \) is a field. Let \( a, b \in \mathbb{Q} \), and consider a nonzero element \( a + b\sqrt{2} \). In order for this number to be nonzero, either \( a \) or \( b \) is nonzero. By rationalizing the denominator, similar to the trick of multiplying by the conjugate in complex numbers, we have

\[
\frac{1}{a + b\sqrt{2}} = \frac{a - b\sqrt{2}}{(a - b\sqrt{2})(a + b\sqrt{2})} = \frac{a - b\sqrt{2}}{a^2 - 2b^2}
\]

\[
= \left(\frac{-a}{a^2 - 2b^2}\right) + \left(\frac{-b}{a^2 - 2b^2}\right)\sqrt{2}.
\]

This calculation makes sense and shows that \( a + b\sqrt{2} \) has an inverse in \( \mathbb{Q}(\sqrt{2}) \) provided that \( a^2 - 2b^2 \neq 0 \); note that the coefficients are rational numbers if this occurs. To verify that \( a^2 - 2b^2 \neq 0 \), recall from the calculation above that \( (a + b\sqrt{2})(a - b\sqrt{2}) = a^2 - 2b^2 \); we can view this equation in \( \mathbb{R} \), which has no zero divisors. We are assuming that \( a + b\sqrt{2} \neq 0 \). If \( a - b\sqrt{2} \) is also nonzero, then \( a^2 - 2b^2 \neq 0 \) since it is a product of two nonzero real numbers. If \( a + b\sqrt{2} = 0 \), then \( b = 0 \) forces \( a = 0 \), but we know that either \( a \) or \( b \) is nonzero. If \( b \neq 0 \), then \( a + b\sqrt{2} = 0 \) yields \( \sqrt{2} = -a/b \), a rational number. However, \( \sqrt{2} \) is irrational. Thus, \( a - b\sqrt{2} \neq 0 \), and so \( a^2 - 2b^2 \neq 0 \). Thus, we do have multiplicative inverses in \( \mathbb{Q}(\sqrt{2}) \) of any nonzero element, and so \( \mathbb{Q}(\sqrt{2}) \) is a field.

It is not at all apparent that this example and the previous example have any relation. However, later on when we discuss the method of building fields from others hinted at in the previous example, we will see that both of these examples can be obtained by this method. The former example is built from \( \mathbb{Z}_2 \) and the current example is built from \( \mathbb{Q} \).

**Definition 3.30.** A commutative ring without zero divisors is called an integral domain.

**Example 3.31.** As the name implies, the ring of integers is an integral domain. Every field is an integral domain.

\footnote{prove this in a homework exercise. Also \( \sqrt{2} \)?}
We can generalize the previous proposition.

**Theorem 3.32.** Every finite integral domain is a field.

**Proof.** Let $D$ be a finite integral domain and $S = \{d_1, d_2, \ldots, d_n\}$ the set of its nonzero elements with $d_i \neq d_j$ for $i \neq j$. We must show that each $d_i$ has a multiplicative inverse in $S$. To that end, denote by $d_i S$ the set $\{d_i d_1, d_i d_2, \ldots, d_i d_n\}$ and observe that $d_i S \subseteq S$ because $D$ is an integral domain. But in fact $d_i S = S$ since otherwise there would be repetition among the elements $d_i d_j$. But $d_i d_j = d_i d_k$ holds only for $j = k$ by the cancellation law for multiplication. Thus $d_i S = S$ and because $1 \in S$, in particular $1 = d_i d_j$ for some $d_j$. \qed
Chapter 4

Linear Algebra and Linear Codes

In this chapter we review the main ideas of linear algebra. The one twist is that we allow our scalars to come from any field instead of just real numbers. In particular, the notion of vector space over the field $\mathbb{Z}_2$ will be essential in our study of coding theory. We will also need to look at other finite fields when we discuss Reed-Solomon codes. One benefit to working with vector spaces over finite fields is that all sets in question are finite, and so computers can be useful in working with them.

4.1 Vector Spaces

Let us recall the idea of a vector space. To discuss this notion, we have two types of objects, vectors and scalars. For example, $\mathbb{R}^3$, the usual three-dimensional space, is an example of the set of vectors, and the set of scalars in this case is $\mathbb{R}$, the set of real numbers. Probably you have only seen $\mathbb{R}$ acting as the set of scalars in previous courses. However, in any computation or definition about vector spaces, what you needed was that you could add, subtract, multiply, and divide scalars, and that these arithmetic operations had some appropriate properties. In fact, there is nothing we need about $\mathbb{R}$ other than it is a field to do everything you have seen about vector spaces. Therefore, we will simply review these concepts, except that we assume the set of scalars is some field $F$ instead of $\mathbb{R}$.

Recall that the main idea of a vector space $V$ is that we have two operations, (vector) addition and scalar multiplication. The addition operation is an ordinary binary operation on $V$. Scalar multiplication is somewhat different. This operation takes a pair consisting of a scalar and a vector and returns a vector. If $F$ is the set of scalars, then this operation is really a function from $F \times V$ to $V$.

**Definition 4.1.** Let $F$ be a field. An $F$-vector space is a nonempty set $V$ together with an operation $+$ on $V$ and an operation of scalar multiplication, such that for all $u, v, w \in V$ and $\alpha, \beta \in F$,

- $u + v = v + u$;
• \( u + (v + w) = (u + v) + w; \)
• there is a vector \( \mathbf{0} \) with \( u + \mathbf{0} = u \) for any \( u \in V \);
• for every \( u \in V \) there is a vector \( -u \in V \) with \( u + (-u) = \mathbf{0} \);
• \( \alpha(u + v) = \alpha u + \alpha v; \)
• \( (\alpha + \beta)u = \alpha u + \beta u; \)
• \( (\alpha \beta)u = \alpha(\beta u) \)
• \( 1 \cdot u = u. \)

Before we look at examples, we point out that the properties in the list above that involve only addition are exactly the same properties as in the definition of a ring. Therefore, only multiplication distinguishes a vector space from a ring. Later on we will focus on sets with only one operation.

We now illustrate the definition of a vector space with several examples.

**Example 4.2.** The most basic example we will use of an \( F \)-vector space is the space \( F^n \) of all \( n \)-tuples of elements of \( F \). That is,

\[ F^n = \{ (a_1, \ldots, a_n) : a_i \in F \}. \]

The operations on \( F^n \) that make \( F^n \) into a vector space are the pointwise operations:

\[ (a_1, \ldots, a_n) + (b_1, \ldots, b_n) = (a_1 + b_1, \ldots, a_n + b_n), \]
\[ \alpha(a_1, \ldots, a_n) = (\alpha a_1, \ldots, \alpha a_n). \]

**Example 4.3.** Let \( V \) be the set of all \( n \times m \) matrices with real number entries. We will denote by \( (a_{ij}) \) the matrix whose \( i, j \) entry is \( a_{ij} \). Recall that matrix addition and scalar multiplication are defined by

\[ (a_{ij}) + (b_{ij}) = (a_{ij} + b_{ij}), \]
\[ \alpha(a_{ij}) = (\alpha a_{ij}). \]

In other words, you add and do scalar multiplication componentwise. You should refer to a linear algebra text if necessary to recall that matrix arithmetic satisfies the properties required to see that \( V \) is an \( \mathbb{R} \)-vector space.

**Example 4.4.** We generalize the previous example a little. Let us replace \( \mathbb{R} \) by any field \( F \), and make \( V \) to be the set of all \( n \times m \) matrices with entries from \( F \). We use the same formulas to define addition and scalar multiplication. The proof that shows the set of real valued matrices of a given size forms a vector space also shows, without any changes, that the set of matrices with entries in \( F \) also is a vector space.
Example 4.5. Let $F$ be a field. We denote by $F[x]$ the set of polynomials with coefficients in $F$. Our usual definitions of addition and multiplication in $\mathbb{R}[x]$ work fine for $F[x]$. As a special example of multiplication, we can define scalar multiplication by $\alpha \sum_{i=0}^n a_i x^i = \sum_{i=0}^n \alpha a_i x^i$. With the operations of addition and scalar multiplication, a routine argument will show that $F[x]$ is an $F$-vector space. In fact, $F[x]$, together with addition and multiplication, is also a commutative ring. Thus, $F[x]$ gives us an example of both a ring and a vector space.

We now prove two simple properties of vector spaces. These are in spirit similar to statements we proved for rings. First of all, note that the proof that the additive identity and the additive inverse of an element of a ring are unique works word for word for vector spaces. This is because the proof only involved addition, and with respect to addition, there is no difference between a ring and a vector space.

Lemma 4.6. Let $F$ be a field and let $V$ be an $F$-vector space. If $v \in V$, then $0 \cdot v = 0$.

Proof. Let $v \in V$. To show that $0 \cdot v = 0$, we must show that $0 \cdot v$ satisfies the definition of the additive identity. By using one of the distributive laws, we see that

$$0 \cdot v + v = 0 \cdot v + 1 \cdot v = (0 + 1) \cdot v = 1 \cdot v = v = 0 + v.$$ 

By the cancellation law of addition, which holds for vector spaces just as for rings, we get $0 \cdot v = 0$.

Lemma 4.7. Let $F$ be a field and let $V$ be an $F$-vector space. If $v \in V$, then $(-1) \cdot v = -v$.

Proof. Let $v \in V$. To prove that $(-1) \cdot v$ is equal to $-v$, we need to prove that it is the additive inverse of $v$. We have

$$v + (-1) \cdot v = 1 \cdot v + (-1) \cdot v = (1 + (-1)) \cdot v = 0 \cdot v = 0$$

by the previous lemma. These equations do indeed show that $(-1) \cdot v$ is the additive inverse of $v$.

To consider the types of examples that most often arise, we need the notion of a subspace.

Definition 4.8. Let $F$ be a field and let $V$ be an $F$-vector space. A nonempty subset $W$ of $V$ is called a subspace of $V$ if the operations on $V$ induce operations on $W$, and if $W$ is an $F$-vector space under these operations.

Let us explain this definition. Addition on $V$ is a function from $V \times V$ to $V$. The product $W \times W$ is a subset of $V \times V$, so we can restrict the addition function to this smaller domain. The image of this function need not be inside $W$. However, if it is; that is, if addition restricts to a function $W \times W$ to $W$, then this restriction is a binary operation on $W$. When
this occurs, we say that $W$ is closed under the addition on $V$. To say this more symbolically, to say that $W$ is closed under addition means that for all $v, w \in W$, the sum $v + w$ must also be an element of $W$. Similarly, the scalar multiplication operation on $V$, which is a function from $F \times V$ to $V$, may restrict to a function $F \times W$ to $W$. If this happens, we say that $W$ is closed under scalar multiplication. Symbolically, this means that for all $\alpha \in F$ and all $v \in W$, the element $\alpha v$ must be an element of $W$. When $W$ is closed under addition and scalar multiplication, then we have appropriate operations on $W$ to discuss whether or not $W$ is a vector space. In fact, as the following lemma shows, if $W$ is closed under both addition and scalar multiplication, then $W$ is automatically a subspace of $V$.

**Lemma 4.9.** Let $F$ be a field and let $V$ be an $F$-vector space. Let $W$ be a nonempty subset of $V$, and suppose that (i) for each $v, w \in W$, the sum $v + w \in W$; and (ii) for each $\alpha \in F$ and $v \in W$, the product $\alpha v \in W$. Then $W$ is a subspace of $V$.

**Proof.** Suppose $W$ satisfies statements (i) and (ii) in the lemma. We must verify that $W$, together with its induced operations satisfies the definition of a vector space. Thus, we must verify the eight properties of the definition. Six of these properties are immediate from the fact that $V$ is a vector space. For instance, the commutative law of addition for $V$ yields $v + w = w + v$ for all $v, w \in V$. In particular, if $v, w \in W$, then $v$ and $w$ are also in $V$, so $v + w = w + v$. In a similar way, the associative property, the two distributive properties, and the property $\alpha(\beta v) = (\alpha \beta)v$ for $\alpha, \beta \in F$ and $v \in W$ all hold. We therefore need to check the two remaining properties. First, we need the existence of an additive identity. It is sufficient to show that if $0$ is the identity of $V$, then $0 \in W$. However, we are assuming that $W$ is nonempty; suppose $w \in W$ is any element of $W$. Since $W$ is closed under scalar multiplication, if $\alpha \in F$ is any scalar, then $\alpha w \in W$. In particular, choose $\alpha = 0$. Then $0 \cdot w \in W$. But, $0 \cdot w = 0$. Therefore, $0 \in W$, as desired. Finally, we need to have additive inverses of each element of $W$ to be inside $W$. Let $w \in W$. Again, by closure of scalar multiplication, we see that $(-1) \cdot w \in W$. But, since $(-1) \cdot w = -w$, the additive inverse of $w$ lies in $W$. This completes the proof. $\square$

**Example 4.10.** Let $u_1 = (1, 2, 3)$ and $u_2 = (4, 5, 6)$, two elements of the $\mathbb{R}$-vector space $\mathbb{R}^3$. Let $W$ be the set

$$W = \{\alpha u_1 + \beta u_2 : \alpha, \beta \in \mathbb{R}\}.$$  

This is the set of all linear combinations of $v$ and $w$. We claim that $W$ is a subspace of $\mathbb{R}^3$. Note that $W$ is nonempty, since both $u_1$ and $u_2$ are in $W$. That $u_1$ is in $W$ is because $u_1 = 1 \cdot u_1 + 0 \cdot u_2$. A similar reason holds for $u_2$. To see that $W$ is a subspace, we then need to verify that $W$ is closed under addition and scalar multiplication. For addition, let $v, w \in W$. By definition of $W$, there are scalars $\alpha, \beta, \gamma, \delta$ with $v = \alpha u_1 + \beta u_2$ and $w = \gamma u_1 + \delta u_2$. Then

$$v + w = (\alpha u_1 + \beta u_2) + (\gamma u_1 + \delta u_2)$$

$$= (\alpha + \gamma)u_1 + (\beta + \delta)u_2.$$
Since \( \alpha + \gamma \) and \( \beta + \delta \) are real numbers, the vector \( v + w \) has the form necessary to show that it is in element of \( W \). Therefore, \( W \) is closed under addition. Next, for scalar multiplication, let \( v \in W \) and let \( c \) be a scalar. Again, we may write \( v = \alpha u_1 + \beta u_2 \) for some scalars \( \alpha, \beta \). Then
\[
.cv = c(\alpha u_1 + \beta u_2) = c(\alpha u_1) + c(\beta u_2) \\
= (c\alpha)u_1 + (c\beta)u_2,
\]
which shows that \( cv \) is an element of \( W \). Therefore, by Lemma 4.9, \( W \) is a subspace of \( \mathbb{R}^3 \).

Notice that we only used that \( u_1 \) and \( u_2 \) are vectors in a vector space, and not that the vector space is \( \mathbb{R}^3 \) and that \( u_1 \) and \( u_2 \) are prescribed. This will allow us to easily generalize this example in an important way.

We can be a little more precise about what is \( W \). Since \( v = (1, 2, 3) \) and \( w = (4, 5, 6) \), we have
\[
W = \{ \alpha(1, 2, 3) + \beta(4, 5, 6) : \alpha, \beta \in \mathbb{R} \} \\
= \{ (\alpha + 4\beta, 2\alpha + 5\beta, 3\alpha + 6\beta : \alpha, \beta \in \mathbb{R} \}.
\]
In other words, \( W \) consists of all three-tuples that can be obtained as the product
\[
\begin{pmatrix}
1 & 4 \\
2 & 5 \\
3 & 6
\end{pmatrix}
\begin{pmatrix}
\alpha \\
\beta
\end{pmatrix}
\]
for any \( \alpha, \beta \in \mathbb{R} \). This observation was already used in the study of linear codes and will continue to be important for the further study of coding theory. If you recall your multivariable calculus, you can see a more geometric description of \( W \) in terms of the cross product. Calculating \( v \times w \) yields \((-3, 6, -3)\). This vector is perpendicular to both \( v \) and \( w \). Thus, using the dot product \( \cdot \), we have \( v \cdot (-3, 6, -3) = 0 \) and \( w \cdot (-3, 6, -3) = 0 \). In fact, some computation shows that \( W \) consists exactly of the set of vectors that satisfy the equation \( x \cdot (-3, 6, -3) = 0 \). Geometrically, this set is a plane that passes through the origin.

**Example 4.11.** Let
\[
A = \begin{pmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9
\end{pmatrix}.
\]
Consider the set \( W \) of all elements \( x = (\alpha, \beta, \gamma) \) of \( \mathbb{R}^3 \) that satisfy \( Ax = 0 \). We claim that \( W \) is a subspace of \( \mathbb{R}^3 \). To do this we will use two properties of matrix multiplication. First, note that \( W \) is nonempty since the zero vector \((0, 0, 0)\) satisfies the equation \( Ax = 0 \). Next, let \( v, w \in W \). Then \( Av = 0 \) and \( Aw = 0 \). Thus, \( A(v + w) = Av + Aw = 0 + 0 = 0 \), so \( v + w \in W \). Finally, if \( v \in W \) and \( \alpha \in \mathbb{R} \), then \( A(\alpha v) = \alpha(Av) = \alpha 0 = 0 \). Thus, \( \alpha v \in W \), and so \( W \) is a subspace of \( \mathbb{R}^3 \). As with the previous example, we did not use that \( A \) was a specific \( 3 \times 3 \) matrix, only that \( A \) was a matrix.
The properties of matrix multiplication we used in the proof above are (i) \( A(B + C) = AB + AC \), and (ii) \( A(\alpha B) = \alpha(AB) \) for any matrices \( A, B, C \) and any scalar \( C \). These facts can be found in any linear algebra text.

Exercises

1. Let \( w \) be a fixed vector in \( \mathbb{R}^3 \), and let \( W = \{ v \in \mathbb{R}^3 : w \cdot v = 0 \} \). Show that \( W \) is a subspace of \( \mathbb{R}^3 \). Find a matrix \( A \) so that \( W \) is the nullspace of \( A \).

2. Let \( n \) be a positive integer, and let \( W \) be the set of all words in \( \mathbb{Z}_2^n \) of even weight. Show that \( W \) is a subspace of \( \mathbb{Z}_2^n \).

3. Let \( V \) be a vector space over a field \( F \). If \( v_1, \ldots, v_n \in V \), show that the set \( W = \{ a_1v_1 + \cdots + a_nv_n : a_i \in F \} \) is a subspace of \( V \).

4.2 Linear Independence, Spanning, and Bases

In the previous section we defined vector spaces over an arbitrary field, gave some examples, and proved some simple properties about them. In this section we discuss the most important notions of vector spaces, those that lead to the idea of a basis. From both a theoretical and computational point of view, a basis is extremely important. Using a basis allows many computations to be simplified, and it also is an aide for proving results. As we will see, this and other definitions are the same as those for vector spaces over the real numbers. The primary notion we will use is that of a linear combination. If \( \{v_1, \ldots, v_n\} \) is a collection of vectors, then any vector of the form \( \sum_{i=1}^{n} a_i v_i \) is called a linear combination of the \( v_i \), where the \( a_i \) are scalars.

**Definition 4.12.** Let \( F \) be a field and let \( V \) be an \( F \)-vector space. A collection \( \{v_1, \ldots, v_n\} \) of elements of \( V \) is said to be linearly independent if whenever there are scalars \( a_1, \ldots, a_n \) with \( \sum_{i=1}^{n} a_i v_i = 0 \), then each \( a_i = 0 \).

**Definition 4.13.** Let \( F \) be a field and let \( V \) be an \( F \)-vector space. A collection \( \{v_1, \ldots, v_n\} \) of elements of \( V \) is said to be a spanning set for \( V \) if every element of \( V \) can be expressed in the form \( \sum_{i=1}^{n} a_i v_i \) for some choice of scalars \( a_1, \ldots, a_n \).

These two definitions together give us the notion of a basis.

**Definition 4.14.** Let \( F \) be a field and let \( V \) be an \( F \)-vector space. An ordered list \( [v_1, \ldots, v_n] \) of elements of \( V \) is said to be a basis for \( V \) if this set is both linearly independent and is a spanning set for \( V \).

To help understand these definitions, we look at several examples.
4.2. LINEAR INDEPENDENCE, SPANNING, AND BASES

Example 4.15. Let \( i = (1,0,0) \), \( j = (0,1,0) \), and \( k = (0,0,1) \), three elements of the \( \mathbb{R} \)-vector space \( \mathbb{R}^3 \). Then \([i,j,k]\) is a basis. To verify this we need to prove that this set is linearly independent and spans \( \mathbb{R}^3 \). First, for independence, suppose there are scalars \( a, b, \) and \( c \) with \( ai + bj + ck = 0 \). Then

\[
(0,0,0) = ai + bj + ck = a(1,0,0) + b(0,1,0) + c(0,0,1)
= (a,0,0) + (0,b,0) + (0,0,c) = (a,b,c).
\]

This vector equation yields \( a = b = c = 0 \). Thus, \([i,j,k]\) is indeed linearly independent. For spanning, if \( (x,y,z) \) is a vector in \( \mathbb{R}^3 \), then the equation \( ai + bj + ck = (a,b,c) \) above shows us that we can write this vector in terms of \( i, j, \) and \( k \) as \( (x,y,z) = xi + yj + zk \). Thus, any vector in \( \mathbb{R}^3 \) is a linear combination of \([i,j,k]\), so this set spans \( \mathbb{R}^3 \). Since this set both spans \( \mathbb{R}^3 \) and is independent, it is a basis of \( \mathbb{R}^3 \).

Example 4.16. Let \( W = \{(a,b,0) : a, b \in \mathbb{R}\} \), a subset of \( \mathbb{R}^3 \). In fact, \( W \) is a subspace of \( \mathbb{R}^3 \), as a calculation shows. The vectors \( i = (1,0,0) \) and \( j = (0,1,0) \) are elements of \( W \). We claim that \([i,j]\) is a basis for \( W \). An argument similar to that of the previous example will verify this; to summarize the ideas, we have \((a,b,0) = ai + bj\) is the unique way of expressing the vector \((a,b,0)\) as a linear combination of \( i \) and \( j \). This proves spanning, and the uniqueness will show linear independence since if we start with \((0,0,0)\), then \( a = b = 0 \), so the uniqueness shows that if \( xi + yj = (0,0,0)\), then \( x = y = 0 \).

Example 4.17. Recall the Hamming code of Chapter 2. It is the nullspace of

\[
H = \begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 1
\end{pmatrix},
\]

i.e. the set of solutions to the matrix equation \( Hx = 0 \). \( V \) is a subspace of \( \mathbb{Z}_2^7 \), to verify this, first suppose that \( v, w \in V \). Then \( Hv = Hw = 0 \). Thus, \( H(v+w) = Hv + Hw = 0 + 0 = 0 \), so \( v + w \in V \). Also, if \( v \in V \) and \( \alpha \in \mathbb{Z}_2 \), then \( H(\alpha v) = \alpha (Hv) = \alpha 0 = 0 \). Thus, \( V \) is closed under addition and scalar multiplication, so it is indeed a subspace of \( \mathbb{Z}_2^7 \). We have seen that an arbitrary solution \((x_1,x_2,x_3,x_4,x_5,x_6,x_7)\) to \( Hx = 0 \) satisfies

\[
\begin{align*}
x_1 &= x_3 + x_5 + x_7, \\
x_2 &= x_3 + x_6 + x_7, \\
x_4 &= x_5 + x_6 + x_7, \\
x_3,x_5,x_6,x_7 \text{ are arbitrary.}
\end{align*}
\]

We get four solutions by setting one arbitrary variable equal to 1 and the other three to 0. Doing so yields

\[
(1,1,1,0,0,0,0), (1,0,0,1,1,0,0), (0,1,0,1,0,1,0), (1,1,0,1,0,0,1).
\]
We claim that these form a basis for $V$. We can see this by noting that
\[
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4 \\
x_5 \\
x_6 \\
x_7 \\
\end{pmatrix} = \begin{pmatrix}
x_3 + x_5 + x_7 \\
x_3 + x_6 + x_7 \\
x_3 \\
x_5 + x_6 + x_7 \\
x_5 \\
x_6 \\
x_7 \\
\end{pmatrix} = x_3 \begin{pmatrix}
1 \\
1 \\
0 \\
0 \\
0 \\
0 \\
0 \\
\end{pmatrix} + x_5 \begin{pmatrix}
1 \\
0 \\
1 \\
0 \\
0 \\
0 \\
0 \\
\end{pmatrix} + x_6 \begin{pmatrix}
0 \\
0 \\
0 \\
1 \\
0 \\
0 \\
0 \\
\end{pmatrix} + x_7 \begin{pmatrix}
0 \\
0 \\
0 \\
1 \\
0 \\
0 \\
0 \\
\end{pmatrix},
\]
which shows that every solution is a linear combination of these four solutions. It also shows that if we have a linear combination of these four vectors that equals the zero vector, then the four coefficients are all 0. This proves that they form a basis for $V$.

**Example 4.18.** Let $V$ be a line through the origin in $\mathbb{R}^3$. From calculus or analytic geometry, this line can be described as the set of points of the form $t(a, b, c)$, where $t$ is an arbitrary scalar, and $(a, b, c)$ is a fixed nonzero vector on the line. Then $[(a, b, c)]$ forms a basis for $V$. First, it is obvious that it spans $V$; this is by the description given for $V$. For independence, if $\alpha(a, b, c) = 0$, then since $(a, b, c) \neq 0$, one of the entries is nonzero. Since $(0, 0, 0) = (\alpha a, \alpha b, \alpha c)$, we have $\alpha a = \alpha b = \alpha c = 0$. Either $a = b = c = 0$, which is false, or $\alpha = 0$.

**Example 4.19.** Let $V$ be a plane through the origin in $\mathbb{R}^3$. Recalling some facts from multivariable calculus, we see that the points $(x, y, z)$ in $V$ are those that satisfy the equation $ax + by + cz = 0$ for an appropriate choice of $a, b, c$. Suppose we have fixed values of $a, b,$ and $c$ that determine $V$ from this equation. For convenience, suppose that $a \neq 0$. We produce two specific elements of $V$. Let
\[
v = (-b/a, 1, 0),
\]
\[
w = (-c/a, 0, 1).
\]
These vectors were found by setting $y = 1$ and $z = 0$ and solving for $x$ to find $v$, and setting $y = 0$ and $z = 1$, and solving for $x$ to find $w$. The $\{v, w\}$ is a basis for $V$. First, for independence, if $\alpha$ and $\beta$ are scalars with $\alpha v + \beta w = 0$, then
\[
(0, 0, 0) = \alpha(-b/a, 1, 0) + \beta(-c/a, 0, 1) = (-\alpha b/a - \beta c/a, \alpha, \beta).
\]
Setting the second and third components equal yields $\alpha = 0$ and $\beta = 0$. Next, for spanning, let $(x, y, z)$ be a point on $V$. Then $ax + by + cz = 0$. Solving for $x$, we have $x = (-b/a)y + (-c/a)z$. Thus,
\[
(x, y, z) = ((-b/a)y + (-c/a)z, y, z) = ((-b/a)y, y, 0) + ((-c/a)z, 0, z)
\]
\[
= y(-b/a, 1, 0) + z(-c/a, 0, 1) = yv + zw.
\]
Thus, any point on $V$ is a linear combination of $v$ and $w$. Thus, $[v, w]$ both spans $V$ and is independent, so it forms a basis of $V$.
Example 4.20. Let $F$ be any field, and let $V = F^2$. Then $[(1,0), (0,1)]$ is a basis for $V$, which can be shown by an argument similar to that of the example of $\mathbb{R}^3$ above. However, we can produce lots of other bases for $V$. For example, let $a, b$ be nonzero elements of $F$. We claim that, if $-1 \neq 1$ in $F$, then $[(a,b), (a,-b)]$ is a basis for $V$. First, to see that this set is independent, suppose that $x, y \in F$ with

$$x(a,b) + y(a,-b) = (0,0).$$

Then $(xa + ya, xb - yb) = (0,0)$. This yields two equations

$$(x+y)a = 0,$$

$$(x-y)b = 0,$$

which has only the trivial solution $x = y = 0$ since $a$ and $b$ are both nonzero. This also uses the assumption that $-1 \neq 1$ in $F$, since adding the equations $x + y = 0$ and $x - y = 0$ yields $2x = 0$. Then $2 \neq 0$ if and only if $-1 \neq 1$, so $2x = 0$ forces $x = 0$. Similarly, $y = 0$.

From this example, we see that there are lots of bases for $F^2$. However, what is common to all the bases we constructed is that they contain two elements. This is a special case of a more general fact; we state the general result now.

Theorem 4.21. Let $F$ be a field and let $V$ be an $F$-vector space. Then $V$ has a basis. Moreover, any two bases have the same number of elements.

Instead of proving this theorem, we refer the reader to any intermediate level book on linear algebra.

Definition 4.22. Let $F$ be a field and let $V$ be an $F$-vector space. Then the dimension of $V$ is the number of elements in a basis.

We will use the following result to determine the number of codewords of an error correcting code. The codes we will consider will be $\mathbb{Z}_2$-vector spaces.

Proposition 4.23. Suppose that $V$ is a $\mathbb{Z}_2$-vector space. If $\dim(V) = n$, then $|V| = 2^n$.

Proof. We give two arguments for this. First, suppose that $[v_1, \ldots, v_n]$ is a basis for $V$. Then every element of $V$ is of the form $a_1v_1 + \cdots + a_nv_n$. So, to produce elements of $V$, we choose the scalars. Since $|\mathbb{Z}_2| = 2$, we have 2 choices for $a_1$, and 2 choices for $a_2$, and so on. Linear independence shows that these choices are independent, so, by a standard counting principle, we have $2 \cdot 2 \cdot \cdots \cdot 2 = 2^n$ total choices for the $a_i$. Each choice corresponds to exactly one vector of $V$, and since any vector occurs in such a form, we conclude that $|V| = 2^n$. For an alternative argument, recall that if $A$ and $B$ are finite sets, then $|A \times B| = |A| \cdot |B|$. Using induction, it follows that $|\mathbb{Z}_2^n| = 2^n$. Now, given a basis $[v_1, \ldots, v_n]$ for $V$, the function $\mathbb{Z}_2^n \rightarrow V$ given by $(a_1, \ldots, a_n) \mapsto a_1v_1 + \cdots + a_nv_n$ is onto since this set spans $V$, and it is 1-1 since the set is linearly independent. So, this function is a bijection. Thus, the two sets $\mathbb{Z}_2^n$ and $V$ have the same size. We conclude that $|V| = 2^n$. \qed
Exercises

1. If \( w = (1, 2, 3) \), find a basis for the subspace \( W \) of Problem 1 of section 4.1.

2. Find a basis for the linear code \( C = \{00000, 11010, 01101, 10111\} \).

3. Find a basis for the linear code

\[
C = \{000000, 110011, 101110, 101001, 011010, 011101, 000111\}
\]

and express each element of the code as a linear combination of the basis vectors.

4. Let \( C \) be the code spanned by the vectors 11110, 10101, 01011, 11011, 10000. Find a basis for \( C \).

5. Let \( W \) be a subspace of a vector space \( V \). If \( \dim(W) = \dim(V) \), show that \( W = V \). Conclude that if \( W \) is a proper subspace of \( V \) (meaning that \( W \) is a proper subset of \( V \)), then \( \dim(W) < \dim(V) \).

6. Let \( V \) be a vector space over \( \mathbb{Z}_2 \).

   (a) If \( u, v \in V \) are both nonzero and distinct, show that \( \{u, v\} \) is linearly independent.

   (b) If \( u, v, w \in V \) are all nonzero and distinct, show that \( \{u, v, w\} \) is linearly independent if \( w \neq u + v \).

7. Let \( \{v_1, \ldots, v_k\} \) be a set of linearly independent vectors in an \( F \)-vector space \( V \), and let \( w \in V \). If \( w \) is not in the span of \( \{v_1, \ldots, v_k\} \), show that \( \{v_1, \ldots, v_k, w\} \) is linearly independent.

8. Let \( C \) be the nullspace in \( \mathbb{Z}_2^8 \) of the matrix

\[
\begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\end{pmatrix}
\]

Determine the codewords in \( C \), its dimension, distance, and error correction capability. Find a generator matrix for \( C \).

---

1 exercises/theorems: to find a basis
(1) make \( A \) whose rows span \( C \). reduce \( A \) and take nonzero rows.
(2) pick \( v_1 \) nonzero. Pick \( v_2 \) not in the span, if possible. Pick \( v_3 \) not in the span of \( \{v_1, v_2\} \), etc.
4.3 Linear Codes

A linear code is a code $C$ which, as a subset of $\mathbb{Z}_2^n$, is subspace of $\mathbb{Z}_2^n$.

**Example 4.24.** If $A$ is an $m \times n$ matrix over $\mathbb{Z}_2$, then the nullspace of $A$ is a linear code of length $n$, since we know that the nullspace of a matrix is always a vector space. In particular, the Hamming code and the Golay code are linear codes.

**Example 4.25.** $\{(1,0),(0,1)\}$ is not a linear code. The smallest linear code containing $\{(1,0),(0,1)\}$ is the span of this set, which is equal to $\mathbb{Z}_2^2$. Similarly, $\{11100,00111\}$ is not a linear code. Its span is a linear code, and the span is $\{00000,11100,00111,11011\}$.

Recall that the distance $d$ of a code is defined to be $d = \min \{D(u,v) : u, v \in C, u \neq v\}$. Recall that $D(u,v) = \text{wt}(u+v)$.

**Lemma 4.26.** If $C$ is a linear code with distance $d$, then $d = \min \{\text{wt}(v) : v \in C, v \neq 0\}$.

**Proof.** Let $e = \min \{\text{wt}(v) : v \in C, v \neq 0\}$. We may write $e = \text{wt}(u)$ for some nonzero vector $u \in C$. Then $e = D(u,0)$. Since $0 \in C$, we see that $d \leq e$ by definition of $d$. Conversely, we may write $d = D(v,w)$ for some distinct $v,w \in C$. Then $d = \text{wt}(v+w)$. Since $C$ is a linear code, $v+w \in C$, and $v+w \neq 0$ since $v \neq w$. Therefore, by definition of $e$, we see that $e \leq d$. Thus, $d = e$, as desired.

This lemma simplifies finding the distance of a linear code. For example, the Hamming code is

$$C = \{111111,0011001,0000000,1110000,1101001,1100110,0110011,0100101,0010110,0111100,1011010,1010101,1000011,0101010,1001100,0001111\}.$$

A quick inspection shows that the smallest weight of a nonzero codeword of $C$ is 3; therefore, $d = 3$.

Let $C$ be a linear code of length $n$ and distance $d$. If the dimension of $C$ is $k$, then we refer to $C$ as an $(n,k,d)$ linear code. Any matrix whose rows constitute a basis for $C$ is called a generator matrix for $C$. Thus a generator matrix for an $(n,k,d)$ linear code is an $n \times k$ matrix whose rank is $k$. The rows of any such matrix $G$ span an $(n,k,d)$ linear code $C$, and any matrix row equivalent to $G$ is also a generator matrix for $C$.

The reason for the terminology generator matrix is the following theorem.

**Theorem 4.27.** Let $C$ be an $(n,k,d)$ linear code with generator matrix $G$. Then $C = \{vG : v \in \mathbb{Z}_2^k\}$.

**Proof.** For $v \in \mathbb{Z}_2^k$, thought of as a row vector, $vG$ is a linear combination of the rows of $G$. Allowing $v$ range over all of $\mathbb{Z}_2^k$ results in the span of the rows of $G$ which is, by definition, $C$. □
For any row vector \( v \in \mathbb{Z}_2^k \), \( vG \) is the linear combination of the rows of \( G \) whose coefficients are the entries of \( v \). Therefore \( \{vG : v \in \mathbb{Z}_2^k \} \) is the row space of \( G \) which, by definition is \( C \).

The generator matrix thus “encodes” \( v \) as \( vG \) in \( C \).

**Example 4.28.** Recall that the Hamming code \( C \) is defined to be the nullspace of the matrix

\[
H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}.
\]

A matrix calculation shows that the solutions to the system equation \( Hx = 0 \) are

\[
x_1 = x_3 + x_5 + x_7,
\]
\[
x_2 = x_3 + x_6 + x_7,
\]
\[
x_4 = x_5 + x_6 + x_7,
\]
\[
x_3, x_5, x_6, x_7 \text{ are arbitrary}.
\]

Thus, the elements of \( C \), written as columns, are the vectors of the form

\[
\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = x_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + x_5 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + x_6 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + x_7 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}.
\]

The four vectors in the right hand side of this equation, written as rows, form a basis for \( C \). That is,

\[
[1110000, 1001100, 0101010, 1101001]
\]

is a basis for \( C \). Thus, the matrix

\[
\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \end{pmatrix}
\]

is a generator matrix for \( C \). Alternatively, if we row reduce this matrix, we obtain another generator matrix. Doing so, the matrix

\[
G = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}
\]
is another generator matrix. Thus, we may describe the Hamming code as \( C = \{ vG : v \in \mathbb{Z}_2^4 \} \). We may view encoding a four-tuple \( v \) as \( vG \) as appending three check digits to \( v \), since for any \( v \), the product \( vG \) has \( v \) as the first 4 components; this is because the left 4 \times 4 submatrix of \( G \) is the identity matrix. For example, if \( v = 1011 \), then \( vG = 1011010 \).

Thus, in general, by using a generator matrix in reduced echelon form, we may view encoding \( v \) as \( vG \) as adding a certain number (which is \( n - k \)) of extra digits to \( v \). In this way, linear codes can be constructed in a way that, in some sense, generalizes the identification number schemes we discussed in the first chapter.

We defined a generator matrix of an \((n, k, d)\) linear code \( C \) to be a matrix \( G \) whose rows form a basis for \( C \). Then \( G \) is a \( k \times n \) matrix of rank \( k \). Instead of producing generator matrices from linear codes, we can produce linear codes from matrices.

**Proposition 4.29.** Let \( G \) be a \( k \times n \) matrix over \( \mathbb{Z}_2 \) of rank \( k \). If \( C = \{ vG : v \in \mathbb{Z}_2^n \} \), then \( C \) is an \((n, k, d)\) linear code for some \( d \), and \( G \) is a generator matrix for \( C \).

**Proof.** We note that, by thinking about how we multiply matrices, if \( v = a_1 \cdots a_k \) and if \( r_1, \ldots, r_k \) are the \( k \) rows of \( G \), then \( vG = \sum_i a_ir_i \). Moreover, this equation shows that any linear combination of rows of \( G \) can be written in the form \( vG \). Thus, \( C \) is the linear span of the rows of \( C \); thus, \( C \) is a subspace of \( \mathbb{Z}_2^n \). Moreover, since the rows of \( G \) are linearly independent, they form a basis for \( C \), and so \( C \) has dimension \( k \). If \( d \) is the distance of \( C \), then \( C \) is an \((n, k, d)\) code. Moreover, \( G \) is a generator matrix since its rows form a basis of \( C \). \( \square \)

We initially constructed codes as nullspaces of matrices. We would like to make some simple restrictions on which matrices do we use to build codes in this way. If \( C \) is an \((n, k, d)\) code, then a matrix \( H \) is called a **parity check matrix** for \( C \) if \( H \) is a \((n - k) \times n\) matrix for which \( C \), when writing its elements as columns, is the nullspace of \( H \). Alternatively, we may say that \( H \) is a parity check matrix for \( C \) if \( C \) is the nullspace of \( H \) and if the rows of \( H \) are linearly independent. We leave the proof of the equivalence of these two statements as an exercise. To keep straight the difference between vectors and column matrices, if \( H \) is a parity check matrix for \( C \), then by writing \( A^T \) for the transpose of a matrix,

\[
C = \{ v \in \mathbb{Z}_2^n : Hv^T = 0 \}.
\]

**Example 4.30.** The Hamming matrix \( H \) is a parity check matrix for the Hamming code \( C \), since \( C \) is the nullspace of \( H \) and since the rows of \( H \) are seen to be linearly independent. Similarly, the matrix \( A = [I : B] \) defined in Section 2.5 is a parity check matrix for the Golay code since the code is the nullspace of \( A \) and since the 12 rows of \( A \) are linearly independent.

The generator and parity check matrices of a code both have the property that their rows are linearly independent. Thus, a parity check matrix of a code can be taken to be a
generator matrix of another code. We can be more specific about this other code. To do so, we define an analogue of the dot product. If \( u, v \in \mathbb{Z}_2^n \), define \( u \cdot v \) by
\[
 u_1 \cdots u_n \cdot v_1 \cdots v_n = u_1 v_1 + \cdots + u_n v_n.
\]
It an easy exercise to see that if \( u, v, w \in \mathbb{Z}_2^n \) and \( a \in \mathbb{Z}_2 \) that \( (u + w) \cdot v = u \cdot v + w \cdot v \), and \( u \cdot (av) = a(u \cdot v) \). If \( C \) is a code, we define the dual code
\[
 C^\perp = \{ v \in \mathbb{Z}_2^n : v \cdot c = 0 \text{ for all } c \in C \}.
\]

**Lemma 4.31.** Let \( C \) be a linear code. Then \( C^\perp \) is a linear code. Moreover, if \( [c_1, \ldots, c_k] \) is a basis for \( C \), then \( v \in C^\perp \) if and only if \( v \cdot c_i = 0 \) for each \( i \).

**Proof.** To show that \( C^\perp \) is linear we must show that it is closed under addition. Let \( v, w \in C^\perp \). Then, for each \( c \in C \), we have \( v \cdot c = 0 \) and \( w \cdot c = 0 \). Therefore, \( (v + w) \cdot c = v \cdot c + w \cdot c = 0 + 0 = 0 \). Thus, \( v + w \in C^\perp \). Next, let \( [c_1, \ldots, c_k] \) be a basis for \( C \). If \( v \in C^\perp \), then \( v \cdot c = 0 \) for all \( c \in C \). Since the \( c_i \) are elements of \( C \), this implies that \( v \cdot c_i = 0 \) for all \( i \). Conversely, if \( v \) is a word such that \( v \cdot c_i = 0 \) for all \( i \), then let \( c \in C \). We may write \( c = a_1 c_1 + \cdots + a_k c_k \) for some \( a_i \in \mathbb{Z}_2 \). Then
\[
 v \cdot c = v \cdot (a_1 c_1 + \cdots + a_k c_k) = v \cdot (a_1 c_1) + \cdots + v \cdot (a_k c_k) = a_1 (v \cdot c_1) + \cdots + a_k (v \cdot c_k) = a_1 0 + \cdots + a_k 0 = 0.
\]
Thus, \( v \in C^\perp \). This proves that \( C^\perp = \{ v \in \mathbb{Z}_2^n : v \cdot c_i = 0 \text{ for each } i \} \). \( \square \)

**Theorem 4.32.** Let \( C \) be a linear code of length \( n \) and dimension \( k \) with generator matrix \( G \) and parity check matrix \( H \).

1. \( HG^T = 0 \).
2. \( H \) is a generator matrix for \( C^\perp \) and \( G \) is a parity check matrix for \( C^\perp \).
3. The linear code \( C^\perp \) has dimension \( n - k \).
4. \((C^\perp)^\perp = C\).
5. If \( H' \) is a generator matrix for \( C^\perp \) and \( G' \) is a parity check matrix for \( C^\perp \), then \( H' \) is a parity check matrix for \( C \) and \( G' \) is a generator matrix for \( C \).

**Proof.** Let \( v \in \mathbb{Z}_2^k \). Then \( vG \in C \). Thus, by definition of parity check matrix, \( H(vG)^T = 0 \). But, \( (vG)^T = G^Tv^T \). Therefore, \( (HG^T)v^T = 0 \). Since \( v \) can be any element of \( \mathbb{Z}_2^k \), we see that \( (HG^T)x = 0 \) for any column vector \( x \). If \( e_i \) is the column vector with a 1 in the \( i \)-th coordinate and 0 elsewhere, we see that \( 0 = (HG^T)e_i \), and this product is the \( i \)-th column of \( HG^T \). Since this is true for all \( i \), we see that \( HG^T = 0 \). This proves the first statement.

For the second, we note that we only need to prove that \( C^\perp \) is the nullspace of \( G \) and that the rows of \( H \) form a basis of \( C^\perp \). For the first part, let \( v \in C^\perp \). To see that \( v \) is in the
nullspace, we must see that $Gv^T = 0$. However, the $i$-th entry of $Gv^T$ is the product of the $i$-th row $r_i$ of $G$ with $v^T$. In other words, it is $r_i \cdot v$. However, since the rows of $G$ form a basis for $C$, we see that $r_i \cdot v = 0$. Thus, $Gv^T = 0$. Conversely, if $v$ is in the nullspace of $G$, then $Gv^T = 0$. This means that $r_i \cdot v = 0$ for all $i$. The rows of $G$ form a basis for $C$; thus, by the lemma, $v \in C^\perp$.

As a consequence of knowing that the nullspace of $G$ is $C^\perp$, we prove Statement 3. From the rank plus nullity theorem, $n$ is the sum of the dimensions of the nullspace and the row space of $G$. By definition, the row space of $G$ is $C$, so its dimension is $k$. Thus, the nullspace has dimension $n - k$. But, this space is $C^\perp$. So, $\dim(C^\perp) = n - k$. To finish the proof of Statement (2) we show that the rows of $H$ form a basis of $C^\perp$. Because we know that the rows are independent, we only need to see that they span $C^\perp$. We first show that the rows are elements of $C^\perp$. Since $H$ is a parity check matrix for $C$, if $v \in C$, then $Hv^T = 0$. The $i$-th entry of $Hv^T$ is the dot product of the $i$-th row of $H$ with $v$; thus, this dot product is 0. Since this is true for all $v \in C$, we see that the rows of $H$ are all elements of $C^\perp$. So, the row space of $H^\perp$ is contained in $C^\perp$. However, both of these spaces have dimension $n - k$; we are using the assumption that the rows of $H$ are linearly independent to see that the row space of $H$ has dimension $n - k$. It is an exercise to show that if a subspace of a vector space has the same dimension as the vector space, then they are equal. From this we conclude that $C^\perp$ is the row space of $H$. Thus, the rows of $H$ form a basis for $C^\perp$, which says that $H$ is a generator matrix for $C^\perp$.

To prove (4), we note that the inclusion $C \subseteq (C^\perp)^\perp$ follows from the definitions: If $c \in C$, then $c \cdot v = 0$ for all $v \in C^\perp$ by definition of $C^\perp$. This is what we need to see that $c \in (C^\perp)^\perp$. However, from (3) we see that the dimension of $(C^\perp)^\perp$ is $n - (n - k) = k$, which is the dimension of $C$. From the inclusion $C \subseteq (C^\perp)^\perp$ we then conclude $C = (C^\perp)^\perp$. Finally, (5) follows from the previous statements, since if $H'$ is a generator matrix for $C^\perp$, then it is a parity check matrix for $(C^\perp)^\perp = C$, and if $G'$ is a parity check matrix for $C^\perp$, then it is a generator matrix for $(C^\perp)^\perp = C$. $\square$

Exercises

1. Define a dot product on $\mathbb{Z}_2^n$ by $a_1 \cdots a_n \cdot b_1 \cdots b_n = a_1 b_1 + \cdots + a_n b_n$, mimicking the dot product on $\mathbb{R}^3$.

(a) If $C \subseteq \mathbb{Z}_2^n$ is a code, define $C^\perp = \{v \in \mathbb{Z}_2^n : v \cdot c = 0 \text{ for all } c \in C\}$. Show that $C^\perp$ is a subspace of $\mathbb{Z}_2^n$.

(b) If $[c_1, \ldots, c_k]$ is a basis for $C$, show that $C^\perp = \{v \in \mathbb{Z}_2^n : v \cdot c_i = 0 \text{ for all } i\}$.

(c) If $[c_1, \ldots, c_k]$ is a basis for $C$ and if $A$ is the matrix whose $i$-th row is $c_i$, show that $C^\perp$ is the nullspace of $A$.

The code $C^\perp$ is called the dual code to $C$. 
2. Find a basis for the Hamming code $C$, and determine the dual code $C^\perp$.

3. A generator matrix for a code $C$ has for its rows a basis of $C$, and a parity check matrix has $C$ as its nullspace. Find a generator matrix $G$ and a parity check matrix $H$ for the code of the previous problem. Check that $HG^T = 0$. Then prove that a parity check matrix and generator matrix for a code must satisfy this property.

4. Let

$$G = \begin{pmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 0
\end{pmatrix}$$

be a generator matrix for a code $C$. Write out the codewords of $C$ and find a parity check matrix for $C$. 

Chapter 5

Field Extensions and Ruler and Compass Constructions

One remarkable application of abstract algebra arises in connection with four classical questions about geometric constructions:

1. Is it possible to trisect an arbitrary angle?
2. Is it possible to construct a square whose area is equal to that of a given circle?
3. Which regular polygons are constructible?
4. Is it possible to construct a cube of exactly twice the volume of a given cube?

Questions 1-3 refer to constructions in the real plane $\mathbb{R}^2$ and question 4 to constructions in $\mathbb{R}^3$. The tools we are allowed are only a compass and an uncalibrated ruler, i.e. devices for constructing circles and straight line segments, without measuring. The only constructions we can make come from a sequence of basic ones:

1. Connecting two given points with a straight line,
2. Drawing the circle centered at a given point passing through another given point,
3. Constructing a point as the intersection of two lines,
4. Constructing a point(s) as the intersection of a line and a circle, and
5. Constructing a point(s) as the intersection of two circles.

As an example of a construction, consider this familiar one:

**Bisecting a line segment**: Suppose that $A$ and $B$ are two distinct points in the plane. To bisect the line segment $AB$ construct the circle centered at $A$ passing through $B$ and the one centered at $B$ and passing through $A$. These circles intersect at two points and the line segment joining them bisects $AB$. 
Suppose that we identify two points $A$ and $B$ in the plane and declare that the distance between them is equal to 1. By drawing the line through $A$ and $B$ and using our unit distance $AB$ as a reference, we can construct all points on this line whose distance from point $A$ is an integer. Our first construction enables us to construct a sequence $\{A_i\}$ of points whose distance from $A$ is $(\frac{1}{2})^i$ for every positive integer $i$ and similar sequences for each of the new points we have constructed. However, we can also construct points separated by an irrational distance.

**Exercise 5.1.** Let $A$ and $B$ be points in the plane which we declare to be one unit apart. Denote by $L$ the perpendicular bisector of $AB$ and $C$ the midpoint of $AB$. Construct the circle centered at $A$ and passing through $B$. Its radius is equal to 1 so that it intersects $L$ at a point $D$ one unit away from $A$. Calculate the length of the line segment $CD$ and observe that it is equal to $\sqrt{2}$.

**Exercise 5.2.** Prove that $\sqrt{2}$ is irrational. (Hint: If $\sqrt{2}$ is rational, then it can be written as $\frac{a}{b}$ with $a$ and $b$ relatively prime integers. From

$$\sqrt{2} = \frac{a}{b}, \text{ or } b\sqrt{2} = a,$$

deduce that both $a$ and $b$ are divisible by 2).

Our investigation of constructibility proceeds by first using familiar construction techniques to establish a coordinate system on the plane, i.e. by constructing the points with integer coordinates. Then we examine the coordinates of point(s) lying on the intersection of two lines joining constructed points, on the intersection of a line with a circle of radius equal to a constructed distance centered at a constructed point, and on the intersection of two such circles. The connection with algebra comes from the fact that the collection of coordinates of all constructed points constitutes a field strictly between the fields of rational numbers and real numbers. From the properties of this field we will be able to investigate solutions of the classical constructibility problems. The next section is a brief digression into field theory.
5.1 Field Extensions

Definition 5.3. If $F$ is a field, then a field extension of $F$ is a field $K$ that contains $F$, and for which the field operations restrict to the field operation of $F$. Under these circumstances we also say that $F$ is a subfield of $K$.

For example, $\mathbb{C}$ is a field extension of $\mathbb{R}$ since $\mathbb{C}$ is a field containing $\mathbb{R}$ and the field operations on the real elements of $\mathbb{C}$ are precisely those of the field $\mathbb{R}$. Similarly, both $\mathbb{C}$ and $\mathbb{R}$ are field extensions of $\mathbb{Q}$. In this chapter we will focus on extensions of the field of rational numbers. For the purposes of investigating error correcting codes, we also consider field extensions, with particular attention to extensions of finite fields. Familiar concepts from linear algebra, (i.e. linear independence, spanning, basis, and dimension) provide essential tools for the study of field extensions.

Exercise 5.4. Suppose that $K$ is an extension field of $F$. Prove that $K$ is an $F$-vector space where the vector space addition in $K$ is addition in $K$ as a field and scalar multiplication $\alpha \beta$ for $\alpha \in F, \beta \in K$ is the multiplication in the field $K$.

Example 5.5. Let $F = \mathbb{Q}$ and $K = \{\alpha + \beta\sqrt{2} : \alpha, \beta \in \mathbb{Q}\}$. Since

$$(\alpha_1 + \beta_1\sqrt{2})(\alpha_2 + \beta_2\sqrt{2}) = (\alpha_1\alpha_2 + 2\beta_1\beta_2) + (\alpha_1\beta_2 + \beta_1\alpha_2)\sqrt{2}$$

we see that $K$ is closed under multiplication. It is clearly closed under addition and contains $\mathbb{Q}$ as the subset $\{\alpha + \beta\sqrt{2} : \alpha \in \mathbb{Q}, \beta = 0\}$. Note that if either $\alpha \neq 0$ or $\beta \neq 0$ then $\alpha^2 + 2\beta^2 \neq 0$. A multiplication reveals that

$$\left(\frac{\alpha}{\alpha^2 + 2\beta^2} - \frac{\beta}{\alpha^2 + 2\beta^2}\sqrt{2}\right)(\alpha + \beta\sqrt{2}) = 1$$

so that $\frac{\alpha}{\alpha^2 + 2\beta^2} - \frac{\beta}{\alpha^2 + 2\beta^2}\sqrt{2}$ is the multiplicative inverse of $\alpha + \beta\sqrt{2}$ and is an element of $K$. Thus nonzero elements of $K$ all have multiplicative inverses. This field $K$ is usually denoted by $\mathbb{Q}(\sqrt{2})$.

Definition 5.6. Let $K$ be a field extension of $F$. We denote the dimension of $K$ as an $F$-vector space by $[K : F]$.

Exercise 5.7.

Example 5.8. The dimension of $\mathbb{Q}(\sqrt{2})$ as a $\mathbb{Q}$ vector space is 2 since 1 and $\sqrt{2}$ form a basis. Thus, $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$.

Given a field extension $K$ of $F$ and an element $\alpha \in K$ we denote by $F(\alpha)$ the smallest field extension of $F$ which contains $\alpha$. By smallest we mean that for any other field extension $L$ of $F$ with $\alpha \in L$, we have $F(\alpha) \subseteq L$. For $\alpha, \beta \in K$ we write $F(\alpha, \beta)$ instead of $F(\alpha)(\beta)$. Similarly for any finite collection $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ of elements of $K$ we write $F(\alpha_1, \alpha_2, \ldots, \alpha_n)$ for the smallest field extension of $F$ containing these elements.
Exercise 5.9. Verify that \( \mathbb{Q}(\sqrt{2}) \) as defined above is the smallest field extension of \( \mathbb{Q} \) which contains \( \sqrt{2} \).

Exercise 5.10. Given a field extension \( K \) of \( F \) and an element \( \alpha \in K \) prove that \( F(\alpha) \) is the intersection of all fields \( L \) with \( F \subseteq L \subseteq K \) and \( \alpha \in L \).

Example 5.11. With \( i \) denoting a complex square root of \(-1\), \( \mathbb{R}(i) = \mathbb{C} \).

Example 5.12. The complex cube roots of \( 1 \) solve \( x^3 - 1 = 0 \). Since \( x^3 - 1 = (x - 1)(x^2 + x + 1) \), the three cube roots of \( 1 \) are

\[
1, \quad \omega_1 = \frac{-1 + i\sqrt{3}}{2}, \quad \text{and} \quad \omega_2 = \frac{-1 - i\sqrt{3}}{2}.
\]

Exercise 5.13. Verify that \( \omega_2 = \omega_1^2 \), so that \( \mathbb{Q}(\omega_1) \) contains all three cube roots of \( 1 \).

Exercise 5.14. Verify that \( \omega_1 = e^{2\pi i/3} \).

Exercise 5.15. Show for any positive integer \( n \) that the set of all \( n^{th} \) roots of \( 1 \) lies in the field \( \mathbb{Q}(e^{2\pi i/n}) \).

Theorem 5.16 (Dimension Formula). Let \( K \) be a field extension of \( F \) and \( L \) a field extension of \( K \). Then

\[
[L : F] = [L : K][K : F].
\]

Proof. If either \([L : K]\) or \([K : F]\) is infinite, then so is \([L : F]\) and there is nothing to show. So suppose that \([K : F] = m\) and \([L : K] = n\). Let \( \{a_1, \ldots, a_m\} \) and \( \{b_1, \ldots, b_n\} \) be bases for \( K \) over \( F \) and \( L \) over \( K \) respectively. We show that the elements \( a_ib_j \in L \), for \( 1 \leq i \leq m, \ 1 \leq j \leq n \), constitute a basis for \( L \) as an \( F \) vector space.

(Linear independence): Suppose that

\[
\sum_{1 \leq i \leq m, 1 \leq j \leq n} \gamma_{ij}a_ib_j = 0
\]

for some \( \gamma_{ij} \in F \). Then

\[
\sum_{1 \leq j \leq n} \left( \sum_{1 \leq i \leq m} \gamma_{ij}a_i \right)b_j = 0
\]

gives a dependence relation among \( b_1, \ldots, b_n \) over \( K \). Since they are linearly independent, we must have the coefficient of each \( b_j \) equal to 0, i.e.,

\[
\sum_{1 \leq i \leq m} \gamma_{ij}a_i = 0 \quad \text{for each} \quad j.
\]

But the linear independence of the \( a_i \) then forces each \( \gamma_{ij} = 0 \).

(Spanning) Let \( c \in L \). Since \( \{b_1, \ldots, b_n\} \) spans \( L \) over \( K \) we can write \( c = \sum_{1 \leq j \leq n} \alpha_jb_j \) for some \( \alpha_j \in K \). Each \( \alpha_j \) can be expanded as an \( F \) linear combination of the \( a_i \) as

\[
\alpha_j = \sum_{1 \leq i \leq m} \beta_{ij}a_i.
\]

Combining these expressions we obtain

\[
c = \sum_{1 \leq j \leq n} \alpha_jb_j = \sum_{1 \leq j \leq n} \sum_{1 \leq i \leq m} \beta_{ij}a_ib_j.
\]
Exercises

1. Determine $[\mathbb{Q}(\sqrt{2}, e^{2\pi i}) : \mathbb{Q}]$.

2. If $K$ is a field extension of $F$, prove that $K = F$ if and only if $[K : F] = 1$.

5.2 Ruler and Compass Constructions

Our first goal is to construct a coordinate system by choosing a length to be the unit length and to construct all of the points in the real plane $\mathbb{R}^2$ with integer coordinates. Begin with two points $A$ and $B$ in the plane and draw the line $L$ joining them. Declare $L$ to be the horizontal axis and the line $L'$ bisecting the line segment $AB$ (constructed at the beginning of the chapter) the vertical axis. The point of intersection of the two axes is denoted $O$ and referred to as the origin of the coordinate system. Declare the length of the line segment $OA$ to be equal to 1. The following construction applied to $L$ and $L'$ enables the construction of our coordinate system.

Given a line $L$ and a point $A$ not on $L$, constructing the unique line though $A$ that is parallel to $L$:

\[ A \]

\[ L \]

Define a point $(a, b)$ in the plane to be constructible if it can be reached from the integer grid points by a sequence of the basic constructions 1-5 above. Define a real number $a$ to be constructible if $(a, 0)$ is a constructible point. Note that the first exercise of this chapter can be used to show that $\sqrt{3}/2$ is a constructible number and our bisection procedure implies that if $a$ is constructible, then so is $a/2$. Thus the constructible numbers form a set of real numbers, properly larger than the integers, that contains at least some noninteger rational numbers and some irrationals. Our goal is to find an algebraic description of the set of constructible numbers.

Exercises

1. (Dropping a perpendicular) Construct two points $P, Q$ on $L$ by intersecting $L$ with a circle centered at $A$. Bisect $PQ$ as above and explain why the line joining $A$ to the midpoint $B$ of $PQ$ is perpendicular to $L$.

2. Bisect $QA$ and construct the line joining $B$ and the midpoint of $QA$.

3. Intersect the circle centered at $B$ whose radius is the length of $QA$ with the line of the previous exercise. Now complete the construction of the line through $A$ parallel to $L$. 
4. Explain how to construct all points \((a, b)\) in \(\mathbb{R}^2\) where \(a, b\) are integers.

### 5.3 Constructions and Field Extensions

Notice that if \(a\) is constructible, then so is \(-a\): construct the lines joining \((a, 0)\) and \((1, 0)\) and \((0, 1)\) and their parallels through \((0, 1)\) and \((1, 0)\) respectively. Furthermore, if \(L\) is a line, \(P\) a constructible point on \(L\), and \(a\) a constructible number, then the points on \(L\) that are at distance \(a\) from \(P\) are also constructible: using \(L\) as the horizontal axis and \(P\) as the origin of the coordinate system, the sequence of constructions producing \((\pm a, 0)\) yield the desired points. In particular, the constructible numbers are closed under addition. Call a line constructible if it joins two constructible points and a circle constructible if its center is a constructible point and its radius a constructible number. Thus we can “mark off” any constructible length on a constructible line (and construct a constructible circle centered at any constructible point).

**Exercise 5.17.** Given that \(A\) and \(B\) are constructible numbers, with \(B \neq 0\), explain why the points and parallel line segments \(L_1\) and \(L_2\) in the diagram below are constructible.

![Diagram](image)

**Exercise 5.18.** Use what you know about similar triangles to deduce the value of \(x\) and then explain why \(\frac{A}{B}\) is constructible.

Since \(AB = \frac{A}{B}\) we find that the set of constructible numbers is closed under multiplication as well.

**Theorem 5.19.** The set of constructible numbers is a field extension of the rational numbers.

**Proof.** The set of constructible numbers \(C\) contains the integers and is closed under real number addition, multiplication, subtraction and division by nonzero elements. Thus \(C\) constitutes a subfield of the real numbers. Since \(C\) contains the integers and their reciprocals, and is closed under multiplication, we obtain an extension of the field of rational numbers. 

**Corollary 5.20.** Any field containing the integers is an extension field of \(\mathbb{Q}\).
The field of constructible numbers is closed under one more operation.

**Exercise 5.21.** Suppose that $C$ is a constructible number. Explain why the labeled points in the diagram below are constructible:

![Diagram](image_url)

**Exercise 5.22.** From the remarks above, the circle $C$ of radius $\frac{C+1}{2}$ centered at $(\frac{C+1}{2}, 0)$ is constructible. Write down its equation.

**Theorem 5.23.** If $C$ is a constructible number then so is $\sqrt{C}$.

**Exercise 5.24.** Construct the intersection of the line through $(1, 0)$ that is parallel to $L_2$ with the circle $C$. Call this point $(1, y)$. Evaluate $y$ to conclude that $\sqrt{C}$ is constructible.

![Diagram](image_url)

**Example 5.25.** $\alpha = \frac{1+\sqrt{3}}{2+\sqrt{5}}$ is constructible.

**Exercise 5.26.** With $\alpha$ as above, rationalize the denominator and calculate $[\mathbb{Q}(\alpha) : \mathbb{Q}]$. 
We can view the field $\mathbb{C}$ of constructible numbers as being built up in stages from the rational numbers. If a construction results in a point $P = (\alpha, \beta)$ for which at least one of $\alpha$ or $\beta$ is not rational, then we obtain the field extension $F = \mathbb{Q}(\alpha, \beta)$ all of whose elements are constructible. Moreover, any point in the plane $\mathbb{R}^2$ whose coordinates lie in $F$ are constructible and we can make further constructions using these points to obtain a larger field extension. A field $K$ so constructed will be called a *constructible field* and the points of $\mathbb{R}^2$ whose coordinates lie in $K$, all of which are constructible, will be called the plane of $K$. How are the fields related?

To answer this question let $K$ be a field of constructible numbers, and $\mathbb{P}$ the plane of $K$. Let’s determine what we have to adjoin to $K$ in order to obtain the coordinates of points produced by our three basic constructions:

1. The intersection of two nonparallel lines in $\mathbb{P}$ (i.e. lines joining points whose coordinates lie in $K$),

2. The intersection of a line in $\mathbb{P}$ with a circle in $\mathbb{P}$ (i.e. a line as in 1 and a circles whose center is in $\mathbb{P}$ and whose radius is in $K$), and

3. The intersection of two circles in $\mathbb{P}$.

For 1. suppose that the lines $L_1$ and $L_2$ pass through $(a_1, b_1)$ and $(c_1, d_1)$ and $(a_2, b_2)$ and $(c_2, d_2)$ respectively. Then an equation for $L_1$ is

$$y - d_1 = \frac{(d_1 - db_1)}{(c_1 - a_1)}(x - c_1)$$

and an equation for $L_2$ is

$$y - d_2 = \frac{(d_2 - b_2)}{(c_2 - a_2)}(x - c_2).$$

They intersect where

$$d_1 + \frac{(d_1 - b_1)}{(c_1 - a_1)}(x - c_1) = d_2 + \frac{(d_2 - b_2)}{(c_2 - a_2)}(x - c_2).$$

Write this as

$$d_1 + m_1(x - c_1) = d_2 + m_2(x - c_2),$$

noting that the slopes $m_1$ and $m_2$ are in $K$. Solving for $x$, we obtain,

$$d_1 - d_2 + m_2 c_2 - m_1 c_1 = x(m_2 - m_1)$$

$$d_1 - d_2 + m_2 c_2 - m_1 c_1 = \frac{(m_2 - m_1)}{x}.$$
Exercise 5.27. Show that a line in \( \mathbb{P} \) has an equation of the form \( ax + by + c = 0 \) where \( a, b, c \) all lie in \( K \).

Exercise 5.28. Show that a circle in \( \mathbb{P} \) has an equation of the form \( x^2 + y^2 + dx + ey + f = 0 \) where \( d, e, f \) all lie in \( K \).

Exercise 5.29. Show that the coordinates of the points of intersection of a line \( L \) in \( \mathbb{P} \) with a circle \( C \) in \( \mathbb{P} \) solve a quadratic equation whose coefficients lie in \( K \). First consider the case where the equation of \( L \) is \( ax + by + c = 0 \) with \( a \neq 0 \) and then the case where \( a = 0 \).

Exercise 5.30. Given that \( C_1 \) and \( C_2 \) are intersecting circles in the plane of \( K \), suppose that \( C_1 \) has equation
\[
x^2 + y^2 + ax + by + c = 0
\]
and \( C_2 \) has equation
\[
x^2 + y^2 + dx + ey + f = 0
\]
where all of \( a, b, c, d, e, f \) lie in \( F \). If \( a \neq d \) show that

1. The \( x \) coordinates of the points of intersection have the form
\[
x = \frac{(c - f) + (b - e)y}{d - a}
\]

2. The \( y \) coordinates of the points of intersection of \( C_1 \) and \( C_2 \) solve a quadratic equation whose coefficients lie in \( K \).

Exercise 5.31. If \( a = d \), show that the \( y \) coordinates lie in \( K \) and that the \( x \) coordinates solve a quadratic equation whose coefficients lie in \( F \).

Recall that a polynomial with rational coefficients factors into linear factors if and only if all of its roots are rational numbers. The same statement holds for an arbitrary field of coefficients. The following lemma and theorem relate to the special way in which constructible numbers arise.

Lemma 5.32. Let \( F \) be a field extension of \( \mathbb{Q} \) and \( f = ax^2 + bx + c \) a quadratic polynomial with coefficients in \( F \). If \( \alpha \) is a root of \( f \), then \( [F(\alpha) : F] = 1 \) if \( f \) is reducible and \( [F(\alpha) : F] = 2 \) if \( f \) is irreducible.

Proof. Apply the quadratic formula to \( f \) to obtain that
\[
\alpha_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{and} \quad \alpha_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}
\]
are the roots of $f$. Set $d = b^2 - 4ac$ and $\delta = \sqrt{d}$. Observe that $F(\alpha_1) = F(\delta) = F(\alpha_2)$ since any field extension of $F$ that contains $\delta$ must also contain $\alpha_1$ and vice versa (similarly for $\alpha_2$). Reducibility of $f$ is equivalent with $\delta \in F$ and therefore with $F(\delta) = F$ (i.e. $[F(\delta) : F] = 1$). Assume that $f$ is irreducible. Any element of $F(\delta)$ can be written as $e + f\delta$ for some $e, f \in F$ since we can “rationalize” the denominator of any fraction just as we do when $F = \mathbb{Q}$. Thus $1, \delta$ spans $F(\delta)$. A dependence relation over $F$ between 1 and $\delta$ would result in $\delta \in F$. 

We have seen that given a constructible field $K$ and its plane $\mathbb{P}$, the coordinates of points that arise from constructions in $\mathbb{P}$ solve linear or quadratic polynomials with coefficients in $K$. This gives us a numerical criterion that distinguishes the constructible numbers:

**Theorem 5.33.** If $\alpha$ is a constructible number, then $[\mathbb{Q}(\alpha) : \mathbb{Q}]$ is a power of 2.

**Proof.** Since $\alpha$ is constructible there a sequence of basic constructions will produce the point $(\alpha, 0)$. The constructions yield a sequence of field extensions

$$\mathbb{Q} = F_0 \subseteq F_1 \subseteq F_2 \subseteq \cdots \subseteq F_n$$

with $\alpha \in F_n$ and $F_{i+1}$ obtained from $F_i$ by solving a quadratic polynomial with coefficients in $F_i$. From the lemma we know that $[F(\alpha) : F] = 1$ or 2 so that repeated applications of the Dimension Formula yield $[F_n : \mathbb{Q}] = 2^m$ for some $m \leq n$. Since $\mathbb{Q} \subseteq \mathbb{Q}(\alpha) \subseteq F_n$ another application of the Dimension Formula shows that $[\mathbb{Q}(\alpha) : \mathbb{Q}]$ must be a power of 2. □

### 5.4 Classical Construction Problems

The numerical criterion proved in Theorem 5.33 above enables us to solve all of the classical constructibility problems — the first three in the negative and the fourth through a complete description of the regular polygons which are constructible.

#### 5.4.1 Angle Trisection

While certain angles can clearly be trisected (e.g. trisection of a 180° angle was accomplished in Exercise 4.1) there is no sequence of constructions that can be applied to trisect an arbitrary angle. To see this, we prove the following result.

**Theorem 5.34.** An angle of 60° cannot be trisected by ruler and compass constructions.

**Proof.** If it were possible to trisect a 60° then it would be possible to construct the right triangle $OPQ$ below:
5.4. CLASSICAL CONSTRUCTION PROBLEMS

Since the coordinates of $P$ are $(\cos 20^\circ, \sin 20^\circ)$, these numbers would also be constructible. An exercise in trigonometry using multiple angle formulas reveals that $\alpha = \cos 20^\circ$ is a root of the cubic polynomial $f = 8x^3 - 6x - 1$. Indeed, for any angle $x$,

\[
\cos 3x = \cos(x + 2x) = \cos x \cos 2x - \sin x \sin 2x
= \cos x (\cos^2 x - \sin^2 x) - 2 \sin^2 x \cos x
= \cos^3 x - 3 \sin^2 x \cos x
= \cos^3 x - 3 \cos x (1 - \cos^2 x)
= 4 \cos^3 x - 3 \cos x.
\]

We obtain for $x = 20^\circ$ that

\[
\frac{1}{2} = 4\alpha^3 - 3\alpha
\]

from which it is clear $\alpha$ is a root of $f$.

Using the rational root test, one can show that this polynomial has no rational roots and is therefore irreducible as a polynomial with rational coefficients. If $\alpha$ was the root of a quadratic polynomial $g$ with rational coefficients, write

\[
f(x) = q(x)g(x) + r
\]

with $q$ a polynomial of degree 1 with rational coefficients and $r \in \mathbb{Q}$ the remainder. Evaluating both sides at $x = \alpha$ reveals that $r = 0$ and $f = qg$, contradicting the irreducibility of $f$. Thus $1, \alpha, \alpha^2$ are linearly independent over $\mathbb{Q}$ in $\mathbb{Q}(\alpha)$ and we find that $[\mathbb{Q}(\alpha) : \mathbb{Q}]$ is divisible by 3 (in fact it is equal to 3). By our numerical criterion, $\alpha$, and therefore the $20^\circ$ angle, are not constructible.

5.4.2 Duplicating a Cube

The problem asks whether, given a constructible cube, it is possible to construct a cube of precisely twice the volume of the given one. To effect such constructions it is necessary of course to extend our constructions from the plane to three space. This is not a serious issue though. The problem is solved in the negative by demonstrating the following result.

\[
\]
Theorem 5.35. The unit cube cannot be duplicated, i.e. a cube whose volume is equal to 2 cannot be constructed with ruler and compass.

Proof. If it were possible to construct a cube whose volume is equal to 2, then each side of the cube would have length $\sqrt[3]{2}$. The theorem is therefore proved if we demonstrate that $\sqrt[3]{2}$ is not constructible. The argument is similar to the one above. Clearly $\sqrt[3]{2}$ is a root of $x^3 - 2$, which can again be demonstrated to be irreducible by the rational root test. Moreover, $\sqrt[3]{2}$ is a root of no polynomial of degree less than three and therefore $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}]$ is divisible by 3 (in fact it is also equal to 3).

5.4.3 Squaring the Circle

At issue is whether, given a constructible circle, it is possible to construct a square with precisely the same area as the given circle. This problem is also resolved in the negative by considering the unit circle, whose area is $\pi$. A square of area equal to $\pi$ would have a side of length $\sqrt{\pi}$ so that constructibility of said square is equivalent to constructibility of $\sqrt{\pi}$. Since the constructible numbers are closed under squaring and square roots, the issue rests on the constructibility of $\pi$. A famous but rather difficult theorem asserts that $\pi$ is a transcendental number, i.e. $\pi$ is a root of no nonzero polynomial with rational coefficients. As a consequence, the infinitely many powers $1, \pi, \pi^2, \ldots$ are all linearly independent over $\mathbb{Q}$. In particular $[\mathbb{Q}(\pi) : \mathbb{Q}]$ is infinite, which is certainly not a power of 2.

5.4.4 Constructible Polygons

Consider a regular $n$-gon (i.e. polygon with $n$ sides of equal length) inscribed within the unit circle with one vertex at the point $(1, 0)$:
It is most convenient here to use radian measure for angles so that the set of its vertices is exactly \( \{ (\cos \frac{2\pi j}{n}, \sin \frac{2\pi j}{n}) : j = 0, \ldots, n - 1 \} \). Thus the regular \( n \)-gon is constructible if and only if \( \cos \frac{2\pi j}{n} \) and \( \sin \frac{2\pi j}{n} \) are constructible numbers. With \( i \) denoting as usual the complex square root of -1, we have by de Moivre’s theorem \( e^{i\theta} = \cos \theta + i\sin \theta \), so that \( (\frac{2\pi}{n})^j = \cos \frac{2\pi j}{n} + i\sin \frac{2\pi j}{n} \). It follows that the coordinates of all of the vertices lie in the field \( \mathbb{Q}(\frac{2\pi}{n}, i) \). We also have
\[
\mathbb{Q}(\cos \frac{2\pi}{n}, \sin \frac{2\pi}{n}, i) = \mathbb{Q}(\frac{2\pi}{n}, i),
\]
again from de Moivre’s theorem.

Note that
\[
[\mathbb{Q}(\cos \frac{2\pi}{n}, \sin \frac{2\pi}{n}, i) : \mathbb{Q}(\cos \frac{2\pi}{n}, \sin \frac{2\pi}{n})] = 2
\]
since \( \mathbb{Q}(\cos \frac{2\pi}{n}, \sin \frac{2\pi}{n}) \) is a subfield of the real numbers and 1, \( i \) comprise a basis for the larger field. Similarly, if \( i \) is not in the field \( \mathbb{Q}(\frac{2\pi}{n}) \) then also
\[
[\mathbb{Q}(\frac{2\pi}{n}, i) : \mathbb{Q}(\frac{2\pi}{n})] = 2.
\]

As a consequence we see that
\[
[\mathbb{Q}(\cos \frac{2\pi}{n}, \sin \frac{2\pi}{n}) : \mathbb{Q}] = \frac{[\mathbb{Q}(\cos \frac{2\pi}{n}, \sin \frac{2\pi}{n}, i) : \mathbb{Q}]}{2}
= \frac{[\mathbb{Q}(\frac{2\pi}{n}, i) : \mathbb{Q}]}{2}
= \frac{[\mathbb{Q}(\frac{2\pi}{n}, i) : \mathbb{Q}]}{2} + \frac{[\mathbb{Q}(\frac{2\pi}{n}) : \mathbb{Q}]}{2}.
\]

Thus we have proved the

**Proposition 5.36.** \([\mathbb{Q}(\cos \frac{2\pi}{n}, \sin \frac{2\pi}{n}) : \mathbb{Q}] \) is a power of 2 if and only if \([\mathbb{Q}(\frac{2\pi}{n}) : \mathbb{Q}] \) is a power of 2. In particular, if the regular \( n \)-gon is constructible, then \([\mathbb{Q}(\cos \frac{2\pi}{n}, \sin \frac{2\pi}{n}) : \mathbb{Q}] \) is a power of 2.

The field extensions \( \mathbb{Q}(\frac{2\pi}{n}) \) of \( \mathbb{Q} \), known as the cyclotomic extensions, are very well studied for their applications in number theory and other branches of mathematics. The integer \([\mathbb{Q}(\frac{2\pi}{n}) : \mathbb{Q}] \), which is denoted by \( \varphi(n) \), has great significance in number theory, some of which is investigated in later chapters. Remarkably, \( \varphi(n) \) is equal to the number of positive integers less than \( n \) whose greatest common divisor with \( n \) is equal to 1. So for instance \( \varphi(3) = \varphi(4) = \varphi(6) = 2 \), \( \varphi(5) = \varphi(8) = 4 \), \( \varphi(7) = \varphi(9) = 6 \). In general, if \( n \) is prime, then \( \varphi(n) = n - 1 \).

The contrapositive of the proposition gives us a method to show that certain \( n \)-gons are not constructible.

**Example 5.37.** The regular 9-gon is not constructible. To see this let \( \alpha = \cos \frac{2\pi}{9} \). Note that
\[
-\frac{1}{2} = \cos \frac{2\pi}{3} = \cos 3(\frac{2\pi}{9}) = 4\alpha^3 - 3\alpha.
\]
so that \( \alpha \) is a root of \( 8x^3 - 6x + 1 \). Once again the rational root test shows that this polynomial is irreducible and arguments similar to those used in the trisection problem show that \( \alpha \) is a root of no quadratic polynomial with rational coefficients. We again obtain that \([\mathbb{Q}(\alpha) : \mathbb{Q}]\) is divisible by three so that

\[
[\mathbb{Q}(\cos \frac{2\pi}{9}, \sin \frac{2\pi}{9}) : \mathbb{Q}] = [\mathbb{Q}(\cos \frac{2\pi}{9}, \sin \frac{2\pi}{9}) : \mathbb{Q}(\cos \frac{2\pi}{9})][\mathbb{Q}(\cos \frac{2\pi}{9}) : \mathbb{Q}]
\]

is not a power of 2.

**Exercise 5.38.** Use the Pythagorean theorem to show that \([\mathbb{Q}(\cos \theta, \sin \theta) : \mathbb{Q}(\cos \theta)] \leq 2\). Conclude that it suffices to consider \([\mathbb{Q}(\cos \frac{2\pi}{n}) : \mathbb{Q}]\) in the proposition.

While the proposition shows that if the regular \( n \)-gon is constructible, \( \varphi(n) \) must be a power of 2, the converse is also true, namely:

**Theorem 5.39.** \( \mathbb{Q}(e^{\frac{2\pi i}{n}}) \) is a constructible field, and therefore the regular \( n \)-gon is constructible, if and only if \( \varphi(n) \) is a power of 2.

A complete proof of this theorem is beyond the scope of this book, but using it we see that the 7-gon is not constructible, while the 17-gon is.
Chapter 6

Quotient Rings and Field Extensions

In this chapter we describe a method for producing field extension of a given field. If $F$ is a field, then a field extension is a field $K$ that contains $F$ and for which the field operations restrict to the field operation of $F$. For example, $\mathbb{C}$ is a field extension of $\mathbb{R}$ since $\mathbb{C}$ is a field containing $\mathbb{R}$ and the field operations on the real elements of $\mathbb{C}$ are precisely those of the field $\mathbb{R}$. Similarly, $\mathbb{C}$ is a field extension of $\mathbb{Q}$. For coding theory we need field extensions of $\mathbb{Z}_2$. To produce a field extension of a field $F$ we will use a polynomial $f(x)$ with coefficients in $F$, and we will produce it by mimicking the idea of producing the integers modulo $n$ by starting with the integers and a fixed integer $n$. In order to do this we need to know that the arithmetic of polynomials is sufficiently similar to the arithmetic of integers. In the first section of this chapter we see that notions relating to divisibility work just as well for polynomials over a field as for the integers.

6.1 Arithmetic of Polynomial Rings

Let $F$ be a field, and let $F[x]$ be the ring of polynomials in the indeterminate $x$. High school students study the arithmetic of this ring without saying so in so many words, at least for the case $F = \mathbb{R}$. In this section we make a formal study of this arithmetic, seeing that much of what we did for integers above can be done in the ring $F[x]$. We start with the most basic definition.

**Definition 6.1.** Let $f$ and $g$ be polynomials in $F[x]$. Then we say that $f$ divides $g$, or $g$ is divisible by $f$, if there is a polynomial $h$ with $g = fh$.

The greatest common divisor of two integers $a$ and $b$ is the largest integer dividing both $a$ and $b$. This definition needs to be modified a little for polynomials. While we cannot talk about “largest” polynomial in the same manner as for integers, we can talk about the degree of a polynomial. Recall that the degree of a nonzero polynomial $f$ is the largest integer $m$ for which the coefficient of $x^m$ is nonzero. If $f(x) = a_n x^n + \cdots + a_1 x + a_0$ with $a_n \neq 0$, then the degree of $f(x)$ is $n$. We write $\text{deg}(f)$ for the degree of $f$. The degree function allows
us to measure size of polynomials. However, there is one extra complication. For example, any polynomial of the form \( ax^2 \) with \( a \neq 0 \) divides \( x^2 \) and \( x^3 \). Thus, there isn’t a unique polynomial of highest degree that divides a pair of polynomials. To pick one out, we consider monic polynomials, polynomials whose leading coefficient is 1. For example, \( x^2 \) is the monic polynomial of degree 2 that divides both \( x^2 \) and \( x^3 \), while \( 5x^2 \) is not monic. As a piece of terminology, we will refer to an element \( f \in F[x] \) as a polynomial over \( F \).

**Definition 6.2.** Let \( f \) and \( g \) be polynomials over \( F \), not both zero. Then a greatest common divisor of \( f \) and \( g \) is a monic polynomial of largest degree that divides both \( f \) and \( g \).

The problem with the definition above has to do with uniqueness. Could there be more than one greatest common divisor of a pair of polynomials? The answer is no, and we will prove this after we prove the analogue of the division algorithm.

The main reason for assuming that the coefficients of our polynomials lie in a field is to ensure that the division algorithm is valid. Before we prove it, we need a simple lemma to make the statement in the following lemma and other results as simple as possible.

**Lemma 6.3.** Let \( F \) be a field and let \( f \) and \( g \) be polynomials over \( F \). Then \( \deg(fg) = \deg(f) + \deg(g) \).

**Proof.** If either \( f = 0 \) or \( g = 0 \), then the equality \( \deg(fg) = \deg(f) + \deg(g) \) is true by our convention above. So, suppose that \( f \neq 0 \) and \( g \neq 0 \). Write \( f = a_nx^n + \cdots + a_0 \) and \( g = b_mx^m + \cdots + b_0 \) with \( a_n \neq 0 \) and \( b_m \neq 0 \). Therefore, \( \deg(f) = n \) and \( \deg(g) = m \). The definition of polynomial multiplication yields

\[
f \cdot g = (a_n b_m)x^{n+m} + (a_n b_{m-1} + a_{n-1} b_m)x^{n+m-1} + \cdots + a_0 b_0.
\]

Now, since the coefficients come from a field, which has no zero divisors, we can conclude that \( a_n b_m \neq 0 \), and so \( \deg(fg) = n + m = \deg(f) + \deg(g) \), as desired. \( \square \)

**Proposition 6.4** (Division Algorithm). Let \( F \) be a field and let \( f \) and \( g \) be polynomials over \( F \) with \( f \) nonzero. Then there are unique polynomials \( q \) and \( r \) with \( g = qf + r \) with \( \deg(r) < \deg(f) \).

**Proof.** Let

\[
S = \{ t \in F[x] : t = g - qf \text{ for some } q \in F[x] \}.
\]

Then \( S \) is a nonempty set of polynomials, since \( g \in S \). Thus, by the well ordering property of the integers, there is a polynomial \( r \) of least degree in \( S \). By definition, there is a \( q \in F[x] \) with \( r = g - qf \), so \( g = qf + r \). We need to see that \( \deg(r) < \deg(f) \). If, on the other hand, \( \deg(r) \geq \deg(f) \), say \( n = \deg(f) \) and \( m = \deg(r) \). If \( f = a_nx^n + \cdots + a_0 \) and \( r = r_mx^m + \cdots + r_0 \) with \( a_n \neq 0 \) and \( r_m \neq 0 \), then by thinking about the method of
long division of polynomials, we realize that we may write \( r = (r_m a_n^{-1}) x^{m-n} f + r' \) with \( \deg(r') < m = \deg(r) \). But then
\[
g = q f + r = q f + (r_m a_n^{-1}) x^{m-n} f + r' = (q + r_m a_n^{-1} x^{m-n}) f + r',
\]
which shows that \( r' \in \mathcal{S} \). Since \( \deg(r') < \deg(r) \), this would be a contradiction to the choice of \( r \). Therefore, \( \deg(r) \geq \deg(f) \) is false, so \( \deg(r) < \deg(f) \), as we wanted to prove. This proves existence of \( q \) and \( r \). For uniqueness, suppose that \( q = f + r \) and \( g = q f + r' \) for some polynomials \( q, q' \) and \( r, r' \) in \( F[x] \), and with \( \deg(r), \deg(r') < \deg(f) \). Then \( q f + r = q' f + r' \), so \( (q - q') f = r' - r \). Taking degrees and using the lemma, we have
\[
\deg(q' - q) + \deg(f) = \deg(r' - r).
\]
Since \( \deg(r) < \deg(f) \) and \( \deg(r') < \deg(f) \), we have \( \deg(r' - r) < \deg(f) \). However, if \( \deg(q' - q) \geq 0 \), this is a contradiction to the equation above. The only way for this to hold is for \( \deg(q' - q) = \deg(r' - r) = -\infty \). Thus, \( q' - q = 0 = r' - r \), so \( q' = q \) and \( r' = r \), proving uniqueness.

We now prove the existence of greatest common divisors of polynomials, and also prove the representation theorem analogous to Proposition 1.12.

**Proposition 6.5.** Let \( F \) be a field and let \( f \) and \( g \) be polynomials over \( F \), not both zero. Then \( \gcd(f, g) \) exists and is unique. Furthermore, there are polynomials \( h \) and \( k \) with \( \gcd(f, g) = hf + kg \).

**Proof.** We will prove this by proving the representation result. Let
\[
\mathcal{S} = \{hf + kg : h, k \in F[x]\}.
\]
Then \( \mathcal{S} \) contains nonzero polynomials as \( f = 1 \cdot f + 0 \cdot g \) and \( g = 0 \cdot f + 1 \cdot g \) both lie in \( \mathcal{S} \). Therefore, there is a nonzero polynomial \( d \in \mathcal{S} \) of smallest degree by the well ordering principle. Write \( d = hf + kg \) for some \( h, k \in F[x] \). By dividing by the leading coefficient of \( d \), we may assume that \( d \) is monic without changing the condition \( e \in \mathcal{S} \). We claim that \( d = \gcd(f, g) \). To show that \( e \) is a common divisor of \( f \) and \( g \), first consider \( f \). By the division algorithm, we may write \( f = qd + r \) for some polynomials \( q \) and \( r \), and with \( \deg(r) < \deg(d) \). Then
\[
r = f - qd = f - q(hf + kg)
= (1 - qh)f + (-qk)g.
\]
This shows \( r \in \mathcal{S} \). If \( r \neq 0 \), this would be a contradiction to the choice of \( d \) since \( \deg(r) < \deg(d) \). Therefore, \( r = 0 \), which shows that \( f = qd \), and so \( d \) divides \( f \). Similarly, \( d \) divides \( g \). Thus, \( d \) is a common divisor of \( f \) and \( g \). If \( e \) is any other common divisor of \( f \) and \( g \), then \( e \) divides any combination of \( f \) and \( g \); in particular, \( e \) divides \( hf + kg = d \). This forces
\[ \deg(e) \leq \deg(d) \text{ by Lemma 6.3. Thus, } d \text{ is the monic polynomial of largest degree that divides } f \text{ and } g, \text{ so } d \text{ is a greatest common divisor of } f \text{ and } g. \text{ This proves everything but uniqueness. For that, suppose that } d \text{ and } d' \text{ are both monic common divisors of } f \text{ and } g \text{ of largest degree. By the proof above, we may write both } d \text{ and } d' \text{ as combinations of } f \text{ and } g. \text{ Also, from this, the argument above shows that } d \text{ divides } d' \text{ and vice-versa. If } d' = ad \text{ and } d = bd', \text{ then } d = bd' = abd. \text{ Taking degrees shows that } \deg(ab) = 0, \text{ which means that } a \text{ and } b \text{ are both constants. But, since } d \text{ and } d' \text{ are monic, for } d' = ad \text{ to be monic, } a = 1. \text{ Thus, } d' = ad = d. \] 

**Exercises**

1. Let \( F \) be a field and let \( f, g, q, r \) be polynomials in \( F[x] \) such that \( g = qf + r \). Prove that \( \gcd(f, g) = \gcd(f, r) \).

2. Let \( F \) be a field. If \( f \in F[x] \) has a multiplicative inverse in \( F[x] \), prove that \( \deg(f) = 0 \). Conversely, show that any nonzero polynomial of degree 0 in \( F[x] \) has a multiplicative inverse in \( F[x] \).

   (This problem shows that the units of \( F[x] \) are the nonzero constant polynomials.)

3. Calculate, by using the Euclidean algorithm, the greatest common divisor in \( \mathbb{R}[x] \) of \( x^5 + 5x^3 + x^2 + 4x + 1 \) and \( x^4 - x^3 - x - 1 \), and write the greatest common divisor as a linear combination of the two polynomials. You may check your work with Maple, but do the calculation by hand.

4. Calculate and express the greatest common divisor of \( x^{15} + x^{14} + x^{13} + x^{12} + x^{11} + x^{10} + x^9 + x^8 + x^7 + x^6 + x^5 + x^4 + x^3 + x^2 + x + 1 \) and \( x^9 - x^6 - x^4 - x^2 - x - 1 \) as a linear combination of the two polynomials. You are welcome and encouraged to do this in Maple; if you do so, look at the worksheet Problem4Assign6.mws.

5. Prove the **Remainder Theorem**: if \( f(x) \) is a polynomial over \( F \) and \( a \in F \), if we write \( f(x) = q(x)(x - a) + r(x) \) by the division algorithm, then \( r(x) \) is equal to the constant polynomial \( f(a) \). Conclude that \( a \) is a root of \( f \) if and only if \( x - a \) divides \( f(x) \).

   (Hint: you may use the following results: if \( f, g \) are polynomials over \( F \) and \( a \in F \), then \( (fg)(a) = f(a)g(a) \) and \( (f + g)(a) = f(a) + g(a) \).)

6. Let \( f(x) = (x - a)g(x) \) for some \( a \in F \) and some \( g \in F[x] \). If \( b \neq a \) is a root of \( f \), show that \( g(b) = 0 \).

7. Prove that a polynomial of degree \( n \) has at most \( n \) roots in \( F \).

   (Hint: use induction together with the previous problems.)

8. Let \( f \in F[x] \) be a polynomial of degree 2 or 3. Prove that if \( f \) can be factored nontrivially as \( f = gh \), then \( f \) has a root in \( F \).
9. Give an example of a polynomial $f \in \mathbb{R}[x]$ of degree 4 that can be factored nontrivially as $f = gh$ but for which $f$ does not have a root in $\mathbb{R}$.

10. **Rational Root Test**: let $f(x) = a_n x^n + \cdots + a_0$ be a polynomial with integer entries. If $r = b/c$ is a rational root of $f$ written in reduced form, show that $b$ divides $a_0$ and $c$ divides $a_n$.

   (Hint: clear denominators in the equation $f(b/c) = 0$, then move the constant term to one side and the rest on the other, and factor. This will lead you to the first statement.)

11. Factor completely and determine the roots of the polynomial $x^5 - 8x^4 - 4x^3 + 76x^2 - 5x + 84$ as

   (a) a polynomial over $\mathbb{R}$;
   
   (b) a polynomial over $\mathbb{C}$;
   
   (c) a polynomial over $\mathbb{Z}_2$.

   (Feel free to do this with Maple; the file Problem7Assign7.mws has some information about factoring polynomials over different fields.)

### 6.2 Ideals and Quotient Rings

We will construct extension fields of a field $F$ by starting with an ideal of the polynomial ring $F[x]$ and constructing the associated quotient ring. We must therefore begin by defining ideals.

**Definition 6.6.** Let $R$ be a ring. An ideal $I$ is a nonempty subset of $R$ such that (i) if $a, b \in I$, then $a + b \in I$, and (ii) if $a \in I$ and $r \in R$, then $ar \in I$ and $ra \in I$.

This definition says that an ideal is a subset of $R$ closed under addition that satisfies a strengthened form of closure under multiplication. Not only is the product of two elements of $I$ also in $I$, but that the product of an element of $I$ and any element of $r$ is an element of $I$.

**Example 6.7.** Let $R = \mathbb{Z}$. If $n$ is an integer, let $n\mathbb{Z}$ be the set of all multiples of $\mathbb{Z}$. That is, $n\mathbb{Z} = \{na : a \in \mathbb{Z}\}$.

To see that this set is an ideal, first consider addition. if $x, y \in n\mathbb{Z}$, then there integers $a$ and $b$ with $x = na$ and $y = nb$. Then $x + y = na + nb = n(a + b)$. Therefore, $x + y \in n\mathbb{Z}$.

Second, for multiplication, let $x = na \in n\mathbb{Z}$ and let $r \in \mathbb{Z}$. Then $rx = x r = r(na) = n(ra)$, a multiple of $n$. Therefore, $rx \in n\mathbb{Z}$. This proves that $n\mathbb{Z}$ is an ideal. If $n > 0$, notice that $n\mathbb{Z} = \{0, n, 2n, \ldots, -n, -2n, \ldots\}$ is the same as the equivalence class of 0 under the relation congruence modulo $n$. This is an important connection that we will revisit.
Example 6.8. Let $R = F[x]$ be the ring of polynomials over a field, and let $f \in F[x]$. Let
$$I = \{gf : g \in F[x]\},$$
the set of all multiples of $f$. This set is an ideal of $F[x]$ by the same calculation as in the previous example. However, we repeat this calculation. For closure under addition, let $h, k \in I$. Then $h = gf$ and $k = g'f$ for some polynomials $g$ and $g'$. Then $h + k = gf + g'f = (g + g')f$, a multiple of $f$, so $h + k \in I$. For multiplication, let $h = gf \in I$, and let $a \in F[x]$. Then $ah = ha = agf = (ag)f$, a multiple of $f$, so $ah \in I$. Thus, $I$ is an ideal of $F[x]$. This ideal is typically denoted by $(f)$.

Example 6.9. Let $R$ be any commutative ring, and let $a \in R$. Let
$$aR = \{ar : r \in R\}.$$ We can consider $aR$ to be the set of multiples of $a$. We show that $aR$ is an ideal of $R$. First, let $x, y \in aR$. Then $x = ar$ and $y = as$ for some $r, s \in R$. Then $x + y = ar + as = a(r + s)$, so $x + y \in aR$. Next, let $x = ar \in aR$ and let $z \in R$. Then $xz = arz = a(rz) \in aR$. Also, $zx = xz$ since $R$ is commutative, so $zx \in aR$. Therefore, $aR$ is an ideal of $R$. This construction generalizes the previous two examples. The ideal $aR$ is typically called the ideal generated by $a$, or the principal ideal generated by $a$. It is often written as $(a)$.

Example 6.10. Let $R$ be any commutative ring, and let $a, b \in R$. Set
$$I = \{ar + bs : r, s \in R\}.$$ To see that $I$ is an ideal of $R$, first let $x, y \in I$. Then $x = ar + bs$ and $y = ar' + bs'$ for some $r, s, r', s' \in R$. Then
$$x + y = (ar + bs) + (ar' + bs') = (ar + ar') + (bs + bs') = a(r + r') + b(s + s') \in I$$
by the associative and distributive properties. Next, let $x \in I$ and $z \in R$. Again, $x = ar + bs$ for some $r, s \in R$. Then
$$xz = (ar + bs)z = (ar)z + (bs)z = a(rz) + b(sz).$$ This calculation shows that $xz \in I$. Again, since $R$ is commutative, $zx = xz$, so $zx \in I$. Thus, $I$ is an ideal of $R$. We can generalize this example to any finite number of elements of $R$: given $a_1, \ldots, a_n \in R$, if
$$J = \{a_1r_1 + \cdots + a_nr_n : r_i \in R \text{ for each } i\},$$
then a similar argument will show that $J$ is an ideal of $R$. The ideal $J$ is typically referred to as the ideal generated by the elements $a_1, \ldots, a_n$, and it is often denoted by $(a_1, \ldots, a_n)$.  

The division algorithm has a nice application to the structure of ideals of $\mathbb{Z}$ or of $F[x]$. We prove the result for polynomials, leaving the analogous result for $\mathbb{Z}$ to the reader.

**Theorem 6.11.** Let $F$ be a field. Then any ideal of $F[x]$ can be generated by a single polynomial. That is, if $I$ is an ideal of $F[x]$, then there is a polynomial $f$ with $I = (f) = \{fg : g \in F[x]\}$.

**Proof.** Let $I$ be an ideal of $F[x]$. If $I = \{0\}$, then $I = (0)$. So, suppose that $I$ is nonzero. Let $f \in I$ be a nonzero polynomial of least degree. We claim that $I = (f)$. To prove this, let $g \in I$. By the division algorithm, there are polynomials $q, r$ with $g = qf + r$ and $\deg(r) < \deg(f)$. Since $f \in I$, the product $qf \in I$, and thus $g - qf \in I$ as $g \in I$. We conclude that $r \in I$. However, the assumption on the degree of $f$ shows that the condition $\deg(r) < \deg(f)$ forces $r = 0$. Thus, $g = qf \in (f)$. This proves $I \subseteq (f)$. Since every multiple of $f$ is in $I$, the reverse inclusion $(f) \subseteq I$ is also true. Therefore, $I = (f)$. \qed

We can give an ideal theoretic description of greatest common divisors in $\mathbb{Z}$ and in $F[x]$. Suppose that $f$ and $g$ are polynomials over a field $F$. If $\gcd(f, g) = d$, then we have proved that $d = fh + gk$ for some polynomials $h, k$. Therefore, $d$ is an element of the ideal $I = \{fs + gt : s, t \in F[x]\}$. However, since $d$ divides $f$ and $g$, it follows that $d$ divides every element of $I$. Therefore, $I = (d)$ is simply the set of multiples of $d$. Therefore, one can identify the greatest common divisor of $f$ and $g$ by identifying a monic polynomial $d$ satisfying $I = (d)$.

We now use ideals to define quotient rings. In order to define them, we first need to specify what are their elements. These are cosets, which we now define. We have seen cosets when we discussed decoding with the Hamming code. These cosets arose from a subspace of a vector space. The idea here is essentially the same; the only difference is that we start with an ideal of a ring instead of a subspace of a vector space.

**Definition 6.12.** Let $R$ be a ring and let $I$ be an ideal of $R$. If $a \in R$, then the coset $a + I$ is defined as $a + I = \{a + x : x \in I\}$.

Recall the description of equivalence classes for the relation congruence modulo $n$. For example, if $n = 5$, then we have five equivalence classes, and they are

\[
\begin{align*}
\bar{0} &= \{0, 5, 10, \ldots, -5, -10, \ldots\}, \\
\bar{1} &= \{1, 6, 11, \ldots, -4, -9, -14, \ldots\}, \\
\bar{2} &= \{2, 7, 12, \ldots, -3, -8, -13, \ldots\}, \\
\bar{3} &= \{3, 8, 13, \ldots, -2, -7, -12, \ldots\}, \\
\bar{4} &= \{4, 9, 14, \ldots, -1, -6, -11, \ldots\}.
\end{align*}
\]

By the first example above, the set $5\mathbb{Z}$ of multiples of 5 forms an ideal of $\mathbb{Z}$. These five
equivalence classes can be described as cosets, namely,
\[
\bar{0} = 0 + 5\mathbb{Z}, \\
\bar{1} = 1 + 5\mathbb{Z}, \\
\bar{2} = 2 + 5\mathbb{Z}, \\
\bar{3} = 3 + 5\mathbb{Z}, \\
\bar{4} = 4 + 5\mathbb{Z}.
\]
In general, for any integer \(a\), we have \(a + 5\mathbb{Z} = \bar{a}\). Thus, cosets for the ideal \(5\mathbb{Z}\) are the same as equivalence classes modulo 5. In fact, more generally, if \(n\) is any positive integer, then the equivalence class \(\bar{a}\) of an integer \(a\) modulo \(n\) is the coset \(a + n\mathbb{Z}\) of the ideal \(n\mathbb{Z}\).

We have seen that an equivalence classes can have different names. Modulo 5, we have \(1 = 6\) and \(2 = 3 = 22\), for example. Similarly, cosets can be represented in different ways.

If \(R = F[x]\) and \(I = xR\), the ideal of multiples of the polynomial \(x\), then \(0 + I = x + I = x^2 + I = 4x^{17} + I\). Also, \(1 + I = (x + 1) + I\). For some terminology, we refer to \(a\) as a coset representative of \(a + I\). One important thing to remember is that the coset representative is not unique, as the examples above demonstrate.

When we defined operations on \(\mathbb{Z}_n\), we defined them with the formulas \(\bar{a} + \bar{b} = a + b\) and \(\bar{a} \cdot \bar{b} = ab\). Since these equivalence classes are the same thing as cosets for \(n\mathbb{Z}\), this leads us to consider a generalization. If we replace \(\mathbb{Z}\) by any ring and \(n\mathbb{Z}\) by any ideal, we can mimic these formulas to define operations on cosets. First, we give a name to the set of cosets.

**Definition 6.13.** If \(I\) is an ideal of a ring \(R\), let \(R/I\) denote the set of cosets of \(I\). In other words, \(R/I = \{a + I : a \in R\}\).

We now define operations on \(R/I\) in a manner like the operations on \(\mathbb{Z}_n\). We define
\[
(a + I) + (b + I) = (a + b) + I, \\
(a + I) \cdot (b + I) = (ab) + I.
\]
In other words, to add or multiply two cosets, first add or multiply their coset representatives, then take the corresponding coset. As with the operations on \(\mathbb{Z}_n\), we have to check that these formulas make sense. In other words, the name we give to a coset should not affect the value we get when adding or multiplying. We first need to know when two elements represent the same coset. To help with the proof, we point out two simple properties. If \(I\) is an ideal, then \(0 \in I\). Furthermore, if \(r \in I\), then \(-r \in I\). The proofs of these facts are left as exercises.

**Lemma 6.14.** Let \(I\) be an ideal of a ring \(R\). If \(a, b \in R\), then \(a + I = b + I\) if and only if \(a - b \in I\).

**Proof.** Let \(a, b \in R\). First suppose that \(a + I = b + I\). From \(0 \in I\) we get \(a = a + 0 \in a + I\), so \(a \in b + I\). Therefore, there is an \(x \in I\) with \(a = b + x\). Thus, \(a - b = x \in I\). Conversely, suppose that \(a - b \in I\). If we set \(x = a - b\), an element of \(I\), then \(a = b + x\). This shows
a \in b + I$. So, for any $y \in I$, we have $a + y = b + (x + y) \in I$, as $I$ is closed under addition. Therefore, $a + I \subseteq b + I$. The reverse inclusion is similar; by using $-x = b - a$, again an element of $I$, we will get the inclusion $b + I \subseteq a + I$, and so $a + I = b + I$. \qed

In fact, we can generalize the fact that equivalence classes modulo $n$ are the same thing as cosets for $n\mathbb{Z}$. Given an ideal, we can define an equivalence relation by mimicking congruence modulo $n$. To phrase this relation in a new way, $a \equiv b \mod n$ if and only if $a - b$ is a multiple of $n$, so $a \equiv b \mod n$ if and only if $a - b \in n\mathbb{Z}$. Thus, given an ideal $I$ of a ring $R$, we may define a relation by $x \equiv y \mod I$ if $x - y \in I$. One can prove in the same manner as for congruence modulo $n$ that this is an equivalence relation, and that, for any $a \in R$, the coset $a + I$ is the equivalence class of $a$.

**Lemma 6.15.** Let $I$ be an ideal of a ring $R$. Let $a, b, c, d \in R$.

1. If $a + I = c + I$ and $b + I = d + I$, then $a + b + I = c + d + I$.
2. If $a + I = c + I$ and $b + I = d + I$, then $ab + I = cd + I$.

**Proof.** Suppose that $a, b, c, d \in R$ satisfy $a + I = c + I$ and $b + I = d + I$. To prove the first statement, by the lemma we have elements $x, y \in I$ with $a - c = x$ and $b - d = y$. Then

\[
(a + b) - (c + d) = a + b - c - d = (a - c) + (b - d) = x + y \in I .
\]

Therefore, again by the lemma, $(a + b) + I = (c + d) + I$. For the second statement, we rewrite the equations above as $a = c + x$ and $b = d + y$. Then

\[
ab = (c + x)(d + y) = c(d + y) + x(d + y) = cd + (cy + xd + xy).
\]

Since $x, y \in I$, the three elements $cy, xd, xy$ are all elements of $I$. Thus, the sum $cy + xd + xy \in I$. This shows us that $ab - cd \in I$, so the lemma yields $ab + I = cd + I$. \qed

The consequence of the lemma is exactly that our coset operations make sense. Thus, we can ask whether or not $R/I$ is a ring. The answer is yes, and the proof is easy, and is exactly parallel to the proof for $\mathbb{Z}_n$.

**Theorem 6.16.** Let $I$ be an ideal of a ring $R$. Then $R/I$, together with the operations of coset addition and multiplication, forms a ring.
Proof. We have several properties to verify. Most follow immediately from the definition of the operations and from the ring properties of \( R \). For example, to prove that coset addition is commutative, we see that for any \( a, b \in R \), we have
\[
(a + I) + (b + I) = (a + b) + I = (b + a) + I = (b + I) + (a + I).
\]
This used exactly the definition of coset addition and commutativity of addition in \( R \). Most of the other ring properties hold for similar reasons, so we only verify those that are a little different. For existence of an additive identity, we have the additive identity \( 0 \) of \( R \), and it is natural to guess that \( 0 + I \) is the identity for \( R/I \). To see that this is indeed true, let \( a + I \in R/I \). Then
\[
(a + I) + (0 + I) = (a + 0) + I = a + I.
\]
Thus, \( 0 + I \) is the additive identity for \( R/I \). Similarly \( 1 + I \) is the multiplicative identity, since
\[
(a + I) \cdot (1 + I) = (a \cdot 1) + I = a + I
\]
and
\[
(1 + I) \cdot (a + I) = (1 \cdot a) + I = a + I
\]
for all \( a + I \in R/I \). Finally, the additive inverse of \( a + I \) is \( -a + I \) since
\[
(a + I) + (-a + I) = (a + (-a)) + I = 0 + I.
\]
Therefore, \( R/I \) is a ring.

The ring \( R/I \) is called a quotient ring of \( R \). This idea allows us to construct new rings from old rings. For example, the ring \( \mathbb{Z}_n \) is really the same thing as the quotient ring \( \mathbb{Z}/n\mathbb{Z} \), since we have identified the equivalence classes modulo \( n \); that is, the elements of \( \mathbb{Z}_n \), with the cosets of \( n\mathbb{Z} \); i.e., the elements of \( \mathbb{Z}/n\mathbb{Z} \). It is this construction applied to polynomial rings that we will use to build extension fields. We recall Proposition 3.27 above that says \( \mathbb{Z}_n \) is a field if and only if \( n \) is a prime. To generalize this result to polynomials, we first need to define the polynomial analogue of a prime number.

Definition 6.17. Let \( F \) be a field. A nonconstant polynomial \( f \in F[x] \) is said to be irreducible over \( F \) if whenever \( f \) can be factored as \( f = gh \), then either \( g \) or \( h \) is a constant polynomial.

Before we give some examples, recall that a constant polynomial is simply a polynomial of degree 0; that is, it is a polynomial of the form \( f(x) = a \) for some \( a \in F \). Any such polynomial has degree 0 if it is not the zero polynomial.
Example 6.18. The terminology irreducible over \( F \) in the definition above is used because irreducibility is a relative term. The polynomial \( x^2 + 1 \) factors over \( \mathbb{C} \) as \( x^2 + 1 = (x - i)(x + i) \). However, we show that \( x^2 + 1 \) is irreducible over \( \mathbb{R} \). One way to do this would be to write \( x^2 + 1 = (ax + b)(cx + d) \), and obtain a system of nonlinear equations in \( a, b, c, d \), and show there is no solution to this system. However, we do it in an easier way, although one that uses a fact left as an exercise. Since \( \deg(x^2 + 1) = 2 \), if it factors over \( \mathbb{R} \), then it must have a root in \( \mathbb{R} \). However, \( x^2 + 1 \) clearly has no roots in \( \mathbb{R} \). Thus, \( x^2 + 1 \) is irreducible over \( \mathbb{R} \).

Example 6.19. The polynomial \( x \) is irreducible. For, if we can factor \( x = gh \), taking degrees of both sides gives \( 1 = \deg(g) + \deg(h) \). Thus, one of the degrees of \( g \) and \( h \) is 1 and the other is 0. The one with degree 0 is a constant polynomial. Thus, we cannot factor \( x \) with both factors nonconstant. So, \( x \) is irreducible. This argument shows that any polynomial of degree 1 is irreducible.

Example 6.20. Consider \( x^2 + 1 \) as a polynomial in \( \mathbb{Z}_5[x] \). Unlike the case of \( \mathbb{Q}[x] \), this polynomial does factor over \( \mathbb{Z}_5 \), since \( x^2 + 1 = (x - 2)(x - 3) \) in \( \mathbb{Z}_5[x] \). In particular, \( x^2 + 1 \) has two roots in \( \mathbb{Z}_5 \). However, for \( F = \mathbb{Z}_3 \), the polynomial \( x^2 + 1 \) is irreducible since \( x^2 + 1 \) has no roots in \( \mathbb{Z}_3 \); it is easy to see that none of the three elements \( 0, 1, \) and \( 2 \) are roots of \( x^2 + 1 \).

We now show that extension fields can be produced from irreducible polynomials.

Proposition 6.21. Let \( F \) be a field, and let \( f \in F[x] \) be a polynomial. If \( I = (f) \) is the ideal generated by an irreducible polynomial \( f \), then \( F[x]/I \) is a field.

Proof. Let \( F \) be a field, and let \( f \in F[x] \) be irreducible. Set \( I = (f) \). We wish to prove that \( F[x]/I \) is a field. We know it is a commutative ring, so we only need to prove that every nonzero element has a multiplicative inverse. Let \( g + I \in F[x]/I \) be nonzero. Then \( g + I \neq 0 + I \), so \( g \notin I \). This means \( f \) does not divide \( g \). Since \( f \) is irreducible, we can conclude that \( \gcd(f, g) = 1 \). Thus, there are \( h, k \in F[x] \) with \( 1 = hf + kg \). Because \( hf \in I \), \( kg - 1 \in I \), so \( kg + I = 1 + I \). By the definition of coset multiplication, this yields \( (k + I)(g + I) = 1 + I \). Therefore, \( k + I \) is the multiplicative inverse of \( g + I \). Because we have proved that an arbitrary nonzero element of \( F[x]/I \) has a multiplicative inverse, this commutative ring is a field.

The converse of this result is also true: if \( F[x]/(f) \) is a field, then \( f \) is an irreducible polynomial. We leave the verification of this fact to an exercise.

To help work with these quotient rings, we see how the division algorithm can help us write elements of \( F[x]/(f) \). Set \( I = (f) \). Given \( g \in F[x] \), by the division algorithm we may write \( g = qf + r \) for some \( g, r \in F[x] \) and with \( \deg(r) < \deg(f) \). Then \( g - r = qf \in I \), so \( g + I = r + I \). This argument shows that any coset \( g + I \) is equal to a coset \( r + I \) for some polynomial \( r \) with \( \deg(r) < \deg(f) \). Thus, \( F[x]/(f) = \{r + I : r \in F[x], \deg(r) < \deg(f)\} \).

This result is the analogue of the description \( \mathbb{Z}_n = \{a : 0 \leq a < n\} = \{a + (n) : 0 \leq a < n\} \).
Example 6.22. Let $F = \mathbb{R}$, and consider the irreducible polynomial $f = x^2 + 1$. In this example we will relate the field $\mathbb{R}[x]/(x^2 + 1)$ to the field of complex numbers $\mathbb{C}$. As in the previous example, the division algorithm implies that every element of this quotient ring can be written in the form $a + bx + I$, where $I = (x^2 + 1)$. Addition in this ring is given by

$$(a + bx + I) + (c + dx + I) = (a + c) + (b + d)x + I.$$ 

For multiplication, we have

$$(a + bx + I)(c + dx + I) = (a + bx)(c + dx) + I = ac + bd x^2 + (ad + bc)x + I = (ac - bd) + (ad + bc)x + I;$$

the simplification in the last equation comes from the equation $bdx^2 + I = -bd + I$. Since $x^2 + 1 \in I$, we have $x^2 + I = -1 + I$, so multiplying both sides by $bd + I$ yields this equation. If you look at these formulas for the operations in $\mathbb{R}[x]/(x^2 + 1)$, you may see a similarity between the operations on $\mathbb{C}$:

$$(a + bi) + (c + di) = (a + c) + (b + d)i$$

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i.$$ 

In fact, one can view this construction as a way of building the complex numbers from the real numbers and the polynomial $x^2 + 1$.

Example 6.23. Let $F = \mathbb{Z}_2$, and consider $f = x^2 + x + 1$. This is an irreducible over $\mathbb{Z}_2$ since it is quadratic and has no roots in $\mathbb{Z}_2$; the only elements of $\mathbb{Z}_2$ are 0 and 1, and neither is a root. Consider $K = \mathbb{Z}_2[x]/(x^2 + x + 1)$. This is a field by the previous proposition. We write out an addition and multiplication table for $K$ once we write down all elements of $K$. First, by the comment above, any coset in $K$ can be represented by a polynomial of the form $ax + b$ with $a, b \in \mathbb{Z}_2$; this is because any remainder after division by $f$ must have degree less than $\deg(f) = 2$. So,

$$K = \{0 + I, 1 + I, x + I, 1 + x + I\}.$$ 

Thus, $K$ is a field with 4 elements. The following tables then represent addition and multi-
If you look closely at these tables, you may see a resemblance between them and the tables of Example 3.28 above. In fact, if you label \( x + I \) as \( a \) and \( 1 + x + I \) as \( b \), along with \( 0 = 0 + I \) and \( 1 = 1 + I \), the tables in both cases are identical. In fact, the tables of Example 3.28 were found by building \( K \), and then labeling the elements as \( 0, 1, a, \) and \( b \) in place of \( 0 + I, 1 + I, x + I, \) and \( 1 + x + I \).

Example 6.24. Let \( f(x) = x^3 + x + 1 \). Then this polynomial is irreducible over \( \mathbb{Z}_2 \). To see this, we first note that it has no roots in \( \mathbb{Z}_2 \) as \( f(0) = f(1) = 1 \). Since \( \deg(f) = 3 \), if it factored, then it would have a linear factor, and so a root in \( \mathbb{Z}_2 \). Since this does not happen, it is irreducible. We consider the field \( K = \mathbb{Z}_2[x]/(x^3 + x + 1) \). We write \( I = (x^3 + x + 1) \) and set \( \alpha = x + I \). We first note an interesting fact about this field; every nonzero element of \( K \) is a power of \( \alpha \). First of all, we have

\[
K = \{ 0 + I, 1 + I, x + I, (x + 1) + I, x^2 + I, (x^2 + 1) + I, (x^2 + x) + I, (x^2 + x + 1) + I \}.
\]

We then see that

\[
\begin{align*}
\alpha &= x + I, \\
\alpha^2 &= x^2 + I, \\
\alpha^3 &= x^3 + I = (x + 1) + I = \alpha + 1, \\
\alpha^4 &= x^4 + I = x(x + 1) + I = (x^2 + x) + I = \alpha^2 + \alpha, \\
\alpha^5 &= x^5 + I = x^3 + x^2 + I = (x^2 + x + 1) + I = \alpha^2 + \alpha + 1, \\
\alpha^6 &= x^6 + I = (\alpha^3)^2 = (x^2 + 1) + I = \alpha^2 + 1, \\
\alpha^7 &= x^7 + I = x(x^2 + 1) + I = x^3 + x + I = 1 + I.
\end{align*}
\]

To obtain these equations we used several calculational steps. For example, we used the definition of coset multiplication. For instance, \( \alpha^2 = (x + I)^2 = x^2 + I \) from this definition.
Next, for \( \alpha^3 = x^3 + I \), since \( x^3 + x + 1 \in I \), we have \( x^3 + I = (x+1) + I \). For other equations, we used combinations of these ideas. For example, to simplify \( \alpha^5 = x^5 + I \), first note that

\[
\alpha^5 = \alpha^3 \cdot \alpha^2 = ((x+1) + I)(x^2 + I) \\
= x^3 + x^2 + I \\
= (x^2 + x + 1) + I
\]

since \( x^3 + x + 1 \in I \).

**Exercises**

1. Let \( F \) be a field, let \( f \in F[x] \), and let \( I = (f) \) be the ideal generated by \( f \).

   (a) Show that, for any \( g \in F[x] \), if \( g = qf + r \) with \( \deg(r) < \deg(f) \), then \( g + I = r + I \). Conclude that any coset \( g + I \) can be represented by a polynomial of degree less than \( \deg(f) \).

   (b) If \( \deg(g) < \deg(f) \) and \( \deg(h) < \deg(f) \), show that \( g + I = h + I \) if and only if \( g = h \).

2. Determine all of the cosets of the ideal \((x^2 + x + 1)\) in the ring \( \mathbb{Z}_2[x] \). Instead of writing out all the elements in a coset (for which there are infinitely many), just find a coset representative for each coset and say why you have produced all cosets.

   (Hint: use the previous problem.)

3. Determine all of the cosets of the ideal \((x^3 + x + 1)\) in the ring \( \mathbb{Z}_2[x] \).

4. Write out the elements of \( F \) as cosets of polynomials of degree at most 2, and then write out the addition and multiplication tables for \( F \). Feel free to take advantage of the commands plus and mult defined in the worksheet.

5. Write out the powers \((\pi)^i\) for \( 1 \leq i \leq 7 \) as cosets of polynomials of degree at most 2. Feel free to use the powers command defined in the worksheet.

6. Express the following cosets as cosets of polynomials of degree at most 2.

   (a) \( \overline{x^8} \)

   (b) \( \overline{x^5 + x + 1} \)

   (c) \( \overline{x^4 + x^3 + x^2} \)

   (d) \( \overline{x^6 + x} \).
Chapter 7

Cyclic Codes

In this chapter we will build codes from quotient rings of $\mathbb{Z}_2[x]$. One advantage of this construction will be that we can guarantee a certain degree of error correction. We first make a connection between words and elements of such a quotient ring.

7.1 Introduction to Cyclic Codes

Let $n$ be a positive integer. If $f(x) \in \mathbb{Z}_2[x]$ is a polynomial of degree $n$, then by the division algorithm, each element of the quotient ring $\mathbb{Z}_2[x]/(f(x))$ can be represented uniquely as the coset of a polynomial of degree less than $n$. In other words, if $g(x) \in \mathbb{Z}_2[x]$, then we may write $g(x) = q(x)f(x) + r(x)$ with $\deg(r(x)) < \deg(f(x)) = n$. Then $g(x) + (f(x)) = r(x) + (f(x))$; this yields the desired representation for the coset $g(x) + (f(x))$. Uniqueness follows from the division algorithm. However, we recall the idea. If $r(x) + (f(x)) = r'(x) + (f(x))$ with $\deg(r), \deg(r') < n$, then $r(x) + r'(x) \in (f(x))$. Thus, $f(x)$ divides $r(x) + r'(x)$. However, this forces $\deg(r + r') \geq n$ unless $r + r' = 0$. Since $\deg(r + r') \not\equiv n$, we must conclude that $r + r' = 0$, so $r' = r$. Therefore, there is a 1-1 correspondence between elements of $\mathbb{Z}_2[x]/(f(x))$ and polynomials of degree $< n$. On the other hand, if $\mathbb{Z}_2[x]_n$ is the set of polynomials of degree $< n$, then identifying a polynomial with its coefficients provides a 1-1 correspondence between $\mathbb{Z}_2^n$ and $\mathbb{Z}_2[x]_n$ given by $a_0 \cdots a_{n-1} \mapsto a_0 + a_1x + \cdots + a_{n-1}x^{n-1}$.

We will use the polynomial $f(x) = x^n + 1$ in our discussion below; we will see soon why this is a useful choice. We then view $\mathbb{Z}_2[x]/(x^n + 1)$ as the set of words of length $n$.

Define a map $\pi : \mathbb{Z}_2^n \to \mathbb{Z}_2^n$ by $\pi(a_1 \cdots a_n) = a_na_1 \cdots a_{n-1}$. This is the (right) shift function. It shifts the components of a vector to the right, and moves the last component to the front. In terms of the representation of $\mathbb{Z}_2^n$ as $\mathbb{Z}_2[x]/(x^n + 1)$, the map $\pi$ is given by

$$\pi(a_0 + \cdots + a_{n-1}x^{n-1} + (x^n + 1)) = a_{n-1} + a_0x + \cdots + a_{n-2}x^{n-1} + (x^n + 1).$$

Definition 7.1. Let $C$ be a linear code of length $n$. Then $C$ is a cyclic code if for every $v \in C$, we have $\pi(v) \in C$. 

103
Cyclic codes with good error correcting and decoding properties arise naturally from quotients of polynomial rings.

Example 7.2. The code \( \{00000, 11111\} \) is a cyclic code.

Example 7.3. The Hamming code is the linear code consisting of the vectors

\[
C = \{0111100, 0100101, 0010011, 1001110, 1101011, 1110011, 0110011, 1000110, 1010101, 1001100, 1110000, 0000000, 1101001\}.
\]

A quick inspection shows that \( 0111100 \in C \) but \( \pi(0111100) = 0011110 \not\in C \). Therefore, \( C \) is not cyclic.

Example 7.4. Recall that the Hamming code is the nullspace of

\[
H = \begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0
\end{pmatrix}.
\]

By rearranging the columns of \( H \), we can produce a cyclic code. Let

\[
A = \begin{pmatrix}
1 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 1
\end{pmatrix}.
\]

Then the nullspace of \( A \) is

\[
C = \{1010001, 0110100, 1101000, 0010111, 1000110, 0100011, 1111111, 0001101, 1001011, 1100101, 1110010, 0111001, 1110000\}.
\]

A calculation will show that this code is cyclic. One way to verify this is to note that if \( v \) is the first vector listed, then \( \pi(v), \pi^2(v), \ldots, \pi^6(v) \) are all in the code. Note that \( \pi^7(v) = v \).

Also, if \( w \) is the fourth vector, then \( \pi(w), \ldots, \pi^6(w) \), \( \pi^7(w) = w \) are also in the code, and are all different than the 7 vectors \( v, \ldots, \pi^6(v) \). Finally, the two remaining vectors are \( 0000000 \) and \( 1111111 \), which are invariant under \( \pi \). The description

\[
C = \{\pi^i(v) : 0 \leq i \leq 6\} \cup \{\pi^i(w) : 0 \leq i \leq 6\} \cup \{0000000, 1111111\}.
\]

together with a short proof shows that \( C \) is cyclic. Admittedly, this ad-hoc demonstration that the modified Hamming code is cyclic is not very natural. A closer examination of the connection between quotients of polynomial rings and cyclic codes will clarify this issue.

The utility of the polynomial \( x^n + 1 \) emerges in the next proposition. We will identify \( \mathbb{Z}_2^n \) with \( \mathbb{Z}_2[x]/(x^n + 1) \), and denote the ideal \( (x^n + 1) \) by \( I \).

Proposition 7.5. Under the identification of \( \mathbb{Z}_2^n = \mathbb{Z}_2[x]/I \), the map \( \pi \) is given by \( \pi(f(x) + I) = xf(x) + I \).
Thus, for \( n \geq 3 \), we have \( a_n - 1 \in I \), so
\[
a_{n-1} + a_0 x + \cdots + a_{n-2} x^{n-2} + I = (a_{n-1} + a_0 x + \cdots + a_{n-2} x^{n-1} + I) + a_{n-1} x^n + I
\]
\[
= a_0 x + \cdots + a_{n-2} x^{n-2} + a_{n-1} x^n + I
\]
\[
= x f(x) + I.
\]

\( \square \)

**Theorem 7.6.** Let \( C \) be a cyclic code of length \( n \). Then \( C \) is an ideal of \( \mathbb{Z}_2[x]/I \). Conversely, any ideal of \( \mathbb{Z}_2[x]/I \) is a cyclic code.

**Proof.** Since \( C \) is closed under addition, to show that \( C \) is an ideal it suffices to show that for \( f + I \in C \) and \( g + I \in \mathbb{Z}_2[x]/I \), \( g f + I \in C \). Writing \( g = \sum_{i=0}^{n-1} a_i x^i \), and using the linearity of \( C \), it suffices to show that \( x^i f + I \in C \) for each \( 0 \leq i \leq n - 1 \). By the proposition, \( x^i f + I = \pi^i (f + I) \) which lies in \( C \) by our assumption that \( C \) is cyclic.

Conversely, if \( C \) is an ideal of \( \mathbb{Z}_2[x]/I \), then \( C \) is a \( \mathbb{Z}_2 \) vector subspace of \( \mathbb{Z}_2[x]/I \) hence a linear code. But closure under multiplication by \( x + I \) and the proposition shows that \( C \) is also cyclic.

The following theorem and its corollaries are instances of the “isomorphism theorems” important throughout modern algebra.

**Theorem 7.7.** Let \( R \) be a commutative ring and \( I \) an ideal of \( R \). Then the ideals of \( R/I \) are in one-to-one correspondence with the ideals of \( R \) which contain \( I \). The correspondence is given as follows: For \( J \) an ideal of \( R \) containing \( I \), the corresponding ideal of \( R/I \) is \( J/I = \{ a + I : a \in J \} \). For \( \overline{J} \) an ideal of \( R/I \), the corresponding ideal of \( R \) is \( \{ a \in R : a + I \in \overline{J} \} \).

**Remark 7.8.** Three things must be shown:

1. Given \( J \) an ideal of \( R \) containing \( I \), \( J/I \) is an ideal of \( R/I \).
2. Given \( \overline{J} \) an ideal of \( R/I \), \( J = \{ a \in R : a + I \in \overline{J} \} \) is an ideal of \( R \).
3. The processes of 1. and 2. are inverse to each other.

**Proof.** For 1. observe that for \( a, b \in J \), \( (a + I) + (b + I) = (a + b) + I \in J/I \) because \( J \) is closed under addition. Similarly, for \( c + I \in R/I \), \( (c + I)(a + I) = ca + I \in J/I \).

For 2. suppose that \( a, b \in J \) (i.e. \( a + I, b + I \in \overline{J} \)), and that \( c \in R \). Then \( (a + b) + I = (a + I) + (b + I) \in \overline{J} \) so that \( a + b \in J \). Also, \( c a + I = (c + I)(a + I) \in \overline{J} \) so that \( c a \in J \).

To establish 3. note first that given an ideal \( J \) of \( R \), \( J \) is certainly contained in \( \{ a \in R : a + I \in J/I \} \). On the other hand, this set is equal to \( \{ a \in R : a + I = b + I \) for some \( b \in J \} \). But \( a + I = b + I \) implies that \( a - b \in I \subset J \) and, since \( b \in J \), we obtain that \( a \in J \) as well. Thus \( J = \{ a \in R : a + I \in J/I \} \).
For an ideal $\mathcal{J}$ of $R/I$ and $J = \{a \in R : a + I \in \mathcal{J}\}$ the corresponding ideal of $R$, we show that $\mathcal{J} = J/I$. Observe that $a + I \in \mathcal{J}$ if and only if $a \in J$, and this holds if and only if $a + I \in J/I$ by definition of this ideal.

We have seen that for $F$ a field, all ideals of $F[x]$ are principal. The same holds for quotient rings of $F[x]$.

**Corollary 7.9.** Let $F$ be a field, $I$ and ideal of $F[x]$, and $\mathcal{J}$ an ideal of $F[x]/I$. Then $\mathcal{J}$ is a principal ideal.

**Proof.** By the theorem, $\mathcal{J} = J/I$ for some ideal $J$ of $F[x]$. Since every ideal of $F[x]$ is principal $J = (g)$ for some $g \in F[x]$. If $a + I \in J/I$, then $a \in J$ so that $a = gh$ for some $h \in F[x]$. Thus $a + I = gh + I = (g + I)(h + I)$, which lies in the principal ideal $(g + I)$ of $F[x]/I$. Since $g \in J$, we obtain immediately that $(g + I) \subseteq J/I$.

**Corollary 7.10.** Let $C$ be a cyclic code of length $n$. Then there is a divisor $g$ of $f = x^n + 1$ in $\mathbb{Z}_2[x]$ so that $C = (g)/(f)$. Conversely every divisor of $f$ gives rise to a cyclic code in this manner.

**Proof.** Since $C$ is an ideal of $\mathbb{Z}_2[x]/(f)$ it is principal generated say by $g$, and $(f) \subseteq (g)$. But $(f) \subseteq (g)$ means precisely that $g$ divides $f$.

**Definition 7.11.** The generator polynomial of a cyclic code $C$ of length $n$ is a generator of $C$ as an ideal of $\mathbb{Z}_2[x]/(x^n + 1)$.

**Proposition 7.12.** Let $C$ be a cyclic code of length $n$ with generator polynomial $g(x)$. Then $[g(x), xg(x), \ldots, x^{n-\deg(g)}g(x)]$ is a basis for $C$; therefore, $\dim(C) = n - \deg(g(x))$.

### 7.2 Finite Fields

Recall Proposition 6.21, which says that if $p(x)$ is an irreducible polynomial in $\mathbb{Z}_2[x]$, then the quotient ring $\mathbb{Z}_2[x]/(p(x))$ is a field. If $\deg(p(x)) = n$, then there are $2^n$ elements in this field. We will use this idea to construct finite fields of size a power of 2 as a tool in building codes. We remark that if $\mathbb{Z}_2[x]/I$ is such a field, then for any element $f(x) + I$, then $(f(x) + I) + (f(x) + I) = (f(x) + f(x)) + I$. However, since $f(x) \in \mathbb{Z}_2[x]$, we have $f(x) + f(x) = 0$. Thus, any element $\alpha$ of $\mathbb{Z}_2[x]/I$ satisfies $\alpha + \alpha = 0$, or $-\alpha = \alpha$. We will use this fact frequently.

We will denote by $\text{GF}(q)$ a finite field with $q$ elements. To produce such fields we need to produce irreducible polynomials in $\mathbb{Z}_2[x]$. Given a polynomial $p(x) \in \mathbb{Z}_2[x]$, we can determine if $p(x)$ is irreducible by using a computer algebra package or by a tedious calculation, if $\deg(p)$ is not too large. However, if $\deg(p) \leq 3$, we have an easy test for irreducibility.

**Proposition 7.13.** Let $F$ be a field, and let $p(x) \in F[x]$ with $2 \leq \deg(p) \leq 3$. Then $p(x)$ is irreducible over $F$ if and only if $p(x)$ has no roots in $F$. 
Proof. One direction is easy. If \( a \in F \) with \( p(a) = 0 \), then \( x - a \) divides \( p(x) \), and so \( p(x) = (x - a)q(x) \) for some polynomial \( q(x) \). The assumption on the degree of \( p(x) \) shows that \( q(x) \) is not a constant polynomial. Therefore, \( p(x) \) is reducible. Conversely, suppose that \( p(x) \) is reducible. Then we may factor \( p(x) = f(x)g(x) \) with each of \( f \) and \( g \) nonconstant. Since \( \deg(p) \leq 3 \) and \( \deg(p) = \deg(f) + \deg(g) \), we conclude that either \( \deg(f) = 1 \) or \( \deg(g) = 1 \). Suppose that \( \deg(f) = 1 \). Then \( f(x) = ax + b \) for some \( a, b \in F \) with \( a \neq 0 \). Then \( f(x) \) has a root \( -ba^{-1} \in F \); this element is then also a root of \( p(x) \). A similar conclusion holds if \( \deg(g) = 1 \).

Example 7.14. The polynomial \( x^2 + x + 1 \) is irreducible over \( \mathbb{Z}_2 \); thus, we obtain a field \( \mathbb{Z}_2[x]/(x^2 + x + 1) \). Its elements are the cosets \( 0 + I, 1 + I, x + I, x + 1 + I \). We write the first two cosets as \( 0 \) and \( 1 \), latter two as \( a \) and \( b \). We then see that + and \( \cdot \) are given by the tables

\[
\begin{array}{cccc}
+ & 0 & 1 & a & b \\
0 & 0 & 1 & a & b \\
1 & 1 & 0 & b & a \\
a & a & b & 0 & 1 \\
b & b & a & 1 & 0 \\
\end{array}
\quad
\begin{array}{cccc}
\cdot & 0 & 1 & a & b \\
0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & a & b \\
a & 0 & a & b & 1 \\
b & 0 & b & 1 & a \\
\end{array}
\]

These tables are exactly those of Example 3.28. Those tables were built from this quotient ring.

One important property of finite fields is the existence of primitive elements.

Definition 7.15. A primitive element of a finite field \( F \) is an element \( \alpha \) so that every nonzero element of \( F \) is a power of \( \alpha \).

For convenience, we write \( F^* \) for the set of nonzero elements of \( F \). Referring to Example 7.14, both \( a \) and \( b \) are primitive elements, since \( F^* = \{1, a, b\} \) and \( b = a^2 \) and \( a = b^2 \). Note that \( 1 = a^0 = b^0 \), so \( 1 \) is always a power of any nonzero element.

Example 7.16. The polynomial \( x^3 + x + 1 \) is irreducible over \( \mathbb{Z}_2 \) since it does not have a root in \( \mathbb{Z}_2 \). Thus, if \( I = (x^3 + x + 1) \), we may form the field \( \mathbb{Z}_2[x]/I \). This field has 8 elements. One representation of these elements is as the cosets of polynomials of degree less than 3. Alternatively, if \( \alpha = x + I \), then we have the following table.

<table>
<thead>
<tr>
<th>power of ( \alpha )</th>
<th>coset representative</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha^0 )</td>
<td>1</td>
</tr>
<tr>
<td>( \alpha^1 )</td>
<td>( x )</td>
</tr>
<tr>
<td>( \alpha^2 )</td>
<td>( x^2 )</td>
</tr>
<tr>
<td>( \alpha^3 )</td>
<td>( x + 1 )</td>
</tr>
<tr>
<td>( \alpha^4 )</td>
<td>( x^2 + x )</td>
</tr>
<tr>
<td>( \alpha^5 )</td>
<td>( x^2 + x + 1 )</td>
</tr>
<tr>
<td>( \alpha^6 )</td>
<td>( x^2 + 1 )</td>
</tr>
<tr>
<td>( \alpha^7 )</td>
<td>1</td>
</tr>
</tbody>
</table>
To help see how to make these calculations, we point out that \( \alpha^3 = (x + I)^3 = x^3 + I \). However, \( x^3 + x + 1 \in I \); therefore, \( x^3 + I = x + 1 + I \). We can use this to calculate further powers of \( \alpha \). For example, \( \alpha^4 = \alpha^3 \cdot \alpha = (x + 1 + I)(x + I) = (x + 1)x + I = x^2 + x \). By using this idea along with using the relation \( x^3 + I = x + 1 + I \), we can obtain all the powers of \( \alpha \). Alternatively, we can use the division algorithm. For instance, since \( \alpha^5 = x^5 + I \), we calculate that \( x^5 = (x^3 + x + 1)(x^2 + 1) + x^2 + x + 1 \), yielding that \( x^5 + I = x^2 + x + 1 + I \).

It is true that every finite field has a primitive element. However, the proof of this fact involves more complicated ideas than we will consider, so we do not give the proof. However, for any example, we will be able to produce a primitive element, either by a hand calculation or by using a computer algebra package.

**Lemma 7.17.** Let \( \alpha \) be a primitive element of \( \text{GF}(2^n) \). Then \( \alpha^{2^n-1} = 1 \). Furthermore, \( \alpha^r \neq 1 \) for all \( r \) with \( 1 \leq r < 2^n - 1 \).

**Proof.** Since \( \text{GF}(2^n)^* = \{\alpha^i : i \geq 1\} \) is a finite set, there are positive integers \( r < s \) with \( \alpha^r = \alpha^s \). Then \( \alpha^{s-r} = 1 \). Thus, by the well-ordering principle of \( \mathbb{Z} \), there is a smallest integer \( t \) satisfying \( \alpha^t = 1 \). We next show that \( \text{GF}(2^n)^* = \{\alpha^i : 0 \leq i < t\} \). To see this, if \( i \in \mathbb{Z} \), then by the division algorithm, there are integers \( q \) and \( r \) with \( i = qt + r \) and \( 0 \leq r < t \). Then \( \alpha^i = \alpha^{qt+r} = (\alpha^q)^r \alpha^r = \alpha^r \) since \( \alpha^t = 1 \). This proves the claim. Furthermore, if \( 0 \leq i < j < t \), then \( \alpha^i \neq \alpha^j \), since if \( \alpha^i = \alpha^j \), then \( \alpha^{j-i} = 1 \), contradicting minimality of \( t \). This proves that \( \{\alpha^i : 0 \leq i < t\} \) has exactly \( t \) elements. Since this set is equal to \( \text{GF}(2^n)^* \), which has \( 2^n - 1 \) elements, we see that \( t = 2^n - 1 \). This proves that \( \alpha^{2^n-1} = \alpha^t = 1 \). Furthermore, we have shown that if \( 0 < i < t \), then \( \alpha^0 \neq \alpha^i \), showing that \( \alpha^i \neq 1 \) if \( 0 \leq i < t = 2^n - 1 \). ∎

**Minimal Polynomials and Roots of Polynomials**

Let \( I \) be the ideal \( (x^3 + x + 1) \) in \( \mathbb{Z}_2[x] \) and \( F \) the field \( \mathbb{Z}_2[x]/I \) described in the previous section. If \( \alpha \) is the coset of \( x \), then we note that

\[
\alpha^2 + \alpha + 1 = (x^2 + I) + (x + I) + (1 + I) = x^2 + x + 1 + I = 0 + I.
\]

Therefore, \( \alpha \) is a root of the polynomial \( x^2 + x + 1 \). In particular, \( \alpha \) is a root of a polynomial over \( \mathbb{Z}_2 \). In fact, we note that each of the three nonzero elements of \( F \) are roots of \( x^3 + 1 = (x + 1)(x^2 + x + 1) \). Moreover, \( 0 \) is the root of the polynomial \( x \). As we will see shortly, each element of a finite field is a root of a nontrivial polynomial in \( \mathbb{Z}_2[x] \). Considering the polynomial of smallest degree for which an element is a root will be an important idea for us.

**Definition 7.18.** Let \( \alpha \) be an element of a finite field \( \text{GF}(2^n) \). The minimal polynomial \( m_\alpha(x) \) of \( \alpha \) is the polynomial of smallest degree for which \( \alpha \) is a root.
7.2. **FINITE FIELDS**

We do not yet know that the minimal polynomial of an element exists. We prove this fact in the following proposition.

**Theorem 7.19.** Let \( \alpha \in \text{GF}(2^n) \). Then the minimal polynomial \( m_\alpha(x) \) of \( \alpha \) exists and is unique. Furthermore,

1. If \( f(x) \in \mathbb{Z}_2[x] \), then \( f(\alpha) = 0 \) if and only if \( m_\alpha(x) \) divides \( f(x) \).
2. \( m_\alpha(x) \) is irreducible over \( \mathbb{Z}_2 \).
3. \( m_\alpha(x) \) divides \( x^{2^n} + x \).

**Proof.** We first show that \( \alpha \) is a root of some nonzero polynomial. If we consider the set \( \{ \alpha^i : i \geq 0 \} \) of powers of \( \alpha \), then we see that this is a finite set since it is a subset of the finite set \( \text{GF}(2^n) \). Since there are infinitely many possible exponents, we see that there are distinct positive integers \( r \) and \( s \) with \( \alpha^r = \alpha^s \). Then \( \alpha^r + \alpha^s = 0 \), so \( \alpha \) is a root of \( x^r + x^s \). This shows that \( \alpha \) is a root of a nontrivial polynomial. Thus, the set \( I = \{ f(x) \in \mathbb{Z}_2[x] : f(\alpha) = 0 \} \) is nonzero. We claim that \( I \) is an ideal. To prove this, we first see that if \( f(x), g(x) \in I \), then \( (f + g)(x) \in I \) since \( (f + g)(\alpha) = f(\alpha) + g(\alpha) = 0 + 0 = 0 \). Also, if \( f(x) \in I \) and \( h(x) \in \mathbb{Z}_2[x] \), then \( (fh)(x) \in I \) since \( (fh)(\alpha) = f(\alpha)h(\alpha) = 0 \cdot h(\alpha) = 0 \). Thus, our claim is proved. By Theorem XYZ, we may write \( I = (p(x)) \) for some polynomial \( p(x) \). We note that, by definition, \( f(x) \in I \) if and only if \( \alpha \) is a root of \( f(x) \). Furthermore, this occurs if and only if \( f(x) \) is divisible by \( p(x) \). This forces \( p(x) \) to be a polynomial of minimal degree for which \( \alpha \) is a root. We will have proven the first two statements along with (1) once we show that \( p(x) \) is unique. However, if \( I = (q(x)) \) for some \( q(x) \), then \( p(x) \) divides \( q(x) \) and vice-versa. If \( p(x) = s(x)q(x) \) and \( q(x) = t(x)p(x) \), then see see that \( p(x) = s(x)t(x)p(x) \). By cancellation, \( s(x)t(x) = 1 \), forcing \( s(x) = t(x) = 1 \), and so \( q(x) = p(x) \). Thus, we have a unique polynomial of smallest degree for which \( \alpha \) is a root. This polynomial is \( m_\alpha(x) \), and so \( m_\alpha(x) = p(x) \) exists and is unique.

To prove irreducibility, suppose that \( m_\alpha(x) = f(x)g(x) \) for some \( f, g \in \mathbb{Z}_2[x] \). Then \( 0 = m_\alpha(\alpha) = f(\alpha)g(\alpha) \). This equation is over the field \( F \). Since \( F \) has no zero divisors, either \( f(\alpha) = 0 \) or \( g(\alpha) = 0 \). This would be a contradiction if \( 1 < \deg(f), \deg(g) < \deg(m_\alpha(x)) \). Therefore, one of \( f(x) \) and \( g(x) \) must be 1 and the other \( m_\alpha(x) \). This proves that \( m_\alpha(x) \) is irreducible over \( \mathbb{Z}_2 \).

Finally, to prove that \( m_\alpha(x) \) divides \( x^{2^n} + x \), we note that if \( \alpha = 0 \), then \( m_\alpha(x) = x \), and this clearly divides \( x^{2^n} + x \). So, suppose that \( \alpha \neq 0 \). We give a proof assuming the existence of a primitive element for every finite field. Suppose that \( \beta \) is a primitive element of \( F \). Then \( \beta^{2^n-1} = 1 \) by Lemma 7.17. Since \( \alpha \neq 0 \), there is an \( i \) with \( \alpha = \beta^i \). Thus,

\[
\alpha^{2^n-1} = (\beta^i)^{2^n-1} = (\beta^{2^n-1})^i = 1^i = 1.
\]

This proves that \( \alpha \) is a root of \( x^{2^n-1} + 1 \). Multiplying by \( x \) yields \( x^{2^n} + x \), and \( \alpha \) is then also a root of this polynomial. Statement 1 then shows that \( m_\alpha(x) \) divides \( x^{2^n} + x \). \( \square \)
We can prove Statement 3 of the previous theorem without using the existence of primitive elements. We give this alternative proof partly to be more complete, and partly because it is a look ahead to group theory. We note that we only used the existence of a primitive element to prove that $\alpha^{2^n-1} = 1$. We verify this result holds without using primitive elements.

**Proposition 7.20.** Let $F$ be a finite field with $|F^*| = q$. Then $\alpha^q = 1$ for each $\alpha \in F^*$.

**Proof.** List the elements of $F^* = \{a_1, \ldots, a_q\}$. We claim that $F^* = \{\alpha a_1, \ldots, \alpha a_q\}$. To see why, we clearly have $\{\alpha a_1, \ldots, \alpha a_q\} \subseteq F^*$ since $\alpha \neq 0$. For the reverse inclusion, we note that, given $i$, the element $a_i \alpha^{-1} \in F^*$, so there is a $j$ with $a_i \alpha^{-1} = a_j$. Then $a_i = \alpha a_j$. This proves the claim. Next, since multiplication in $F^*$ is commutative (and associative), multiplying all of the element of $F^*$ together, we see that

$$a_1 \cdots a_q = (\alpha a_1) \cdots (\alpha a_q) = \alpha^q(a_1 \cdots a_q).$$

Cancellation yields $\alpha^q = 1$, as desired. \qed

**Example 7.21.** Consider the field $GF(8) = \mathbb{Z}_2[x]/(x^3 + x + 1) = \mathbb{Z}_2[x]/I$. As we saw in Example 7.16, $\alpha = x + I$ is a primitive element. By Theorem 7.19, the minimal polynomial of each element of $GF(8)$ divides $x^8 + x$. If we factor this polynomial into irreducible polynomials, we obtain

$$x^8 + x = x(x + 1)(x^3 + x + 1)(x^3 + x^2 + 1).$$

The minimal polynomials of 0 and 1 are $x$ and $x + 1$, respectively. We saw that $x^3 + x + 1$ is the minimal polynomial of $\alpha$. The remaining elements $\alpha^2, \ldots, \alpha^6$ have minimal polynomial equal to one of these two cubics. To see, for example, which is the minimal polynomial of $\alpha^2$, we simply evaluate one at $\alpha^2$. Trying the first cubic, we have;

$$(\alpha^2)^3 + \alpha^2 + 1 = \alpha^6 + \alpha^2 + 1 = (\alpha^2 + 1) + \alpha^2 + 1 = 0;$$

we used that $x^6 + I = x^2 + 1 + I$ to see that $\alpha^6 = \alpha^2 + 1$. Thus, $m_{\alpha^2}(x) = x^3 + x + 1$. Similarly, if we try $\alpha^3$, we have

$$(\alpha^3)^3 + \alpha^3 + 1 = \alpha^9 + \alpha^3 + 1 = \alpha^2 + \alpha^3 + 1 = \alpha^2 + (\alpha + 1) + 1 = \alpha^2 + \alpha \neq 0$$

since $\alpha^7 = 1$ and $\alpha^3 + \alpha + 1 = 0$. Therefore, $m_{\alpha^3}(x) = x^3 + x^2 + 1$, the only possible choice since $\alpha^3$ is not a root of the other three factors of $x^8 + x$. If we were to finish this calculation, we would see that $x^3 + x + 1 = m_\alpha(x) = m_{\alpha^2}(x) = m_{\alpha^4}(x)$ and $x^3 + x^2 + 1 = m_{\alpha^3}(x) = m_{\alpha^6}(x) = m_{\alpha^5}(x)$. By making a simple but quite helpful observation, we can make the process of finding minimal polynomials easier. To help motivate this result, we note that $x^3 + x + 1$ is the minimal polynomial of $\alpha$, $\alpha^2$, and $\alpha^4$, and $x^3 + x^2 + 1$ is the minimal polynomial of $\alpha^3$, $\alpha^6 = (\alpha^3)^2$, and $\alpha^5 = (\alpha^3)^4$. 

\[72x76\]
To help us prove the following proposition, we note that if \( a, b \in \text{GF}(2^n) \), then \((a + b)^2 = a^2 + 2ab + b^2 = a^2 + b^2 \) since \(2c = c + c = 0 \) for any \( c \in \text{GF}(2^n) \). Thus, we have the formula \((a + b)^2 = a^2 + b^2 \). By an induction argument, we see that \((a_1 + \cdots + a_r)^2 = a_1^2 + \cdots + a_r^2\) for any \( r \) and any \( a_i \in \text{GF}(2^n) \).

**Proposition 7.22.** Let \( f(x) \in \mathbb{Z}_2[x] \). If \( \alpha \in \text{GF}(2^n) \) is a root of \( f(x) \), then \( \alpha^2 \) is also a root of \( f(x) \).

Write \( f(x) = b_0 + \cdots + b_{r-1}x^{r-1} + x^r \). Note that each \( b_i \in \mathbb{Z}_2 = \{0, 1\} \). Since \( f(\alpha) = 0 \), we have \( b_0 + b_1\alpha + \cdots + b_{r-1}\alpha^{r-1} + \alpha^r \). Therefore,

\[
 f(\alpha^2) = \sum b_i(\alpha^2)^i = \sum b_i^2\alpha^{2i} = \left(\sum b_i\alpha^i\right)^2 = f(\alpha)^2 = 0.
\]

Going back to the previous example, since \( m_\alpha(x) = x^3 + x + 1 \) has \( \alpha \) as a root, the proposition implies that \( \alpha^2 \) and \( \alpha^4 = (\alpha^2)^2 \) are roots as well. Their minimal polynomials must also be \( x^3 + x + 1 \) since the minimal polynomial of each divides \( x^3 + x + 1 \), and \( x^3 + x + 1 \) is irreducible over \( \mathbb{Z}_2 \), i.e. it has no divisors other than itself and 1. Similarly, once we know that \( m_{\alpha^3}(x) = x^3 + x^2 + 1 \), we conclude that this is also the minimal polynomial of \((\alpha^3)^2 = \alpha^6\) and \((\alpha^3)^4 = \alpha^5\). However, \((\alpha^3)^8 = (\alpha^5)^2 = \alpha^{10} = \alpha^3\), so we do not produce any more roots of \( x^3 + x^2 + 1 \). By the way, this shows that over \( \text{GF}(8) \), we have the factorizations

\[
x^3 + x + 1 = (x - \alpha)(x - \alpha^2)(x - \alpha^4),
\]

\[
x^3 + x^2 + 1 = (x - \alpha^3)(x - \alpha^6)(x - \alpha^5).
\]

Thus, \( x^8 + x = x(x + 1)(x^3 + x + 1)(x^3 + x^2 + 1) = (x - 0)(x - 1)(x - \alpha)(x - \alpha^2)(x - \alpha^3)(x - \alpha^4)(x - \alpha^5)(x - \alpha^6) \).

In other words, \( x^8 = x \) has one linear factor for each element of \( \text{GF}(8) \) as must be the case from the remainder theorem and its consequences.

1. Use both the simp command and the Rem command to check that you get the same thing. Recall that the syntax for the Rem command is Rem\((h, k, x) \mod 2\); pay attention to the fact that the simp command will give an answer in terms of \( a \) while the Rem command will write the answer in terms of \( x \).

2. There are three roots of the polynomial \( x^3 + x^2 + 1 \) in \( F \); find them. Choose one of the roots; call it \( t \). Calculate \( t, t^2, \) and \( t^4 \) to see that \( t, t^2, t^4 \) are the three roots of \( x^3 + x^2 + 1 \).

3. There are three roots of the polynomial \( x^3 + x + 1 \) in \( F \); find them. Choose one of the roots; call it \( t \). Calculate \( t, t^2, \) and \( t^4 \) to see that \( t, t^2, t^4 \) are the three roots of \( x^3 + x + 1 \).
CHAPTER 7. CYCLIC CODES

7.3 Reed-Solomon Codes

The music that you hear on a CD is encoded using a special class of cyclic codes known Reed-Solomon codes. As we saw in Section 7.1, we can produce a cyclic code of length \( m \) by finding a polynomial \( g(x) \) which divides \( x^m + 1 \). We use this idea to produce important codes of length \( 2^n - 1 \) using a primitive element for \( GF(2^n) \). Choose a positive integer \( n \), and let \( t \) be an integer with \( t \leq 2^n - 1 \). For a primitive element of \( GF(2^n) \), let \( g(x) \) be the least common multiple of the minimal polynomials of the elements \( \alpha, \alpha^2, \ldots, \alpha^{2t} \) (i.e. the polynomial of least degree with these roots). The code whose generator polynomial is \( g(x) \) is called a Reed-Solomon code with designated distance \( d = 2t + 1 \). We determine the dimension of this code and see why \( d \) is called the designated distance in the theorem below.

Example 7.23. Let \( n = 3 \), let \( \alpha \) be a primitive element of \( GF(8) \), and let \( C = RS(3, 1, \alpha) \). This means that the generator polynomial \( g(x) \) is the minimal polynomial of \( \alpha \) since this polynomial will vanish on \( \alpha \) and \( \alpha^2 \). Thus, \( \deg(g(x)) = 3 \), so the dimension \( k \) of \( C \) is \( 2^n - 1 - \deg(g(x)) = 7 - 3 = 4 \). To be more explicit, let us represent \( GF(8) \) as \( \mathbb{Z}_2[x]/(x^3 + x + 1) \) and choose \( x \) to be the coset of \( x \). Then \( g(x) = x^3 + x + 1 \). From our association between polynomials and words, the basis \([g(x), xg(x), x^2g(x), x^3g(x)]\) corresponds to the four words \([1101000, 0110100, 0011010, 0001101]\). A generator matrix for \( C \) is then

\[
\begin{pmatrix}
1 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 1
\end{pmatrix}.
\]

If we compute a parity check matrix for \( C \), we can get

\[
H = \begin{pmatrix}
1 & 0 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 1
\end{pmatrix}
\]

to be such a matrix. Looking at \( H \), we see that it closely resembles the Hamming matrix; in fact it is obtained from the Hamming matrix by appropriately rearranging the columns. This Reed-Solomon code is obtained from the \((7,4)\) Hamming code by a permutation. Recall that the \((7,4)\) Hamming code corrects one error.

Theorem 7.24. Let \( C = RS(n, t, \alpha) \). Then the distance \( d \) of \( C \) is at least \( 2t + 1 \). Therefore, \( C \) corrects at least \( t \) errors.

Proof. Set \( m = 2^n - 1 \). The code \( C \) consists of all cosets in \( \mathbb{Z}_2[x]/(x^m + 1) \) of polynomials with roots \( \alpha, \alpha^2, \ldots, \alpha^{2t} \). According to our convention identifying polynomials of degree \( < m \)
with the vector of their coefficients (of length \(m\), \(C\) is the nullspace of the \(2t \times m\) matrix

\[
H = \begin{pmatrix}
1 & \alpha^1 & \alpha^2 & \ldots & \alpha^{m-1} \\
1 & \alpha^2 & (\alpha^2)^2 & \ldots & (\alpha^2)^{m-1} \\
1 & \alpha^3 & (\alpha^3)^2 & \ldots & (\alpha^3)^{m-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \alpha^{2t} & (\alpha^{2t})^2 & \ldots & (\alpha^{2t})^{m-1}
\end{pmatrix}.
\]

The distance of \(C\) is at least \(d\) if every set of \(d - 1 = 2t\) columns of \(C\) is linearly independent (i.e. their sum is nonzero). This condition holds if and only if every submatrix of \(H\) consisting of \(2t\) of its columns has nonzero determinant. To see that this is satisfied, consider such a submatrix:

\[
M = \begin{pmatrix}
\alpha^{j_1} & \alpha^{j_2} & \ldots & \alpha^{j_{2t}} \\
(\alpha^{j_1})^2 & (\alpha^{j_2})^2 & \ldots & (\alpha^{j_{2t}})^2 \\
(\alpha^{j_1})^3 & (\alpha^{j_2})^3 & \ldots & (\alpha^{j_{2t}})^3 \\
\vdots & \vdots & \ddots & \vdots \\
(\alpha^{j_1})^{2t} & (\alpha^{j_2})^{2t} & \ldots & (\alpha^{j_{2t}})^{2t}
\end{pmatrix}.
\]

The following lemma shows that \(\det(M) = (\alpha^{j_1}\alpha^{j_2}\cdots\alpha^{j_{2t}})\Pi\) where \(\Pi\) is the product of all \((\alpha^{j_i} - \alpha^{j_k})\) where \(j_i > j_k\). In particular, since \(2t \leq m < 2^n\), \(\alpha^{j_i} - \alpha^{j_k} \neq 0\) for each such \(j_i, j_k\) and we see that \(\Pi \neq 0\).

**Lemma 7.25.** (Vandermonde determinant) Let \(F\) be a field and \(a_1, a_2, \ldots, a_n\) distinct elements of \(F\). The determinant of the Vandermonde matrix

\[
V = \begin{pmatrix}
1 & 1 & \ldots & 1 \\
a_1 & a_2 & \ldots & a_n \\
\vdots & \vdots & \ddots & \vdots \\
a_1^{n-1} & a_2^{n-1} & \ldots & a_n^{n-1}
\end{pmatrix}
\]

is equal to \((a_n - a_1)(a_n - a_2)\cdots(a_n - a_{n-1}) \cdot (a_{n-1} - a_1) \cdots (a_{n-1} - a_{n-2}) \cdots (a_1 - a_1) = \Pi_{i>j}(a_i - a_j)\).

**Proof.** Argue by induction on \(n\) beginning with \(n = 2\):

\[
\begin{vmatrix}
1 & 1 \\
a_1 & a_2
\end{vmatrix} = a_2 - a_1.
\]
Suppose that the result holds for \( n \) elements of \( F \), \( n \geq 2 \). Consider the matrix
\[
M(x) = \begin{pmatrix}
1 & 1 & \ldots & 1 & 1 \\
a_1 & a_2 & \ldots & a_n & x \\
& \ddots & \ddots & \ddots & \ddots \\
& & \ddots & \ddots & \ddots \\
a_1^n & a_2^n & \ldots & a_n^n & x^n
\end{pmatrix}
\]
Note that \( M(a_{n+1}) \) is the Vandermonde matrix corresponding to \( a_1, a_2, \ldots, a_{n+1} \). The determinant of \( M(x) \) is a polynomial of degree precisely \( n \) with the \( n \) distinct roots \( a_1, a_2, \ldots, a_n \). By repeated use of the remainder theorem, \( \det(M(x)) = \alpha \prod_{i=1}^{n}(x - a_i) \) where \( \alpha \) is the coefficient of \( x^n \). But the coefficient of \( x^n \) is exactly the determinant of the Vandermonde matrix
\[
\begin{pmatrix}
1 & 1 & \ldots & 1 \\
a_1 & a_2 & \ldots & a_n \\
& \ddots & \ddots & \ddots \\
& & \ddots & \ddots & \ddots \\
a_1^{n-1} & a_2^{n-1} & \ldots & a_n^{n-1}
\end{pmatrix}
\]
which, by the induction hypothesis, is equal to \( \prod_{i>j}(a_i - a_j) \).

Remark 7.26. In the proof of the theorem, the factor \( (\alpha^{3i}\alpha^{3j} \cdots \alpha^{2n}) \) comes from factoring \( \alpha^{3i} \) from each of the \( n \) rows of \( M \) before applying the result of the lemma.

Example 7.27. Let’s build a 3 error correcting code of length 15, i.e. \( RS(15, 3, \alpha) \) where \( \alpha = \bar{x} \in \mathbb{Z}_2[x]/(x^4 + x + 1) \) is a primitive element for this concrete realization of \( GF(16) \). The objective is to find the polynomial \( g \in \mathbb{Z}_2[x] \) of least degree with roots \( \alpha, \alpha^2, \ldots, \alpha^6 \) and \( f \in \mathbb{Z}_2[x] \) has \( \beta \) as a root, then \( \beta^{2i} \) is a root of \( f \) for every \( i \) we see that \( \alpha^2 \) and \( \alpha^4 \) are also roots of \( m_\alpha(x) \).

Example 7.28. Now we must determine \( m_{\alpha^3}(x), m_{\alpha^5}(x), \) and \( m_{\alpha^6}(x) \). Arguments similar to those above show that \( h = (x - \alpha^3)(x - \alpha^6)(x - \alpha^9)(x - \alpha^{12}) \) divides \( m_{\alpha^3}(x) \) and therefore \( g \) (note that \( \alpha^{15} = 1 \) so that \( \alpha^{24} = \alpha^9 \)). Thus this product must equal \( m_{\alpha^3}(x) = m_{\alpha^6}(x) \) if it lies in \( \mathbb{Z}_2[x] \). One can either multiply out the factors of \( h \) and check coefficients, or argue as follows. Since the degree of \( h \) is four, \( h = m_{\alpha^3}(x) \) if there is an irreducible polynomial of degree 4 in \( \mathbb{Z}_2[x] \) that has \( \alpha^3 \) as a root. Of the polynomials of degree 4 in \( \mathbb{Z}_2[x] \), only \( x^4 + x + 1, x^4 + x^3 + 1, \) and \( x^4 + x^3 + x^2 + x + 1 \) are irreducible (check this!). Evaluating the last polynomial at \( \alpha^3 \) reveals that \( x^4 + x^3 + x^2 + x + 1 = m_{\alpha^3}(x) \).

Example 7.29. Finally we determine \( m_{\alpha^5}(x) \). Consider \( (x - \alpha^5)(x - \alpha^{10}) \) noting that \( \alpha^{20} = \alpha^5 \). We see that \( (x - \alpha^5)(x - \alpha^{10}) = x^2 + (\alpha^5 + \alpha^{10})x + 1 = x^2 + x + 1 = m_{\alpha^5}(x) \). Thus
g(x) = (x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1)(x^2 + x + 1) generates a 3 error correcting code of dimension 5.

**Remark 7.30.** Reed-Solomon codes are a special subclass of the cyclic codes known as BCH codes. Just as with Reed-Solomon codes, BCH codes are built with a primitive element \(\alpha\) for \(GF(2^n)\), but instead of having length \(2^n - 1\), a general BCH code can have as length any divisor \(m\) of \(2^n - 1\) with generator polynomial vanishing on \(\beta^i, \beta^2, \ldots \beta^{2t}\) where \(\beta = \alpha^{(2^n-1)/m}\).

**Exercises**

1. Consider the \(RS(15, 3, \alpha)\) code constructed above.
   
   (a) Write down any four codeword polynomials in this code.
   
   (b) Determine whether or not \(x^{13} + x^{12} + x^{11} + x^{10} + x^6 + x^3 + x^2 + 1\) is a codeword polynomial.

2. Consider the code of Example 4 of Page 57 of Mackiw. Check that \(r(x) = x^{13} + x^{12} + x^{11} + x^{10} + x^9 + x^8 + x^5 + x^4 + x^3 + x + 1\) is not a codeword. Calculate the \(S_i = r(\alpha^i)\) and write them as powers of \(\alpha\); consult Table 2 on Page 55. Next, obtain the \(2t \times 2t\) matrix \(A\) whose \(i, j\) entry is \(S_{i+j-1}\). Determine the number of errors \(v\) by finding the largest \(l\) for which the \(l \times l\) top left part of \(A\) is invertible. Use Assignment9.mws to help you with these computations.

3. Consider the BCH code of Example 5 of Page 58 of Mackiw. This code can correct up to two errors. Write down two nonzero codewords. Change each by making two errors. Run through, for each word, the steps you took for the previous problem, and see that you recover the original codeword.

4. Using the code of the previous problem, take a valid codeword and change it by making three errors. Go through the decoding procedure for it. Do you recover the original codeword?

5. A fact about \(RS\) codes is that if \(g\) is the generator polynomial for an \(RS\) code defined with a primitive element for \(GF(2^n)\) and \(m = \deg(g)\), then the code has dimension \(2^n - 1 - m\). Determine the number of codewords in the code of Problem 1.

6. If you wish to transmit color pictures using 2048 colors, you will need 2048 codewords. If you wish to do this with an \(RS\) code using the field \(F\) of Problem 1, what degree of a generator polynomial will you need to use? By using the generator procedure, find the largest value of \(t\) for which the corresponding code has at least 2048 codewords (recall that \(t \leq n/2\) in order for the code to correct at least \(t\) errors, so do not consider values of \(t\) larger than \(n/2\)). What percentage of errors in a codeword can be corrected?
7. Continuing the previous problem, if you instead use \( F = \mathbb{Z}_2[x]/(x^5 + x^2 + 1) \), determine the length of a RS code you obtain from this field, and determine the degree of a generator polynomial needed to have 2048 codewords. What is the largest value of \( t \) you can use to have 2048 codewords? By using this larger field, can you correct a larger percentage of errors in a codeword than in the previous problem?

### 7.4 Error Correction for Reed-Solomon Codes

In this section a procedure is developed for the correction of any set of \( t \) or fewer errors for an \( RS(2^n - 1, t, \alpha) \) code. Using our convention identifying words with cosets of polynomials of degree \( < 2^n - 1 \), we’ll consider only polynomial expressions. Thus, we suppose that \( c(x) \) is a codeword that is received as \( r(x) \). If \( r(x) \neq c(x) \), write \( r(x) = c(x) + e(x) \), where \( e(x) \) is the error polynomial. Note that \( e(x) = x^{m_1} + x^{m_2} + \cdots + x^{m_k} \) where \( k \) is equal to the number of errors that have occurred in positions \( m_1, m_2, \ldots, m_k \). Our decoding procedure will first determine \( k \) and then locate the \( m_i \) thus determining \( e(x) \). The codeword \( c(x) \) is then recovered as \( r(x) + e(x) \).

By the construction of the RS code, \( c(\alpha^i) = 0 \) for \( i = 1, 2, \ldots, 2t \). This already gives us some information about \( e(x) \), namely, for \( i = 1, 2, \ldots, 2t \) we have

\[
r(\alpha^i) = c(\alpha^i) + e(\alpha^i) = e(\alpha^i).
\]

In particular, the \( 2t \) elements of \( GF(2^n) \) given by

\[
e(\alpha^i) = (\alpha^{m_1})^i + (\alpha^{m_2})^i + \cdots + (\alpha^{m_k})^i
\]

are known, and from these we will recover \( k \) and the \( m_i \). To that end, for each \( i = 1, \ldots, k \) set \( Y_j \) equal to the as yet unknown powers \( \alpha^{m_j} \) of our primitive element for \( GF(2^n) \), so that

\[
e(\alpha) = \sum_{j=1}^{k} Y_j
\]

\[
e(\alpha^2) = \sum_{j=1}^{k} Y_j^2
\]

\[\vdots\]

\[
e(\alpha^{2t}) = \sum_{j=1}^{k} Y_j^{2t}
\]
To simplify the notation and to emphasize the relationship with the \( Y_i \), set

\[ e(\alpha^i) = S_i = S_i(Y_1, \ldots, Y_k) = \sum_{j=1}^{k} Y_j^i. \]

Thus, our goal is to determine \( k \) and the \( Y_j \) by constructing a polynomial in \( GF(2^n) \) whose roots are precisely the \( Y_j \). Solving for the \( Y_j \) as powers of \( \alpha \) will give the positions of the errors.

Certainly \( L(z) = (z + Y_1)(z + Y_2) \cdots (z + Y_k) \) is the desired polynomial, called the error locator polynomial, but we don’t as yet know its coefficients. We can, however, calculate them from the \( S_i \) together with a version of a result known as Newton’s identities. Collecting coefficients of powers of \( z \) we have

\[ L(z) = \sum_{i=1}^{k} \sigma_{k-i}(Y_1, \ldots, Y_k)z^i \]

where

\[ \sigma_0 = 1 \]

\[ \sigma_1(Y_1, \ldots, Y_k) = Y_1 + \cdots + Y_k = \sum_{j=1}^{k} Y_j \]

\[ \sigma_2(Y_1, \ldots, Y_k) = Y_1^2 + Y_1Y_2 + \cdots + Y_1Y_k + \cdots + Y_{k-1}Y_k + Y_k^2 = \sum_{1 \leq i \leq j \leq k} Y_iY_j \]

\[ \sigma_3(Y_1, \ldots, Y_k) = \sum_{1 \leq i \leq j \leq r \leq k} Y_iY_jY_r \]

\[ \vdots \]

\[ \sigma_k(Y_1, \ldots, Y_k) = Y_1Y_2 \cdots Y_k \]

are known as the elementary symmetric functions in \( Y_1, Y_2, \ldots, Y_k \). The (to be determined) \( \sigma_i \) are related to the (known) \( S_i \) by Newton’s identities:

**Lemma 7.31.** For each \( i = 1, 2, \ldots, 2t \) we have

\[ S_i\sigma_k + S_{i+1}\sigma_{k-1} + \cdots + S_{i+k-1}\sigma_1 + S_{i+k} = 0. \]

**Proof.** For each \( j = 1, \ldots, k \) evaluate \( L(z) = \sum_{i=1}^{k} \sigma_{k-i}(Y_1, \ldots, Y_k)z^i \) at \( z = Y_j \) to obtain

\[ 0 = Y_j^k + \sigma_1Y_j^{k-1} + \sigma_2Y_j^{k-2} + \cdots + \sigma_{k-1}Y_j + \sigma_k. \]

Multiplying each equation by \( Y_j^i \) results in

\[ 0 = Y_j^{i+k} + \sigma_1Y_j^{i+k-1} + \sigma_2Y_j^{i+k-2} + \cdots + \sigma_{k-1}Y_j^{i+1} + \sigma_kY_j^i \]
and summing these \( j \) equations yields

\[
0 = \sum_{j=1}^{k} Y_j^{i+k} + \sigma_1 \sum_{j=1}^{k} Y_j^{i+k-1} + \sigma_2 \sum_{j=1}^{k} Y_j^{i+k-2} + \cdots + \sigma_{k-1} \sum_{j=1}^{k} Y_j^{i+1} + \sigma_k \sum_{j=1}^{k} Y_j^i
\]

which is the desired result.

Recalling once again that we are working over \( \mathbb{Z}_2 \), view the system of equations \( S_i \sigma_k + S_{i+1} \sigma_{k-1} + \cdots + S_{i+k-1} \sigma_1 = S_{i+k} \) for \( 1 \leq i \leq k \) as a linear system in the \( k \) unknowns \( \sigma_j \) with coefficients \( S_{i+k} \). In matrix form this becomes

\[
\begin{bmatrix}
S_1 & S_2 & \cdots & S_k \\
S_2 & S_3 & \cdots & S_{k+1} \\
\vdots & \vdots & \ddots & \vdots \\
S_k & S_{k+1} & \cdots & S_{2k-1}
\end{bmatrix}
\begin{bmatrix}
\sigma_k \\
\sigma_{k-1} \\
\vdots \\
\sigma_1
\end{bmatrix}
= \begin{bmatrix}
S_{k+1} \\
S_{k+2} \\
\vdots \\
S_{2k}
\end{bmatrix}.
\]

(★)

If the matrix of coefficients is invertible we solve it for the \( \sigma_j \), which are the coefficients in the error locator polynomial \( L(z) \), then solve \( L(z) \). Expressing the roots of \( L(z) \) as powers of \( \alpha \) locates the positions of the errors.

All of this development is done under the assumption that \( k \) errors have been made. But as of yet we don’t know the value of \( k \). The next lemma enables us to determine \( k \).

**Lemma 7.32.** Let \( \mathbb{S} \) be the \( t \times t \) matrix

\[
\begin{bmatrix}
S_1 & S_2 & \cdots & S_t \\
S_2 & S_3 & \cdots & S_{t+1} \\
\vdots & \vdots & \ddots & \vdots \\
S_t & S_{t+1} & \cdots & S_{2t-1}
\end{bmatrix}
\]

. Then \( k \) is the largest integer for which the \( k \times k \) submatrix

\[
\mathbb{S}_k=
\begin{bmatrix}
S_1 & S_2 & \cdots & S_k \\
S_2 & S_3 & \cdots & S_{k+1} \\
\vdots & \vdots & \ddots & \vdots \\
S_k & S_{k+1} & \cdots & S_k
\end{bmatrix}
\]

is invertible.

Thus to find \( k \) begin with the matrix \( \mathbb{S} \). If \( \mathbb{S} \) is invertible, then \( k = t \). Otherwise, delete the last column and last row of \( \mathbb{S} \) and check for invertibility. Iteration of this process will result in an invertible matrix \( \mathbb{S}_k \) and hence the linear system ★ can be solved for the \( \sigma_j \).
7.4. ERROR CORRECTION FOR REED-SOLOMON CODES

Proof. (Of the lemma) We show that the matrix \( S_k \) is nonsingular while if \( l > k \) then the matrix \( S_l \) is singular. Note that \( l \leq t \). Recalling that \( S_i = \sum_{j=1}^{k} Y_j^i \), the matrix \( S_l \) factors as

\[
\begin{bmatrix}
S_1 & S_2 & \ldots & S_t \\
S_2 & S_3 & \ldots & S_{l+1} \\
\vdots & \vdots & \ddots & \vdots \\
S_l & S_{l+1} & \ldots & S_l
\end{bmatrix}
\]

Moreover, the entries of \( S_l \) depend only on \( Y_1, \ldots, Y_k \). Thus, if \( l > k \), we may take \( Y_{k+2} = \cdots = Y_t = 0 \) to obtain that \( D \), and therefore \( S_l \), is singular. The \( Y_j \) for \( 1 \leq j \leq k \) are all nonzero though, and distinct as they are distinct powers of the primitive element \( \alpha \). Furthermore, \( A \) is a Vandermonde matrix so that \( |S_k| = |A| |D| |A^T| = |A|^2 |D| = \Pi_{1 \leq i < j \leq k} (Y_i - Y_j)^2 \Pi_{i=1}^{k} Y_i \neq 0 \).

Example 7.33. Let’s see how the decoding procedure is implemented on the \( RS(15, 3, \alpha) \) constructed above. Suppose that the received word is the polynomial \( r(x) = x + x^2 + x^3 + x^4 + x^5 + x^6 + x^7 + x^{10} + x^{12} \). We compute the \( S_i = r(\alpha^i) \), \( i = 1, 2, 3, 4, 5 = 2t - 1 \) using the relations \( \alpha^4 = \alpha + 1 \), \( \alpha^{15} = 1 \).

\[
S_1 = \alpha + \alpha^2 + \alpha^3 + \alpha^4 + \alpha^5 + \alpha^6 + \alpha^7 + \alpha^{10} + \alpha^{12} = \alpha^2 \\
S_2 = \alpha^2 + \alpha^4 + \alpha^5 + \alpha^6 + \alpha^8 + \alpha^{10} + \alpha^{12} + \alpha^{14} + \alpha^{20} + \alpha^{24} = \alpha + 1 \\
S_3 = \alpha^3 + \alpha^6 + \alpha^9 + \alpha^{12} + \alpha^{15} + \alpha^{18} + \alpha^{21} + \alpha^{30} + \alpha^{36} = \alpha^3 + 1 \\
S_4 = \alpha^4 + \alpha^8 + \alpha^{12} + \alpha^{16} + \alpha^{20} + \alpha^{24} + \alpha^{28} + \alpha^{40} + \alpha^{48} = \alpha^2 + 1 \\
S_5 = \alpha^5 + \alpha^{10} + \alpha^{15} + \alpha^{20} + \alpha^{25} + \alpha^{30} + \alpha^{35} + \alpha^{50} + \alpha^{60} = 1.
\]

Consider the matrix

\[
S = \begin{bmatrix}
S_1 & S_2 & S_3 \\
S_2 & S_3 & S_4 \\
S_3 & S_4 & S_5
\end{bmatrix}
\]

\[
S = \begin{bmatrix}
\alpha^2 & \alpha + 1 & \alpha^3 + 1 \\
\alpha + 1 & \alpha^3 + 1 & \alpha^2 + 1 \\
\alpha^3 + 1 & \alpha^2 + 1 & 1
\end{bmatrix}
\]

Since \( |S| = 0 \) there are fewer than 3 errors.

We next consider

\[
S_2 = \begin{bmatrix}
S_1 & S_2 \\
S_2 & S_3
\end{bmatrix}
\]

\[
S_2 = \begin{bmatrix}
\alpha^2 & \alpha + 1 \\
\alpha + 1 & \alpha^3 + 1
\end{bmatrix}
\]
Since \(|S_2| = \alpha^2 + \alpha + 1 \neq 0\), there are exactly two errors, and we can solve the system

\[
S_2 \begin{bmatrix} \sigma_2 \\ \sigma_1 \end{bmatrix} = \begin{bmatrix} S_3 \\ S_4 \end{bmatrix}
\]

for the coefficients of the error locator polynomial \(L(z) = z^2 + \sigma_1 z + \sigma_2\). Our system

\[
\begin{bmatrix} \alpha^2 & \alpha + 1 \\ \alpha + 1 & \alpha^3 + 1 \end{bmatrix} \begin{bmatrix} \sigma_2 \\ \sigma_1 \end{bmatrix} = \begin{bmatrix} \alpha^3 + 1 \\ \alpha^2 + 1 \end{bmatrix}
\]

can be solved by inverting \(S_2\). Note that \(\alpha^2 + \alpha + 1 = \alpha^{10}\), and \(\alpha^{-10} = \alpha^5\) so

\[
S_2^{-1} = \alpha^5 \begin{bmatrix} \alpha^3 + 1 & \alpha + 1 \\ \alpha + 1 & \alpha^2 \end{bmatrix}.
\]

The solutions \(\sigma_1 = \alpha^2, \sigma_2 = \alpha^3 + \alpha^2\) are then obtained by matrix multiplication to obtain \(L(z) = z^2 + \alpha^2 z + \alpha^3 + \alpha^2\). In a general situation, one could search through \(GF(2^m)\) for the roots of \(L(z)\), but in the present context observe that \(L(z) = (z + \alpha)^2 + \alpha^2 (z + \alpha)\) so that either \(z = \alpha\) or \(z = \alpha + \alpha^2 = \alpha^5\). Thus we have determined that the errors appear in positions 1 and 5 and we decode \(r(x) = x + x^2 + x^3 + x^4 + x^5 + x^6 + x^7 + x^{10} + x^{12}\) to \(x^2 + x^3 + x^4 + x^6 + x^7 + x^{10} + x^{12} = x^2 g\).

Use both the simp command and the Rem command to check that you get the same thing. Recall that the syntax for the Rem command is \(\text{Rem}(h, k, x) \mod 2\); pay attention to the fact that the simp command will give an answer in terms of \(a\) while the Rem command will write the answer in terms of \(x\).

1. There are three roots of the polynomial \(x^3 + x^2 + 1\) in \(F\); find them. Choose one of the roots; call it \(t\). Calculate \(t, t^2, t^4\) to see that \(t, t^2, t^4\) are the three roots of \(x^3 + x^2 + 1\).

2. There are three roots of the polynomial \(x^3 + x + 1\) in \(F\); find them. Choose one of the roots; call it \(t\). Calculate \(t, t^2, t^4\) to see that \(t, t^2, t^4\) are the three roots of \(x^3 + x + 1\).
Chapter 8

Cryptography and Group Theory

In this chapter we discuss one of the main methods of encrypting data, the RSA encryption system. The algebraic structure that is at the heart of this method is that of a group. Group theory, perhaps the first algebraic structure to be studied abstractly, is one of the most fundamental of structures. As we shall see, rings, fields, and vector spaces are all special examples of groups. What distinguishes groups from these other structures is that a group has only a single operation.

Cryptography is the subject of transmitting private data in a secure manner. If you make a purchase on the internet, you need to send to the merchant a credit card number. If somebody were to intercept the transmission of this information, they would have your number. Because of this, most internet sites encrypt such data. By doing so, anybody intercepting the transmission will see a useless string of digits instead of a valid credit card number. If, however, the interceptor were to know how the merchant replaces credit card numbers with other numbers, they would have a way of recovering the number. Because of this, merchants must use methods of encryption that are very difficult to “break”. We will discuss one such system, the RSA encryption system.

8.1 The RSA encryption system

The RSA encryption system was invented in 1978 by Ron Rivest, Adi Shamir, and Leonard Adleman, and is one of the most common methods of encrypting data used today. To describe the RSA system, one starts with the following data:

- distinct prime numbers \( p \) and \( q \)
- an integer \( e \) relatively prime to \( (p - 1)(q - 1) \).

From this data we will build an encryption system. Let \( n = pq \). We will restrict our attention to encrypting numbers. This is satisfactory, since any text message can be converted to numbers by replacing each letter with an appropriate number. Let \( M \) be an integer,
considered to be a message we wish to encrypt. We then calculate $M^e \mod n$, the remainder after dividing $n$ into $M^e$. This remainder is our encrypted message.

For example, let

\[
\begin{align*}
p &= 3486784409, \\
q &= 282429536483, \\
e &= 19.
\end{align*}
\]

Then $n = pq = 984770904450021093547$. Also, $(p-1)(q-1) = 984770904164104772656$. To encrypt the message 12345, we calculate

\[
12345^{19} \mod n,
\]

which comes out to be 123355218486796132288. Therefore, if we wish to transmit the number 12345, we would instead transmit 123355218486796132288.

How does somebody receiving 123355218486796132288 know that this number represents 12345? First, by our assumption that $e$ is relatively prime to $(p-1)(q-1)$, we know that $e$ has an inverse modulo $(p-1)(q-1)$; that is, there is an integer $d$ with $ed \equiv 1 \mod (p-1)(q-1)$. If an encrypted number $N$ is received, then one calculates $N^d \mod n$; the result returns the original message. For example, from a Maple computation, we can see that

\[
d = 207320190350337846875,
\]

Thus, to recover the original message 12345, we compute

\[
123355218486796132288^{207320190350337846875} \mod 984770904450021093547 = 12345
\]

While this calculation looks formidable, Maple can do it virtually instantaneously. In fact, on an average personal computer, Maple can calculate $M^d \mod n$ in a couple of seconds even if $d$ and $n$ are 400 digit numbers, so the calculations in the RSA system are easy to do even with very large numbers.

To summarize, the RSA encryption system starts with two prime numbers $p$ and $q$ and an integer $e$ satisfying $\gcd(e, (p-1)(q-1)) = 1$. One then calculates a positive integer $d$ satisfying $ed \equiv 1 \mod (p-1)(q-1)$. One then encrypts an integer $M$ by replacing it by

\[
N = M^e \mod n.
\]

To decrypt $N$, one sees that

\[
M = N^d \mod n.
\]

Why this method works will be addressed in our study of group theory. The theoretical fact that the inventors of the RSA system, Rivest, Shamir, and Adleman, needed was Euler’s theorem. This is a special case of Lagrange’s theorem of group theory. We will introduce groups, see several examples, and prove Lagrange’s theorem in the next section. To prove this theorem we will revisit cosets, a subject we have studied for both vector spaces and for rings.
8.2 Groups

To motivate the definition of a group, we discuss the main example used in the RSA encryption system. Let $R$ be a ring. Recall that a unit of $R$ is an element having a multiplicative inverse. Recall also that if $a, b$ are units of $R$, then so is $ab$, since $ab$ has $b^{-1}a^{-1}$ as its multiplicative inverse. We denote by $R^*$ the set of all units of $R$. In other words,

$$R^* = \{a \in R : \text{there is a } c \in R \text{ with } ac = ca = 1\}.$$ 

By the statement above, if we multiply two elements of $R^*$, the result is another element of $R^*$. Therefore, multiplication induces a binary operation on the set $R^*$. We note three properties of this binary operation: multiplication on $R^*$ is associative; $1 \in R^*$, so $R^*$ has an identity; and each element of $R^*$ has an inverse in $R^*$. It is these properties that make up the definition of a group.

**Definition 8.1.** Let $G$ be a nonempty set together with a binary operation $*$ on $G$. Then the pair $(G, *)$ is said to be a group if

1. $a \ast (b \ast c) = (a \ast b) \ast c$ for all $a, b, c \in G$;
2. there is an $e \in G$ such that $e \ast a = a \ast e = a$ for all $a \in G$;
3. for each $a \in G$ there is an element $b \in G$ with $a \ast b = b \ast a = e$.

We give several examples of groups.

**Example 8.2.** Let $R$ be a ring. Then, as noted above, $(R^*, \cdot)$ is a group. The identity of $R^*$ is the multiplicative identity $1$.

**Example 8.3.** Let $R$ be a ring. Then $(R, +)$ is a group. To see this, we remark that the definition of a ring includes the three properties of a group when considering addition. Note that the identity of $(R, +)$ is $0$, the additive identity of $R$.

**Example 8.4.** Let $V$ be a vector space. Ignoring scalar multiplication and only considering addition, we see that $(V, +)$ is a group. Again, looking at the definition of a vector space and only considering addition, we have the properties in the definition of a group.

**Example 8.5.** For some special cases of the examples above, the following are groups under addition: $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, and $(\mathbb{R}, +)$. Furthermore, under multiplication, $(\mathbb{Z}^*, \cdot)$, $(\mathbb{Q}^*, \cdot)$, and $(\mathbb{R}^*, \cdot)$ are all groups. To better understand these examples, we see that $\mathbb{Z}^* = \{1, -1\}$; the only elements of $\mathbb{Z}$ that have multiplicative inverses in $\mathbb{Z}$ are $1$ and $-1$. However, since $\mathbb{Q}$ and $\mathbb{R}$ are fields, we have $\mathbb{Q}^* = \mathbb{Q} \setminus \{0\}$ and $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$. That is, every nonzero element of $\mathbb{Q}$ or of $\mathbb{R}$ has a multiplicative inverse (in $\mathbb{Q}$ or $\mathbb{R}$, respectively). This leads to the next example.
Example 8.6. Let $F$ be a field. Then $(F - \{0\}, \cdot)$ is a group under multiplication. This is because $F$ is a ring and $F^* = F - \{0\}$ by the definition of a field.

Example 8.7. Let $R = M_n(\mathbb{R})$, the ring of all $n \times n$ matrices with real number entries. Then $R^*$ is usually denoted by $\text{Gl}_n(\mathbb{R})$, and is called the general linear group of $n \times n$ matrices. By recalling properties of determinants, we may write

$$\text{Gl}_n(\mathbb{R}) = \{ A \in M_n(\mathbb{R}) : \det(A) \neq 0 \}.$$  

This is a group under matrix multiplication. Note that multiplication in this group is not commutative.

Example 8.8. Let $X$ be any set. We denote by $P(X)$ the set of all permutations of $X$. That is,

$$P(X) = \{ f : X \to X : f \text{ is a 1-1, onto function} \}.$$  

We define an operation on $P(X)$ by composition of functions; note that if $f, g$ are functions from $X$ to $X$, then so is $f \circ g$, and that if $f$ and $g$ are both 1-1 and onto, then so is $f \circ g$. Therefore, composition is indeed a binary operation on $P(X)$. We leave it as an exercise to show that $P(X)$ is a group under composition.

We now focus on the example needed for the RSA system. Let $n$ be a positive integer, and consider the ring $\mathbb{Z}_n$. We can then form the group $\mathbb{Z}_n^*$, a group under (coset) multiplication. By Corollary 1.16, if $a \in \mathbb{Z}_n$, then there is a solution to the equation $a \cdot x = 1$ if $\gcd(a, n) = 1$. That is, $a$ has a multiplicative inverse in $\mathbb{Z}_n$ if $\gcd(a, n) = 1$. The converse is also true; if $\gcd(a, n) = 1$, then we may write $1 = ax + by$ for some integers $x, y$, by Proposition 1.12. Then $1 = a \cdot x$, so $a$ has a multiplicative inverse. We thus have the following description:

$$\mathbb{Z}_n^* = \{ a \in \mathbb{Z}_n : \gcd(a, n) = 1 \}.$$  

How many elements does this group have? For example, $\mathbb{Z}_4^* = \{ 1, 3 \}$ has two elements, $\mathbb{Z}_6^* = \{ 1, 5 \}$ also has two elements, and $\mathbb{Z}_{11}^* = \mathbb{Z}_{11} - \{ 0 \}$ has ten elements. In general, the number of elements of $\mathbb{Z}_n^*$ is equal to the number of integers $a$ satisfying $1 \leq a \leq n$ and $\gcd(a, n) = 1$. This number is referred to as $\phi(n)$, and the function $\phi$ is called the Euler phi function. While there is a general formula for computing $\phi(n)$ from the prime factorization of $n$, we concern ourselves only with two cases. First, if $p$ is a prime, then $\phi(p) = p - 1$. This is clear since all integers $a$ with $1 \leq a \leq p$ except for $p$ itself are relatively prime to $p$. The next case we state as a lemma.

Lemma 8.9. Let $p$ and $q$ be distinct primes. Then $\phi(pq) = (p - 1)(q - 1)$.

Proof. Set $n = pq$. To count $\phi(n)$, we first count the number of integers between 1 and $n$ and not relatively prime to $n$. If $1 \leq a \leq n$, then $\gcd(a, n) > 1$ only if $p$ divides $a$ or $q$ divides $a$. The multiples of $p$ between 1 and $n$ are then

$$p, 2p, \ldots , (q - 1)p, qp = n,$$
so there are \( q \) multiples of \( p \) between 1 and \( n \). The multiples of \( q \) in this range are

\[ q, 2q, \ldots, (p - 1)q, pq = n, \]

so there are \( p \) multiples of \( q \) in this range. The only number on both lists is \( n \); this follows from unique factorization. Therefore, there are \( p + q - 1 \) integers between 1 and \( n \) that are not relatively prime to \( n \). Since there are \( n = pq \) numbers total in this range, we see that

\[
\phi(n) = pq - (p + q - 1) = pq - p - q + 1
= p(q - 1) - (q - 1) = (p - 1)(q - 1),
\]

as desired.

To summarize, the group \( \mathbb{Z}_n^* \) has \( \phi(n) \) elements, and if \( n = pq \) is the product of two distinct primes, then \( \mathbb{Z}_{pq}^* \) has \( (p - 1)(q - 1) \) elements. The significance of this result and its application to the RSA encryption system will become clear when we prove Lagrange’s theorem. To do this, we first need to discuss subgroups. This concept is the analogue in group theory of subspaces of a vector space.

Before going further, we make a note about notation and terminology. Some of our examples of groups use addition as the operation and other examples use multiplication. Unless we have a specific group in which the operation is addition, we will use multiplicative notation for the operation. In particular, we will write \( ab \) or \( a \cdot b \) for the product of \( a \) and \( b \) and \( a^{-1} \) for the inverse of \( a \). As for terminology, unless the symbol for the operation needs to be specified, we will refer to a group by a single symbol, such as \( G \), rather than by a pair such as \( (G, \cdot) \). We need to be careful to remember that a group is not just a set but is a set together with a binary operation.

**Definition 8.10.** Let \( G \) be a group. A nonempty subset \( H \) of \( G \) is said to be a subgroup of \( G \) if the operation on \( G \) restricts to an operation on \( H \), and if \( H \) is a group with respect to this restricted operation.

For example, consider the group \( \mathbb{Z} \) under addition and let \( H \) be the set of even integers. Addition restricts to an operation on \( H \) because the sum of two even integers is again even. Just as there is a theorem helping us determine when a subset of a vector space is a subspace (Lemma 4.9), there is a result that helps us determine when a subset of a group is a subgroup.

**Lemma 8.11.** Let \( G \) be a group, and let \( H \) be a nonempty subset of \( G \). Then \( H \) is a subgroup of \( G \) provided that the following two conditions hold: (i) if \( a, b \in H \), then \( ab \in H \), and (ii) if \( a \in H \), then \( a^{-1} \in H \).

**Proof.** Suppose a subset \( H \) of a group \( G \) satisfies the two conditions in the statement. The first says that the operation on \( G \) restricts to an operation on \( H \), so we have a binary operation on \( H \). We need to verify for \( H \) the three axioms in the definition of a group. Associativity is clear; if \( a, b, c \in H \), then \( a, b, c \in G \), so \( a(bc) = (ab)c \) since \( G \) is a group.
Next, Condition (ii) ensures that every element of $H$ has an inverse in $H$. So, the only thing remaining is to see that $H$ has an identity. We do this by proving that if $e$ is the identity of $G$, then $e \in H$. To see why this is true, first note that there is an element $a \in H$ because $H$ is nonempty. By Condition (ii), $a^{-1} \in H$. Then, by Condition (i), $a \cdot a^{-1} \in H$. But $a \cdot a^{-1} = e$, so $e \in H$.

There is a particularly nice construction of subgroups. Let $G$ be a group and let $a \in G$. Consider trying to build a subgroup containing $a$. By Condition (i) of the Lemma above, we see that the subgroup has to contain $a \cdot a = a^2$. Using the condition again, the subgroup has to contain $a^2 \cdot a = a^3$. The subgroup must also contain $a^2$. Then by condition (i), it must contain $a^{-1} \cdot a^{-1} = a^{-2}$. Continuing with this idea leads us to the following definition. If $G$ is a group and $a \in G$, then the cyclic subgroup generated by $a$ is the set

$$\langle a \rangle = \{a^n : n \in \mathbb{Z} \}.$$  

To formally define $a^n$, we first define $a^0 = e$, the identity of $G$. If $n$ is a positive integer, then we define, inductively, $a^{n+1} = a^n \cdot a$. Therefore, $a^1 = a^0 \cdot a = e \cdot a = a$, and $a^2 = a \cdot a$, and so on. For negative exponents, if $n > 0$, we set $a^{-n} = (a^n)^{-1}$. The following laws of exponents are consequences of the definition of a group (these are left as homework problems):

$a^n \cdot a^m = a^{n+m}$, $(a^n)^m = a^{nm}$, and $(a^n)^{-1} = (a^{-1})^n$ for any $n, m \in \mathbb{Z}$. From these it follows that $\langle a \rangle$ is a subgroup of $G$ by the lemma.

**Example 8.12.** Let $G = \mathbb{Z}_8^*$. Then $G = \{\bar{1}, \bar{3}, \bar{5}, \bar{7}\}$. We calculate the cyclic subgroups of $G$. First, since $\bar{1}^n = \bar{1}$ for all $n$, we have $\langle \bar{1} \rangle = \{\bar{1} \}$ contains only the identity $\bar{1}$. Next, for $\bar{3}$, we see that

$$\bar{3}^1 = \bar{3},$$
$$\bar{3}^2 = \bar{9} = \bar{1},$$
$$\bar{3}^3 = \bar{27} = \bar{3},$$
$$\bar{3}^4 = \bar{81} = \bar{1}.$$  

You may recognize a pattern. Since $\bar{3}^2 = \bar{1}$, the identity of $G$, if $n = 2m$ is even, then $\bar{3}^n = \bar{3}^{2m} = (\bar{3}^2)^m = \bar{1}^m = \bar{1}$. If $n = 2m + 1$ is odd, then

$$\bar{3}^n = \bar{3}^{2m+1} = \bar{3} \cdot \bar{3}^{2m} = \bar{3} \cdot \bar{1} = \bar{3}.$$  

Therefore, the only powers of $\bar{3}$ are $\bar{1}$ and $\bar{3}$. Thus, $\langle \bar{3} \rangle = \{\bar{1}, \bar{3}\}$. If we do similar calculations for $\bar{5}$ and $\bar{7}$, we will see similar patterns, for $\bar{5}^2 = \bar{1}$ and $\bar{7}^2 = \bar{1}$. Thus, the only powers of $\bar{5}$ will be $\bar{1}$ and $\bar{5}$, and the only powers of $\bar{7}$ are $\bar{1}$ and $\bar{7}$.

**Example 8.13.** Let $G = \mathbb{Z}_{10}^*$. Then $G = \{\bar{1}, \bar{3}, \bar{7}, \bar{9}\}$ has four elements. As with the previous example (and any group in fact), the cyclic subgroup generated by $\bar{1}$ is just $\{\bar{1}\}$.
Next consider \( \bar{3} \). We have
\[
\begin{align*}
\bar{3}^1 &= \bar{3}, \\
\bar{3}^2 &= \bar{9}, \\
\bar{3}^3 &= \bar{27} = \bar{7}, \\
\bar{3}^4 &= \bar{81} = 1.
\end{align*}
\]
We have therefore produced all four elements of \( G \) as powers of \( \bar{3} \); therefore, \( \langle \bar{3} \rangle = G \). A similar calculation will show that \( \langle \bar{7} \rangle = G \); this can also be seen by the fact that \( \bar{7} = \bar{3}^{-1} \) and \( \langle a \rangle = \langle a^{-1} \rangle \) for any group \( G \) and any \( a \in G \). Finally, for \( \bar{9} \), we have
\[
\begin{align*}
\bar{9}^1 &= \bar{9}, \\
\bar{9}^2 &= \bar{81} = 1.
\end{align*}
\]
As with the previous example, even powers of \( \bar{9} \) are 1 and odd powers of \( \bar{9} \) are \( \bar{9} \), so \( \langle \bar{9} \rangle = \{1, \bar{9}\} \).

We make some numerical observations from the previous two examples. The first thing to notice is that the number of elements in a given cyclic subgroup \( \langle a \rangle \) turned out to be the same as the first positive number \( n \) for which \( a^n \) was the identity. Second, the number of elements in each cyclic subgroup was a divisor of the number of elements of the given group. Both of these facts are not coincidences; they are general facts that we will now prove. The second is in fact Lagrange’s theorem. For a piece of notation, we will write \( |X| \) for the number of elements in a set \( X \). For a group \( G \), the number \( |G| \) is often called the order of \( G \).

**Lemma 8.14.** Let \( G \) be a finite group and let \( a \in G \). If \( n = \min \{m : m > 0, a^m = e\} \), then \( n = |\langle a \rangle| \), the number of elements in the cyclic subgroup generated by \( a \).

**Proof.** We will prove the lemma by proving that \( \langle a \rangle = \{a^r : 0 \leq r < n\} \) and that these elements are all distinct. First, any element of \( \langle a \rangle \) is of the form \( a^s \) for some integer \( s \). By the division algorithm, we may write \( s = qn + r \) with \( 0 \leq r < n \). Then
\[
a^s = a^{qn+r} = a^{qn}a^r = (a^n)^q(a^r) = a^r
\]
since \( a^n = e \), so \( (a^n)^q = e \). Therefore, \( a^s \) can be written as a power \( a^r \) of \( a \) with \( 0 \leq r < n \). This proves the first claim. For the second, suppose that \( a^r = a^t \) with \( 0 \leq r, t < n \). Suppose that \( r \leq t \). Then, by the laws of exponents, \( e = a^r a^{-r} = a^{t-r} \). Since \( n \) is the smallest positive integer satisfying \( a^n = e \), and since \( 0 \leq t - r < n \), we must have \( t - r = 0 \). Thus, \( t = r \). So, the elements \( a^0, a^1, \ldots, a^{n-1} \) are all distinct. Since they form \( \langle a \rangle \), we have proved that \( |\langle a \rangle| = n \). \( \square \)
We now consider Lagrange’s theorem. To do so we need to look at a concept we have seen twice before in the case of vector spaces and for rings. If $H$ is a subgroup of a group $(G, \ast)$, and if $a \in G$, then the coset of $H$ generated by $a$ is the set

$$H \ast a = \{ h \ast a : h \in H \}.$$ 

Cosets are equivalence classes for the following equivalence relation: for $a, b \in G$, define $a \sim b$ if $ab^{-1} \in H$. Then $\sim$ is an equivalence relation (see homework problems) and the equivalence class of $a$ is the coset $H \ast a$. Therefore, the cosets of $H$ form a partition for the group $G$. If we write the group operation as multiplication, we will usually write $Ha$ for the coset of $a$.

**Theorem 8.15 (Lagrange).** Let $G$ be a finite group and let $H$ be a subgroup of $G$. Then $|H|$ divides $|G|$.

**Proof.** We prove this by showing that each coset has $|H|$ elements. From this it will follow that $|G|$ is equal to $|H|$ times the number of cosets, and this will prove the theorem. To do this, let $a \in G$. We wish to prove that $|Ha| = |H|$. One way to prove that two sets have the same size is to produce a 1-1 onto function between them. We do this here by defining a function $f : H \rightarrow Ha$ by $f(h) = ha$. This is 1-1 since if $f(h) = f(k)$, then $ha = ka$. Multiplying both sides on the right by $a^{-1}$ yields $h = k$, so $f$ is 1-1. The function $f$ is also onto since, if $x \in Ha$, then $x = ha$ for some $h \in H$, and so $x = f(h)$. Since $f$ is then a 1-1 onto function from $H$ to $Ha$, we have proven that $|H| = |Ha|$, as desired. □

In fact, we can be a little more detailed. We let $[G : H]$ be the number of cosets of $H$ in $G$. This number is often called the index of $H$ in $G$. The proof of Lagrange’s theorem shows that all cosets have the same number of elements. Therefore, $|G|$ is equal to the product of the size of any one coset with the number of cosets. In other words, since $H = He$ is a coset, $|G| = [H] \cdot [G : H]$.

We can combine Lagrange’s theorem with the previous lemma to get a result key in the RSA encryption system.

**Corollary 8.16.** Let $G$ be a finite group with $n = |G|$. If $a \in G$, then $a^n = e$.

**Proof.** Let $m = |\langle a \rangle|$. By Lagrange’s theorem, $m$ divides $n$, so $n = mt$ for some integer $t$. By the lemma, $a^m = e$. Therefore, $a^n = a^{mt} = (a^m)^t = e^t = e$, as desired. □

As a special case of this corollary, we obtain Euler’s theorem.

**Corollary 8.17 (Euler’s Theorem).** Let $n$ be a positive integer. If $a$ is an integer with $\gcd(a, n) = 1$, then $a^{\phi(n)} \equiv 1 \mod n$.

**Proof.** If $\gcd(a, n) = 1$, then $a \in \mathbb{Z}_n^*$, a group of order $\phi(n)$. The previous corollary tells us that $a^{\phi(n)} = 1$. By definition of coset multiplication, $a^{\phi(n)} = a^{\phi(n)}$. The equation $a^{\phi(n)} = 1$ is equivalent to the relation $a^{\phi(n)} \equiv 1 \mod n$. □
We are now in a position to see how group theory will tell us that the method of decrypting in the RSA system recovers the original message. Let $G = \mathbb{Z}_n^*$, where $n = pq$ is the product of two distinct prime numbers. As we have seen, $|G| = \phi(n) = (p - 1)(q - 1)$. We have an integer $e$ satisfying $\gcd(e, (p - 1)(q - 1)) = 1$. Therefore, there is an integer $d$ satisfying $ed \equiv 1 \mod \phi(n)$. We may write $1 = ed + s\phi(n)$ for some integer $s$. The claim of the RSA system is that, for any message $M$, we have $(M^e)^d \equiv 1 \mod n$. Written another way, it claims that $(\overline{M}^e)^d = \overline{M}$. Assuming that $M$ is not divisible by $p$ or $q$, we have $\overline{M} \in \mathbb{Z}_n^*$. Therefore

$$\overline{M} = \overline{M}^{ed + s\phi(n)} = \overline{M}^{ed} \overline{M}^{s\phi(n)} = \overline{M}^{ed}(\overline{M}^{\phi(n)})^s = \overline{M}^{ed} \overline{M}^{\phi(n)}^s = \overline{M}^{ed} \overline{M}^{\phi(n)^s} = \overline{M}^{ed} \overline{M}^1 = \overline{M}^{ed}$$

since $\overline{M}^{\phi(n)} = 1$ by the corollary to Lagrange’s theorem. Thus, $(\overline{M}^e)^d = \overline{M}$, and so the decryption in RSA recovers the original message.

In the argument above, we assumed that $M$ was not divisible by either $p$ or $q$ in order to conclude that decryption would recover $M$. This is not a necessary assumption, but it makes the argument a little simpler. The general argument is left to the exercises.

**Exercises**

1. Prove the cancellation laws for a group if $G$ is a group and $a, b, c \in G$, then prove (i) if $ab = ac$ in a group, then $b = c$, and (ii) if $ba = ca$, then $b = c$.

2. Let $X$ be a set. If $P(X)$ is the set of all 1-1 onto functions from $X$ to $X$, prove that $P(X)$ is a group under composition of functions. You may wish to recall some fact about 1-1 onto functions. Make sure to say why composition does give you a binary operation on $P(X)$.

3. Let $G$ be a group. If $a \in G$ and $n, m$ are positive integers, prove that

   (a) $a^n a^m = a^{n+m}$,

   (b) $(a^n)^m = a^{nm}$.

   (Hint: recall the inductive definition of $a^n$. For (a), use induction on $m$ for (a). Then use (a) and induction on $m$ for (b).)

4. Let $G$ be a group. If $a, b \in G$ with $ab = ba$, show that $(ab)^n = a^n b^n$.

5. Let $G$ be a group. If $a \in G$ and $n$ is a positive integer, prove that $(a^n)^{-1} = (a^{-1})^n$.

6. Let $G$ be a group and let $H$ be a subgroup of $G$. Define a relation $\sim$ on $G$ by $a \sim b$ if $ab^{-1} \in H$. Prove that $\sim$ is an equivalence relation.
7. With notation in the previous problem, if \( a \in G \), show that \( Ha \) is the equivalence class of \( a \).

8. Using your personal RSA data, encrypt the number \( M = 123456789 \). Then decrypt the result and see that you recover \( M \). Recall that, in Maple, to efficiently calculate \( M^e \mod n \), type \( M \&^e \mod n \).

9. Write out the multiplication table for \( \mathbb{Z}_{15} \).

10. Find the cyclic subgroups \( \langle 4 \rangle \) and \( \langle 7 \rangle \) of \( \mathbb{Z}_{15}^* \).

11. Write out all the cosets in \( \mathbb{Z}_{15}^* \) for the subgroup \( \langle 2 \rangle \).

12. If \( G \) is a group with \( |G| = p \), a prime number, show that if \( a \in G \) is different from \( e \), then \( \langle a \rangle = G \).

13. In doing RSA calculations, if you can find \( m = (p - 1)(q - 1) \), then you can break the encryption because it is easy to find the decoding integer (Maple does it via the igcdex command). Show that if you know \( m \) (and the public information \( n = pq \)), then you can find easily \( p \) and \( q \). Give an example of how to do this by finding \( p \) and \( q \) from the values of \( n \) and \( m \) in the Maple worksheet Assignment11.mws.

(Hint: you cannot do this with the factor command applied to the \( n \) in that worksheet because it is too large!)

14. Let \( n = pq \) with \( p, q \) distinct primes. If \( a, b \) are integers, show that \( a \equiv b \mod n \) if and only if both \( a \equiv b \mod p \) and \( a \equiv b \mod q \).

15. Referring to the notation of the previous problem, if the integer \( M \) is divisible by \( p \) or \( q \), verify that \( M^{ed} \equiv M \mod n \) is true.

### 8.3 Secure Signatures with RSA

One issue of data transmission is the ability to verify a person’s identity. If I send a request to a bank to transfer money out of an account, the bank wants to know if I am the owner of the account. If I make the request over the internet, how can the bank check my identity? The RSA encryption system gives a method for checking identities, which is one of the important features of the system.

Suppose that person \( A \) transmits data to person \( B \), and that person \( B \) wants a method to check the identity of person \( A \). To do this, both person \( A \) and \( B \) get sets of RSA data; person \( A \) has a modulus \( n_A \) and an encryption exponent \( e_A \). These are publicly available. That person also has a decryption exponent \( d_A \) that remains private. Person \( B \) similarly has data \( n_B, e_B, \) and \( d_B \). In addition, person \( A \) has a signature, a publicly available number \( S \). To convince person \( B \) of his identity, person \( A \) first calculates \( T = S^{d_A} \mod n_A \) and then
8.3. SECURE SIGNATURES WITH RSA

\[ R = T^{e_B} \mod n_B. \]

He then transmits \( R \) to person \( B \). Person \( B \) then decrypts \( R \) with her data, recovering \( T = R^{d_B} \mod n_B \). Finally, she encrypts \( T \) with person \( A \)'s data, obtaining \( T^{e_A} \mod n_A = S \). By seeing that this result is the signature of person \( A \), the identity has been validated.

For example, suppose that the data for person \( A \) is

\[
\begin{align*}
 n_A &= 2673157 \\
 e_A &= 23 \\
 d_A &= 2437607 \\
 S &= 837361
\end{align*}
\]

and the data for person \( B \) is

\[
\begin{align*}
 n_B &= 721864639 \\
 e_B &= 19823 \\
 d_B &= 700322447
\end{align*}
\]

Person \( A \) then calculates

\[
837361^{2437607} \mod 2673157 = 1216606,
\]

and then

\[
1216606^{19823} \mod 721864639 = 241279367.
\]

Person \( A \) then transmits 241279367 to person \( B \). When person \( B \) receives this, she calculates

\[
241279367^{700322447} \mod 721864639 = 1216606,
\]

and finally recovers \( S \) as \( S = 1216606^{23} \mod 2673157 \).

To explain why this works, we denote by encrypt\(_A\)(\( M \)) and decrypt\(_A\)(\( M \)) the integers \( M^{e_A} \mod n_A \) and \( M^{d_A} \mod n_A \), respectively. We similarly have encrypt\(_B\)(\( M \)) and decrypt\(_B\)(\( M \)). The validity of the RSA system says that

\[
\begin{align*}
 \text{decrypt}_A(\text{encrypt}_A(M)) &= M, \\
 \text{encrypt}_A(\text{decrypt}_A(M)) &= M.
\end{align*}
\]

Similar equations hold for \( B \). With this notation, person \( A \) calculates

\[
R = \text{encrypt}_B(\text{decrypt}_A(S))
\]

and then person \( B \) calculates

\[
\text{encrypt}_A(\text{decrypt}_B(R)).
\]
Therefore, person $B$ will calculate

$$\text{encrypt}_A(\text{decrypt}_B(\text{encrypt}_B(\text{decrypt}_A(S)))) = \text{encrypt}_A(\text{decrypt}_A(S))$$

$$= S$$

because of the equations above. Therefore, person $B$ does recover the signature of person $A$.

The reason that this method validates the identity of person $A$ is because only person $A$ can calculate $\text{decrypt}_A(S)$. If another person tries to claim he is person $A$, tries to substitute a number $F$ in place of $\text{decrypt}_A(S)$, he will transmit $\text{encrypt}_B(F)$ to person $B$. Person $B$ will then calculate

$$\text{encrypt}_A(\text{decrypt}_B(\text{encrypt}_B(F))) = \text{encrypt}_A(F).$$

However, in order to have $\text{encrypt}_A(F) = S$, we must have

$$\text{decrypt}_A(S) = \text{decrypt}_A(\text{encrypt}_A(F))$$

$$= F,$$

which means that this person has to have the correct decrypted number $\text{decrypt}_A(S)$; he cannot send any other number without person $B$ realizing it is a fake number.
Chapter 9

Symmetry

In this chapter we explore a connection between algebra and geometry. One of the main topics of plane geometry is that of congruence; roughly, two geometric figures are said to be congruent if one can be moved to coincide exactly with the other. We will be more precise below in our description of congruence, and investigating this notion will lead us to new examples of groups.

9.1 Isometries

Consider the following pair of triangles.

They are congruent since the first can be moved to coincide with the second. However, the next pair of triangles are not congruent.

In order to move the first to coincide with the second, one would also have to reshape the triangle. One way of seeing that they are not congruence is that the distances between the vertices of the first triangle are not the same as the distances between the vertices of the second. We use the idea of distance to make more precise the idea of congruence.
Definition 9.1. An isometry of the plane is a 1-1 and onto map of the plane to itself that preserves distances. That is, an isometry \( f \) is a 1-1 onto function of the plane such that, for every pair \( P, Q \) of points, the distance between \( P \) and \( Q \) is equal to the distance between \( f(P) \) and \( f(Q) \).

Two geometric figures are then congruent if there is an isometry that maps one figure exactly to the other. By representing points as ordered pairs of real numbers, we can give algebraic formulas for distance between points. First, let \( k_P \) be the distance from \( P \) to the origin; this is the length of \( P \). The distance between \( P = (a, b) \) and \( Q = (c, d) \) is then given by the distance formula

\[
k_P Q = \sqrt{(a - c)^2 + (b - d)^2}.
\]

In terms of groups, what is important about isometries that the composition of two isometries is again an isometry, as we will prove shortly. Note that since an isometry is a 1-1 and onto function, it has an inverse function.

Lemma 9.2. (1) The composition of two isometries of the plane is again an isometry. (2) The inverse of an isometry is an isometry.

Proof. To prove (1), let \( f \) and \( g \) be isometries of the plane and let \( P, Q \) be points. Then

\[
\|(f \circ g)(P) - (f \circ g)(Q)\| = \|g(P) - g(Q)\| = \|P - Q\|
\]

since \( f \) and \( g \) are each isometries. Therefore, \( f \circ g \) preserves distance. Moreover, \( f \circ g \) is the composition of two 1-1 and onto functions, so it is also 1-1 and onto. Therefore, \( f \circ g \) is an isometry. For part (2), let \( f \) be an isometry, and let \( f^{-1} \) be its inverse function. Let \( P \) and \( Q \) be points. We need to prove that \( \|f^{-1}(P) - f^{-1}(Q)\| = \|P - Q\| \). Let \( P' = f^{-1}(P) \) and \( Q' = f^{-1}(Q) \). Then \( P = f(P') \) and \( Q = f(Q') \) by definition of \( f^{-1} \). Since \( f \) is an isometry, \( \|P' - Q'\| = \|f(P') - f(Q')\| \). In other words, \( \|f^{-1}(P) - f^{-1}(Q)\| = \|P - Q\| \). This is exactly what we need to see that \( f^{-1} \) preserves distance. Since \( f^{-1} \) is also 1-1 and onto, it is an isometry.

Let Isom(\( \mathbb{R}^2 \)) be the set of all isometries of the plane \( \mathbb{R}^2 \). The lemma above shows that composition of functions is a binary operation on Isom(\( \mathbb{R}^2 \)). Associativity of composition always holds among functions, the identify function is clearly an isometry, and the lemma also shows that every element of Isom(\( \mathbb{R}^2 \)) has an inverse. Therefore, Isom(\( \mathbb{R}^2 \)) is a group.

9.2 Symmetry Groups

Let \( X \) be a subset of the plane. We associate a group to \( X \), called the symmetry group of \( X \). This notion makes perfect sense for subsets of \( \mathbb{R}^n \) for any \( n \); but, we will restrict our attention to plane figures.
Definition 9.3. If \( X \) is a subset of the plane, then the symmetry group of \( X \) is the set of all isometries \( f \) of the plane for which \( f(X) = X \). This group is denoted by \( \text{Sym}(X) \).

To better understand the definition, we look at it more carefully. If \( f \) is an isometry and \( X \) is a subset of the plane, then \( f(X) = \{ f(P) : P \in X \} \). Therefore, \( f \in \text{Sym}(X) \) if for every point \( P \in X \), we have \( f(P) \in X \) and, for every \( Q \in X \) there is a \( P \in X \) with \( f(P) = Q \). We do not require that \( f(P) = P \) for \( P \in X \). For example, if \( X = \{(0,1), (1,0), (-1,0), (0,-1)\} \), then a rotation of \( 90^\circ \) about the origin sends this set of four points to itself. Thus, this rotation is a symmetry of \( X \). To see, in general, that \( \text{Sym}(X) \) is a subgroup of \( \text{Isom}(\mathbb{R}^2) \), we only need to show that if \( f, g \in \text{Sym}(X) \), then \( f \circ g^{-1} \in \text{Sym}(X) \). However, if \( f, g \in \text{Sym}(X) \), then \( f(X) = g(X) = X \). Therefore, \( g^{-1}(X) = X \), and so \( f(g^{-1}(X)) = f(X) = X \).

We start by giving some examples of symmetries.

**Translations.** Let \( v \) be a fixed element of \( \mathbb{R}^2 \). The function \( f(x) = x + v \) is the translation by the vector \( v \). It is trivial to see that \( f \) is an isometry. Its inverse is translation by \( -v \). We will denote the translation by a vector \( v \) by \( \tau_v \).

![Translation example](https://via.placeholder.com/150)

**Reflections.** Let \( \ell \) be a line in \( \mathbb{R}^2 \). Then the map that reflects points across \( \ell \) is an isometry.
If \( \ell \) is the line through the origin parallel to a vector \( w \), then the reflection across \( \ell \) is given by the formula

\[
f(x) = 2 \left( \frac{x \cdot w}{w \cdot w} \right) w - x.
\]

This formula comes from the formula for projection of one vector onto another that one sees in multivariable calculus. From it one can see that \( f \) is an isometry. A reflection \( f \) satisfies \( f^2 = \text{id} \).

**Rotations.** If \( \theta \) be an angle, then the rotation (counterclockwise) by an angle \( \theta \) about the origin is given in coordinates by

\[
r(x, y) = \left[ \begin{array}{cc} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{array} \right] \left( \begin{array}{c} x \\ y \end{array} \right)^T = (x \cos \theta - y \sin \theta, x \sin \theta + y \cos \theta).
\]

From this formula one can see that a rotation about the origin is an isometry. We can use this to describe a rotation about any point. If \( r' \) is the rotation by \( \theta \) about a point \( P \in \mathbb{R}^2 \), and if \( t \) is the translation by \( P \), then \( r' = t \circ r \circ t^{-1} \). As a consequence, this shows that any rotation is an isometry. Note that \( r^{-1} \) is rotation by \(-\theta\) about the origin. If \( \theta = 2\pi/n \) for some integer \( n \), then \( r^n = \text{id} \).

**Glide Reflections.** We can produce new isometries by composition. One can check that the composition of a rotation and a translation is again a rotation, and that the composition of a rotation and a reflection is either a rotation or a translation. The composition of a reflection and a translation may be a reflection, although it may also be a new type of isometry. We will call such a composition a glide reflection if it is not a reflection.
We will show later that any isometry of $\mathbb{R}^2$ is a composition of a translation with either a rotation or a reflection. Therefore, we have accounted for all types of isometries of the plane.

Exercises

1. Prove that the composition of a rotation and a translation is again a rotation.

2. Let $f_1$ and $f_2$ be reflections about the lines $\ell_1$ and $\ell_2$, respectively. Suppose that $\ell_1$ and $\ell_2$ are parallel and that $b$ is the vector perpendicular to these lines such that translation by $b$ sends $\ell_1$ to $\ell_2$. Show that $f_2f_1$ is translation by $2b$.

3. Let $r$ be a rotation about a point $P$ and let $f$ be a reflection. Prove that $fr$ is a rotation about $P$ if $f(P) = P$, but that $fr$ is a translation if $f(P) \neq P$.

4. Let $f$ be a reflection about a line $\ell$ and let $\tau$ be a translation by a vector $b$. If $b$ is parallel to $\ell$, show that $\tau f$ is a reflection about the line $\ell + b/2$.

5. Let $f$ be a reflection about a line $\ell$ and let $\tau$ be a translation by a vector $b$. If $b$ is not parallel to $\ell$, show that $\tau f$ fixes no point, so is not a reflection. Moreover, if $b = b_1 + b_2$ with $b_1$ parallel to $\ell$ and $b_2$ perpendicular to $\ell$, show that $\tau f$ is the composition $\tau'f'$ of the translation $\tau'$ by $b_2$ and $f'$ is the reflection about the line $\ell + b_1/2$.

(By this problem, it follows that any glide reflection is the composition of a reflection followed by a translation by a vector perpendicular to the reflection line.)

9.3 Examples of Symmetry Groups

The examples we give in this section will help us to understand the definition of symmetry group and to introduce as symmetry groups two important classes of groups, the cyclic and dihedral groups.

Example 9.4. We calculate the symmetry group of an equilateral triangle $T$. For convenience, we view the origin as the center of the triangle.
The identity function is always a symmetry of any figure. Also, the rotations about the origin by an angle of $120^\circ$ or $240^\circ$ are elements of $\text{Sym}(T)$. Furthermore, the three reflections across the three dotted lines of the following picture are also isometries.

We have so far found six isometries of the triangle. We claim that $\text{Sym}(T)$ consists precisely of these six isometries. To see this, first note that any isometry must send the set $\{A, B, C\}$ of vertices to itself. There are only six 1-1 functions that send $\{A, B, C\}$ to itself, and any isometry is determined by what it does to three non-collinear points, by Proposition ??, so we have found all symmetries of the triangle.

If $r$ is the rotation by $120^\circ$ and $f$ is any of the reflections, then an exercise will show that $f, rf, r^2f$ are the three reflections. Moreover, $r^2$ is the $240^\circ$ rotation. Therefore, $\text{Sym}(T) = \{e, r, r^2, f, rf, r^2f\}$. The elements $r$ and $f$ satisfy $r^3 = e$ and $f^2 = e$. Another exercise will show that $r$ and $f$ are related via the relation $fr = r^2f$. Alternatively, this equation may be rewritten as $frf = r^{-1}$, since $f^{-1} = f$ and $r^2 = r^{-1}$.

**Example 9.5.** Now let us determine the symmetry group of a square, centered at the origin for convenience. As with the triangle, we see that any isometry that preserves the square must permute the four vertices. There are 24 permutations of the vertices. However, not all come from isometries. First, the rotations of $0^\circ$, $90^\circ$, $180^\circ$, and $270^\circ$ about the origin are symmetries of the square. Also, the reflections about the four dotted lines in the picture below are also symmetries.
We now have eight symmetries of the square and we claim that this is all. To see this, we give a counting argument. There are four choices for where vertex $A$ can be sent since it must go to one of the vertices. Once a choice has been made, there are just two choices for where $B$ is sent since it must go to a vertex adjacent to the image of $A$; this is forced upon us since isometries preserve distance. After images for $A$ and $B$ have been chosen, the images of the other two vertices are then fixed; $C$ is sent to the vertex across from the image of $A$, and $D$ is sent to the vertex across from the image of $B$. So, there is a total of $4 \cdot 2 = 8$ possible isometries. Since we have found eight, we have them all.

If $r$ is the rotation by $90^\circ$ and $f$ is any reflection, then the four reflections are $f, rf, r^2f, r^3f$; this can be seen by an exercise similar to that needed in the previous example. The four rotations are $r, r^2, r^3, r^4 = e$. Thus, the symmetry group of the square is $\{e, r, r^2, r^3, f, rf, r^2f, r^3f\}$. We have $r^4 = e$ and $f^2 = e$, and an exercise shows that $fr = r^3f$, or $frf = r$. This is the same relation that holds for the corresponding elements in the previous example.

**Example 9.6.** If we generalize the previous two examples by considering the symmetry group of a regular $n$-gon, then we would find that the symmetry group has $2n$ elements. If $r$ is a rotation by $360^\circ/n$ and if $f$ is a reflection, then the rotations in the group are the powers $r, r^2, \ldots, r^{n-1}, r^n = e$ of $r$, and the reflections are $f, rf, \ldots, r^{n-1}f$. Thus, half of the elements are rotations and half are reflections. The elements $r$ and $f$ satisfy the relations $r^n = e$, $f^2 = e$, and $frf = r^{-1}$. This group is commonly called the *Dihedral group*, and is denoted $D_n$. We also consider the subgroup of all rotations in $D_n$. This is the cyclic subgroup $\langle r \rangle = \{e, r, \ldots, r^{n-1}\}$. We denote this group by $C_n$. This group also arises as a symmetry group. We show how $C_4$ arises in the following example.

**Example 9.7.** The following figure has only rotations in its symmetry group; because it has rotations only of $0^\circ$, $90^\circ$, $180^\circ$, and $270^\circ$, its symmetry group is $C_4$. 

\begin{center}
\includegraphics[width=0.2\textwidth]{square.png}
\end{center}
By drawing similar but more complicated pictures, we can represent $C_n$ as a symmetry group of some plane figure.

**Example 9.8.** The Zia symbol has symmetry group $D_4$; we see that besides rotations by $0^\circ$, $90^\circ$, $180^\circ$, and $270^\circ$, it has horizontal, vertical, and diagonal reflections in its symmetry group, and no other symmetry.

![Zia symbol](image)

**Example 9.9.** Consider a circle of radius 1 centered at the origin 0. Any isometry of the circle must map the center to itself. Thus, any such isometry is linear, by Proposition ??%. Conversely, if $\varphi$ is any linear isometry, and if $P$ is any point with $d(P, 0) = 1$, then $d(\varphi(P), \varphi(0)) = d(\varphi(P), 0) = 1$. Thus, $\varphi$ sends the circle to itself. Therefore, the symmetry group of the circle is $O_2(\mathbb{R})$.

**Exercises**

1. Let $G$ be the symmetry group of an equilateral triangle. If $r$ is a $120^\circ$ rotation and $f$ is a reflection, show that $f$, $rf$, and $r^2f$ are the three reflections of $G$.

2. Show that there are exactly six 1-1 functions from a set of 3 elements to itself.

3. Let $G$ be the symmetry group of an equilateral triangle. If $r$ is a $120^\circ$ rotation and $f$ is a reflection, so that $fr = r^2f$. Use this to show that $frf = r^{-1}$.

4. Let $G$ be the symmetry group of a square. If $r$ is a $90^\circ$ rotation and $f$ a reflection, show that the four reflections of $G$ are $f$, $rf$, $r^2f$, and $r^3f$.

5. Let $G$ be the symmetry group of a square. If $r$ is a $90^\circ$ rotation and $f$ a reflection, show that $fr = r^3f$. Use this to show that $frf = r^{-1}$.

**9.4 Symmetry Groups of Bounded Figures**

In this section we classify the symmetry groups of a bounded plane figure. We will see that any such symmetry group is isomorphic to a subgroup of $O_2(\mathbb{Z})$. Moreover, if the figure has a “smallest” rotation, then we will see that its symmetry group is either $C_n$ or $D_n$. 
To start, we will be specific about what type of figures we consider. A subset $S$ of $\mathbb{R}^2$ is said to be bounded if there is a positive number $r$ such that $S$ is a subset of the closed disk $\{P \in \mathbb{R}^2 : |P| \leq r\}$ of radius $r$. All examples in the previous section were bounded figures. We begin with a simple but important lemma restricting the type of group that can arise as the symmetry group of a bounded figure.

**Lemma 9.10.** Let $S$ be a bounded figure, and set $G = \text{Sym}(S)$. Then $G$ does not contain a nontrivial translation.

**Proof.** Suppose that $\tau \in G$ is a translation, and suppose that $\tau$ is translation by the vector $b$. Then, since $G$ is a group, $\tau^n \in G$ for every positive integer $n$. Thus, translation by $nb$ is a symmetry of $S$ for every $n$. Since $S$ is bounded, there is an $r$ such that $|P| \leq r$ for every $P \in S$. But then $\tau^n(P) = P + nb \in S$ for every $n$. The triangle inequality implies that $|nb| \leq |P| + |P + nb| \leq 2r$. This yields $n|b| \leq 2r$. However, since this is true for every $n$, we must have $|b| = 0$; thus, $b = 0$. Therefore, $\tau$ is the identity map. \(\square\)

To prove the main result about symmetry groups of bounded figures, we need some facts about compositions of isometries, whose proofs we leave for exercises.

1. The composition of two reflections about lines intersecting at a point $P$ is a rotation about $P$.
2. The composition of two rotations about a common point $P$ is another rotation about $P$.
3. If $r$ and $s$ are rotations about different centers, then $rsr^{-1}s^{-1}$ is a nontrivial translation.
4. If $r$ is a rotation about $P$ and $f$ is a reflection not fixing $P$, then $rf$ is a glide reflection.
5. Let $r$ be a rotation and $f$ a reflection in $O_2(\mathbb{R})$. Prove that $frf = r^{-1}$.

**Proposition 9.11.** Let $S$ be a bounded figure, and set $G = \text{Sym}(S)$. Then $G$ is a subgroup of $O_2(\mathbb{R})$. In particular, every isometry of $S$ is linear.

**Proof.** By the lemma there are no nontrivial translations in $G$. As a consequence, there are no nontrivial glide reflections, since the square of a nontrivial glide is a nontrivial translation. Thus, $G$ must consist solely of rotations and/or reflections. If $G$ is either the trivial group or is generated by a single reflection, then the result is clear. We may then assume that $G$ contains a nontrivial rotation or two reflections. In either case $G$ contains a nontrivial rotation since the composition of two reflections is a rotation. If $G$ contains two rotations $r$ and $s$, then $rs$ is a rotation if $r$ and $s$ share a common rotation center. However, if they have different centers, then $rsr^{-1}s^{-1}$ is a nontrivial translation. By the lemma, this cannot happen. Therefore, every rotation in $G$ has the same center, which we will view as the origin. If $r \in G$ and if $f \in G$ is a reflection, then $fr$ is a glide reflection if $f$ does preserve
the rotation center of \( r \). Since \( G \) does not contain a glide reflection, \( f \) must preserve the origin. Therefore, all symmetries of \( S \) preserve the origin, and so \( G \subseteq O_2(\mathbb{R}) \) is comprised of linear isometries.

Let \( S \) be a bounded figure. By the proposition, \( G = \text{Sym}(S) \) is a subgroup of \( O_2(\mathbb{R}) \). We can then consider the subgroup \( R \) of rotations of \( S \). We consider the case where \( R \) contains a nontrivial rotation of smallest possible angle.

**Corollary 9.12.** If \( S \) is a bounded figure such that \( \text{Sym}(S) \) contains a rotation of smallest possible angle, then \( \text{Sym}(G) \) is either isomorphic to \( C_n \) or to \( D_n \) for some \( n \).

**Proof.** Let \( r \) be a rotation in \( G = \text{Sym}(S) \) of smallest possible nonzero angle \( \theta \). Recall that \( \theta \) is only unique up to multiples of \( 2\pi \). We may assume that \( 0 < \theta < \pi \), since if \( G \) contains a rotation by \( \theta \), then it contains a rotation by \( -\theta \), and if \( \pi < \theta < 2\pi \), then \( 0 < 2\pi - \theta < \pi \), and \( r^{-1} \) is rotation by \( 2\pi - \theta \). If \( s \) is another rotation in \( G \), and if \( s \) is rotation by \( \phi \), then there is an integer \( m \) with \( m\theta \leq \phi < (m+1)\theta \). The symmetry \( sr^{-m} \) has rotation angle \( \phi - m\theta \), which by the inequality is smaller than \( \theta \). By choice of \( \theta \), we must have \( \phi - m\theta = 0 \), so \( \phi = m\theta \). Thus, \( s = r^m \). Therefore, the group \( R \) of rotations of \( S \) is the cyclic group generated by \( r \). Moreover, we claim that \( \theta = 2\pi/n \) for some \( n \). To prove this, let \( n \) be the smallest integer with \( 2\pi/\theta \leq n \). Then \( n - 1 < 2\pi/\theta \). Multiplying these inequalities by \( \theta \) and rewriting a little gives

\[
2\pi \leq n\theta < 2 + \theta,
\]

so \( 0 \leq n\theta - 2\pi < \theta \). However, the rotation angle in the interval \([0, 2\pi]\) for \( r^n \) is \( n\theta - 2\pi \). Since \( r^n \in G \), minimality of \( \theta \) forces \( n\theta - 2\pi = 0 \), giving \( \theta = 2\pi/n \), as desired. Since we have shown that \( R = \langle r \rangle \), and \( r^n = \text{id} \) but \( r^m \neq \text{id} \) for \( 0 < m < n \), we see that \( |R| = n \), and so \( |R| < \infty \). If \( G = R \), then \( G = C_n \). If not, then \( G \) contains a reflection \( f \). Therefore, \( [G : R] = 2 \) by Lemma ????. In any case, \( [G : R] < \infty \). Since \( |G| = [G : R] \cdot |R| \), the group \( G \) is finite. By Proposition ???, we see that \( G \) is \( C_n \) or \( D_n \) for some \( n \).

**Exercises**

1. Prove that the composition of two reflections about lines intersecting at a point \( P \) is a rotation about \( P \). Moreover, if \( \theta \) is the angle between the two reflection lines, prove that the composition is a rotation by \( \theta/2 \).

2. Prove that the composition of two rotations about a common point \( P \) is another rotation about \( P \). Moreover, if the rotations are by angles \( \theta \) and \( \phi \), prove that the composition is rotation by \( \theta + \phi \).

3. If \( r \) and \( s \) are rotations about different centers, prove that \( rsr^{-1}s^{-1} \) is a nontrivial translation.

4. If \( r \) is a rotation about \( P \) and \( f \) is a reflection not fixing \( P \), prove that \( rf \) is a glide reflection.
Bibliography
List of Symbols
Index

basis, 64
binary operation, 43
cancellation law of addition, 49
check digit, 1
code, 23
  BCH, 112
cyclic, 103
  perfect, 36
  Reed-Solomon, 116
codes
  error correcting, 23
column space, 33
complex numbers, 45
congruence modulo n, 6
congruent, 134
coset, 128
coset representative, 96
cyclic group, 126
$D(v, w)$, 24
degree of a polynomial, 89
dihedral group, 139
dimension, 67
distance function, 24
division algorithm, 7
division algorithm for polynomials, 90
elementary row operations, 28
error correcting codes, 23
Euclidean Algorithm, 14
Euler phi function, 124
Euler’s theorem, 128
field, 55
field extension, 77, 89
free variables, 34
gcd$(a, b)$, 11
general linear group, 124
glide reflection, 136
greatest common divisor, 11
greatest integer function, 26
group, 123
Hamming code, 35
Hamming matrix, 32
ideal, 93
ideal generated by an element, 94
identification number, 1
index, 128
International Standard Book Number, 1
irreducible polynomial, 98
ISBN code, 3
isometry, 134
kernel, 33
Lagrange’s theorem, 128
leading coefficient, 90
linear combination, 64
monic polynomial, 90
nullity, 35
nullspace, 33
order of a group, 127
permutation, 124
prime number, 12
quotient ring, 98
INDEX

rank, 33
reflection, 135
relatively prime, 12
ring, 44
    commutative, 45
rotation, 136
row reduced echelon form, 29
row space, 33
RSA encryption system, 121
signature, 130
spanning set, 64
subgroup, 125
subspace, 61
symmetry group, 135
translation, 135
transposition error, 4
unit of a ring, 123
UPC code, 2
USPS Code, 1
vector space, 59
weight, 24
well ordering property, 7
word, 23
wt(w), 24
zero divisor, 50
0, 23
\mathbb{Z}_n, 6