Before Calculus

Kenneth Kuttler

September 7, 2009
Contents

0.1 Introduction .................................................. 6

1 Numbers ......................................................... 9
  1.1 The Number Line And Field Axioms ....................... 9
  1.2 Exercises .................................................. 13
  1.3 Order ....................................................... 14
  1.4 Set Notation ............................................... 15
  1.5 Exercises .................................................. 16
  1.6 Exercises With Answers ................................... 17
  1.7 The Absolute Value ......................................... 17
  1.8 Exercises .................................................. 20
  1.9 Well Ordering And Archimedian Property .................. 20
  1.10 Exercises .................................................. 23
  1.11 Completeness of $\mathbb{R}$ ............................... 26
    1.11.1 Roots* ............................................... 27
  1.12 Exercises .................................................. 28
  1.13 Counting .................................................. 29
    1.13.1 Combinations .......................................... 29
    1.13.2 The Binomial Theorem ................................ 31
  1.14 Counting And Basic Probability .......................... 32
  1.15 Exercises .................................................. 33

2 Functions ....................................................... 37
  2.1 Generalities ............................................... 37
  2.2 Real Functions ............................................. 38
  2.3 Cartesian Coordinates And Graphs ......................... 41
  2.4 Exercises .................................................. 46
  2.5 Quadratic Functions ....................................... 48
    2.5.1 Maximizing And Minimizing ............................ 48
    2.5.2 Solving Quadratic Equations .......................... 50
  2.6 Exercises .................................................. 52
  2.7 Asymptotes ............................................... 54
  2.8 Exercises .................................................. 56

3 Division ........................................................ 59
  3.1 Division And Numbers ...................................... 59
  3.2 Exercises .................................................. 62
  3.3 Rational Root Theorem ..................................... 63
  3.4 Division And Polynomials ................................... 64
  3.5 Factoring Polynomials ...................................... 67
  3.6 Exercises .................................................. 70
  3.7 Partial Fractions .......................................... 71
0.1 Introduction

This is a book on precalculus in the sense that mastery of the material in this book will make calculus much easier. However, there is a lot here which really isn’t essential for calculus but is very interesting for its own sake.

Most students I have encountered when teaching calculus for the past 30 years or so don’t know many of the things I knew when I took calculus. I believe this is why so many have trouble. Some of these things are: the binomial theorem, exponential functions, basic algebraic concepts like one to one or onto functions, trig. identities, and sequences and series. I have included these things in the book.

I have also given a discussion of the method of partial fractions, including proofs which show why the method works. People sometimes wonder about why the method works when they take calculus but it is not usually discussed in the calculus book. Since the questions are mainly algebraic in nature, having little to do with calculus concepts, a book like this one is a good place to give such an explanation. For those who are interested in learning some more, there is a short section on field extensions. I know this is a strange and unusual place to put this material but why not? It doesn’t require prerequisite knowledge to understand and wouldn’t it be great for those who went through it? After all, there are other courses besides calculus which are important and some introduction to these ideas would be helpful.

I have given a discussion of compound interest and amortization as an application of sequences and series. This is material which is useful in life but which is not typically discussed in calculus. This topic is a great example of the usefulness of geometric series and sequences in addition to being important for its own sake.
0.1. INTRODUCTION

This is the right place to discuss limits of sequences. This topic is inherently easier than the limit of a function of a continuous variable because you don’t have to worry about whether the target point is a limit point of the domain of the function. Some exposure to sequences and series and their limits will go a long way toward helping students get through the harder concepts encountered in calculus.

The last part of the book is an introduction to linear algebra. I believe that some linear algebra understood before calculus will greatly help students to understand those concepts which occur, especially in multivariable calculus, which are essentially built on linear algebra concepts. There is really no excuse for not explaining what the derivative is in multivariable calculus but this omission is typical of calculus books because it is assumed that the students do not know about linear transformations or matrices. We put basic linear algebra in the wrong place. It should come before, not after calculus. (I am not speaking of canonical forms, but the simple ideas of matrix algebra and some introduction to determinants and row operations.)

I assume the reader has had some exposure to algebra in High School, although the book is essentially self contained.
Numbers

An understanding of the properties of the real numbers is essential in order to understand calculus. This section contains a review of the algebraic properties of real numbers.

1.1 The Number Line And Field Axioms

To begin with, consider the real numbers, denoted by \( \mathbb{R} \), as a line extending infinitely far in both directions. In this book, the notation, \( \equiv \) indicates something is being defined. Thus the integers are defined as

\[
\mathbb{Z} \equiv \{\cdots -1, 0, 1, \cdots \},
\]

the natural numbers,

\[
\mathbb{N} \equiv \{1, 2, \cdots \}
\]

and the rational numbers, defined as the numbers which are the quotient of two integers.

\[
\mathbb{Q} \equiv \left\{ \frac{m}{n} \text{ such that } m, n \in \mathbb{Z}, n \neq 0 \right\}
\]

are each subsets of \( \mathbb{R} \) as indicated in the following picture. (When you have a quotient of two numbers or more generally the quotient of any two things, the thing on the top is called the numerator and the thing on the bottom is called the denominator. Thus in \( \frac{a}{b} \) also written \( a/b \), \( a \) is called the numerator and \( b \) is called the denominator. I will also often refer to these as “top” and “bottom”.)

As shown in the picture, \( \frac{1}{2} \) is half way between the number 0 and the number, 1. By analogy, you can see where to place all the other rational numbers. It is assumed that \( \mathbb{R} \) has the following algebra properties, listed here as a collection of assertions called axioms. These properties will not be proved which is why they are called axioms rather than theorems. In general, axioms are statements which are regarded as true. Often these are things which are “self evident” either from experience or from some sort of intuition but this does not have to be the case.

**Axiom 1.1.1** \( x + y = y + x \), (commutative law for addition)
Axiom 1.1.2  \[ x + 0 = x, \text{ (additive identity)}. \]

Axiom 1.1.3  For each \( x \in \mathbb{R} \), there exists \(-x \in \mathbb{R}\) such that \( x + (-x) = 0 \), (existence of additive inverse).

Axiom 1.1.4  \((x + y) + z = x + (y + z)\), (associative law for addition).

Axiom 1.1.5  \(xy = yx\), (commutative law for multiplication).

Axiom 1.1.6  \((xy)z = x(yz)\), (associative law for multiplication).

Axiom 1.1.7  \(1x = x\), (multiplicative identity).

Axiom 1.1.8  For each \( x \neq 0 \), there exists \( x^{-1} \) such that \( xx^{-1} = 1\), (existence of multiplicative inverse).

Axiom 1.1.9  \(x(y + z) = xy + xz\), (distributive law).

These axioms are known as the field axioms and any set (there are many others besides \( \mathbb{R} \)) which has two such operations satisfying the above axioms is called a field. Division and subtraction are defined in the usual way by \( x - y \equiv x + (-y) \) and \( x/y \equiv x (y^{-1}) \). It is assumed that the reader is completely familiar with these axioms in the sense that he or she can do the usual algebraic manipulations taught in high school and junior high algebra courses. The axioms listed above are just a careful statement of exactly what is necessary to make the usual algebraic manipulations valid. A word of advice regarding division and subtraction is in order here. Whenever you feel a little confused about an algebraic expression which involves division or subtraction, think of division as multiplication by the multiplicative inverse as just indicated and think of subtraction as addition of the additive inverse. Thus, when you see \( x/y \), think \( x (y^{-1}) \) and when you see \( x - y \), think \( x + (-y) \). In many cases the source of confusion will disappear almost magically. The reason for this is that subtraction and division do not satisfy the associative law. This means there is a natural ambiguity in an expression like \( 6 - 3 - 4 \). Do you mean \((6 - 3) - 4 = -1\) or \(6 - (3 - 4) = 6 - (-1) = 7\)? It makes a difference doesn’t it? However, the so called binary operations of addition and multiplication are associative and so no such confusion will occur. It is conventional to simply do the operations in order of appearance reading from left to right. Thus, if you see \( 6 - 3 - 4 \), you would normally interpret it as the first of the above alternatives.

In doing algebra, the following theorem is important and follows from the above axioms. The reasoning which demonstrates this assertion is called a proof. Proofs and definitions are very important in mathematics because they are the means by which “truth” is determined. In mathematics, something is “true” if it follows from axioms using a correct logical argument. Truth is not determined on the basis of experiment or opinions and it is this which makes mathematics useful as a language for describing certain kinds of reality in a precise manner. It is also the definitions and proofs which make the subject of mathematics intellectually worth while. Take these away and it becomes a gray wasteland filled with endless tedium and meaningless manipulations.

In the first part of the following theorem, the claim is made that the additive inverse and the multiplicative inverse are unique. This means that for a given number, only one number has the property that it is an additive inverse and that, given a nonzero number, only one number has the property that it is a multiplicative inverse. The significance of this is that if you are wondering if a given number is the additive inverse of a given number, all you have to do is to check and see if it acts like one.

\footnote{There are certainly real and important things which should not be described using mathematics because it has nothing to do with these things. For example, feelings and emotions have nothing to do with math.}
Theorem 1.1.10 The above axioms imply the following.

1. The multiplicative inverse and additive inverses are unique.
2. $0x = 0, \, -(-x) = x$,
3. $(-1)(-1) = 1, \, (-1)x = -x$
4. If $xy = 0$ then either $x = 0$ or $y = 0$.

Proof: Suppose then that $x$ is a real number and that $x + y = 0 = x + z$. It is necessary to verify $y = z$. From the above axioms, there exists an additive inverse, $-x$ for $x$. Therefore,

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

and so by the associative law for addition,

$$((-x) + x) + y = ((-x) + x) + z$$

which implies

$$0 + y = 0 + z.$$  

Now by the definition of the additive identity, this implies $y = z$. You should prove the multiplicative inverse is unique.

Consider 2. It is desired to verify $0x = 0$. From the definition of the additive identity and the distributive law it follows that

$$0x = (0 + 0)x = 0x + 0x.$$  

From the existence of the additive inverse and the associative law it follows

$$0 = (-0x) + 0x = (-0x) + (0x + 0x)$$

$$= ((-0x) + 0x) + 0x = 0 + 0x = 0x.$$

To verify the second claim in 2., it suffices to show $x$ acts like the additive inverse of $-x$ in order to conclude that $-(-x) = x$. This is because it has just been shown that additive inverses are unique. By the definition of additive inverse,

$$x + (-x) = 0$$

and so $x = -(-x)$ as claimed.

To demonstrate 3.,

$$( -1)(1 + (-1)) = (-1)0 = 0$$

and so using the definition of the multiplicative identity, and the distributive law,

$$( -1) + (-1)(-1) = 0.$$  

It follows from 1. and 2. that $1 = -(-1) = (-1)(-1)$. To verify $(-1)x = -x$, use 2. and the distributive law to write

$$x + (-1)x = x(1 + (-1)) = x0 = 0.$$  

Therefore, by the uniqueness of the additive inverse proved in 1., it follows $(-1)x = -x$ as claimed.

To verify 4., suppose $x \neq 0$. Then $x^{-1}$ exists by the axiom about the existence of multiplicative inverses. Therefore, by 2. and the associative law for multiplication,

$$y = (x^{-1}x)y = x^{-1}(xy) = x^{-1}0 = 0.$$  

This proves 4. and completes the proof of this theorem.

Recall the notion of something raised to an integer power. Thus \( y^2 = y \times y \) and \( b^{-3} = \frac{1}{b^3} \) etc.

Also, there are a few conventions related to the order in which operations are performed. Exponents are always done before multiplication. Thus \( xy^2 = x(y^2) \) and is not equal to \((xy)^2\). Division or multiplication is always done before addition or subtraction. Thus \( x - y (z + w) = x - [y(z + w)] \) and is not equal to \((x - y)(z + w)\). Parentheses are done before anything else. Be very careful of such things since they are a source of mistakes. When you have doubts, insert parentheses to resolve the ambiguities.

Also recall summation notation. If you have not seen this, the following is a short review of this topic.

**Definition 1.1.11** Let \( x_1, x_2, \ldots, x_m \) be numbers. Then

\[
\sum_{j=1}^{m} x_j \equiv x_1 + x_2 + \cdots + x_m.
\]

Thus this symbol, \( \sum_{j=1}^{m} x_j \) means to take all the numbers, \( x_1, x_2, \ldots, x_m \) and add them all up. Note the use of the \( j \) as a generic variable which takes values from 1 up to \( m \). This notation will be used whenever there are things which can be added, not just numbers.

As an example of the use of this notation, you should verify the following.

**Example 1.1.12** \( \sum_{k=1}^{6} (2k + 1) = 48 \).

Be sure you understand why

\[
\sum_{k=1}^{m+1} x_k = \sum_{k=1}^{m} x_k + x_{m+1}.
\]

As a slight generalization of this notation,

\[
\sum_{j=1}^{m} x_j \equiv x_1 + \cdots + x_m.
\]

It is also possible to change the variable of summation.

\[
\sum_{j=1}^{m} x_j = x_1 + x_2 + \cdots + x_m
\]

while if \( r \) is an integer, the notation requires

\[
\sum_{j=1+r}^{m+r} x_{j-r} = x_1 + x_2 + \cdots + x_m
\]

and so \( \sum_{j=1}^{m} x_j = \sum_{j=1+r}^{m+r} x_{j-r} \).

Summation notation will be used throughout the book whenever it is convenient to do so.

Another thing to keep in mind is that you often use letters to represent numbers. Since they represent numbers, you manipulate expressions involving letters in the same manner as you would if they were specific numbers.
Example 1.1.13 Add the fractions, \( \frac{x}{x^2+y} + \frac{y}{x-1} \).

You add these just like they were numbers. Write the first expression as \( \frac{x(x-1)}{(x^2+y)(x-1)} \) and the second as \( \frac{y(x^2+y)}{(x-1)(x^2+y)} \). Then since these have the same common denominator, you add them as follows.

\[
\frac{x}{x^2+y} + \frac{y}{x-1} = \frac{x(x-1)}{(x^2+y)(x-1)} + \frac{y(x^2+y)}{(x-1)(x^2+y)} = \frac{x^2 - x + yx^2 + y^2}{(x^2+y)(x-1)}.
\]

1.2 Exercises

1. Consider the expression \( x + y (x + y) - x (y - x) \equiv f (x, y) \). Find \( f (-1, 2) \).

2. Show \(- (ab) = (-a) b \).

3. Show on the number line the effect of adding two positive numbers, \( x \) and \( y \).

4. Show on the number line the effect of subtracting a positive number from another positive number.

5. Show on the number line the effect of multiplying a number by \(-1\).

6. Add the fractions \( \frac{x}{x^2+y} + \frac{x-1}{x+1} \).

7. Find a formula for \( (x + y)^2 \), \( (x + y)^3 \), and \( (x + y)^4 \). Based on what you observe for these, give a formula for \( (x + y)^8 \).

8. When is it true that \( (x + y)^n = x^n + y^n \)?

9. Find the error in the following argument. Let \( x = y = 1 \). Then \( xy = y^2 \) and so \( xy - x^2 = y^2 - x^2 \). Therefore, \( x (y - x) = (y - x) (y + x) \). Dividing both sides by \( y - x \) yields \( x = x + y \). Now substituting in what these variables equal yields \( 1 = 1 + 1 \).

10. Find the error in the following argument. \( \sqrt{x^2+1} = x + 1 \) and so letting \( x = 2 \), \( \sqrt{5} = 3 \). Therefore, \( 5 = 9 \).

11. Find the error in the following. Let \( x = 1 \) and \( y = 2 \). Then \( \frac{1}{x} = \frac{1}{x+y} = \frac{1}{x+y} = 1 + \frac{1}{2} = \frac{3}{2} \). Then cross multiplying, yields \( 2 = 9 \).

12. Simplify \( \frac{x^2 y^2 + n}{x+y} \).

13. Simplify the following expressions using correct algebra. In these expressions the variables represent real numbers.

\[
\begin{align*}
(a) \quad & \frac{x^2+y^2+x}{x} \\
(b) \quad & \frac{x^2+y^2+x}{xy} \\
(c) \quad & \frac{x^3+2x^2-x-2}{x+1}
\end{align*}
\]

14. Find the error in the following argument. Let \( x = 3 \) and \( y = 1 \). Then \( 1 = 3 - 2 = 3 - (3 - 1) = x - y (x - y) = (x - y) (x - y) = 2^2 = 4 \).
15. Verify the following formulas.

(a) \((x - y)(x + y) = x^2 - y^2\)

(b) \((x - y)(x^2 + xy + y^2) = x^3 - y^3\)

(c) \((x + y)(x^2 - xy + y^2) = x^3 + y^3\)

16. Find the error in the following.

\[
\frac{xy + y}{x} = y + y = 2y.
\]

Now let \(x = 2\) and \(y = 2\) to obtain

\[
3 = 4
\]

17. Show the rational numbers satisfy the field axioms. You may assume the associative, commutative, and distributive laws hold for the integers.

### 1.3 Order

The real numbers also have an order defined on them. This order can be defined very precisely in terms of a short list of axioms but this will not be done here. Instead, properties which should be familiar are listed here as axioms.

**Definition 1.3.1** The expression, \(x < y\), in words, \((x\) is less than \(y\)) means \(y\) lies to the right of \(x\) on the number line.

\[
\begin{array}{cc}
  x & y \\
\end{array}
\]

The expression \(x > y\), in words \((x\) is greater than \(y\)) means \(x\) is to the right of \(y\) on the number line.

\[
\begin{array}{cc}
  y & x \\
\end{array}
\]

\(x \leq y\) if either \(x = y\) or \(x < y\). \(x \geq y\) if either \(x > y\) or \(x = y\). A number, \(x\), is positive if \(x > 0\).

If you examine the number line, the following should be fairly reasonable and are listed as axioms, things assumed to be true. I suggest you plug in some numbers to reassure yourself about these axioms.

**Axiom 1.3.2** The sum of two positive real numbers is positive.

**Axiom 1.3.3** The product of two positive real numbers is positive.

**Axiom 1.3.4** For a given real number \(x\), one and only one of the following alternatives holds. Either \(x\) is positive, \(x = 0\), or \(-x\) is positive.

**Axiom 1.3.5** If \(x < y\) and \(y < z\) then \(x < z\) (Transitive law).

**Axiom 1.3.6** If \(x < y\) then \(x + z < y + z\) (addition to an inequality).

**Axiom 1.3.7** If \(x \leq 0\) and \(y \leq 0\), then \(xy \geq 0\).

**Axiom 1.3.8** If \(x > 0\) then \(x^{-1} > 0\).

**Axiom 1.3.9** If \(x < 0\) then \(x^{-1} < 0\).
Axiom 1.3.10 If $x < y$ then $xz < yz$ if $z > 0$, (multiplication of an inequality by a positive number).

Axiom 1.3.11 If $x < y$ and $z < 0$, then $xz > yz$ (multiplication of an inequality by a negative number).

Axiom 1.3.12 Each of the above holds with $>$ and $<$ replaced by $\geq$ and $\leq$ respectively except for 1.3.8 and 1.3.9 in which it is also necessary to stipulate that $x \neq 0$.

Axiom 1.3.13 For any $x$ and $y$, exactly one of the following must hold. Either $x = y$, $x < y$, or $x > y$ (trichotomy).

Note that trichotomy could be stated by saying $x \leq y$ or $y \leq x$.

Example 1.3.14 Solve the inequality $2x + 4 \leq x - 8$

Subtract $2x$ from both sides to yield $4 \leq -x - 8$. Next add 8 to both sides to get $12 \leq -x$. Then multiply both sides by $(-1)$ to obtain $x \leq -12$.Alternatively, subtract $x$ from both sides to get $x + 4 \leq -8$. Then subtract 4 from both sides to obtain $x \leq -12$.

Example 1.3.15 Solve the inequality $(x + 1)(2x - 3) \geq 0$.

If this is to hold, either both of the factors, $x + 1$ and $2x - 3$ are nonnegative or they are both non-positive. The first case yields $x + 1 \geq 0$ and $2x - 3 \geq 0$ so $x \geq -1$ and $x \geq \frac{3}{2}$ yielding $x \geq \frac{3}{2}$. The second case yields $x + 1 \leq 0$ and $2x - 3 \leq 0$ which implies $x \leq -1$ and $x \leq \frac{3}{2}$. Therefore, the solution to this inequality is $x \leq -1$ or $x \geq \frac{3}{2}$.

Example 1.3.16 Solve the inequality $x(x + 2) \geq -4$

Here the problem is to find $x$ such that $x^2 + 2x + 4 \geq 0$. However, $x^2 + 2x + 4 = (x + 1)^2 + 3 \geq 0$ for all $x$. Therefore, the solution to this problem is all $x$ for any $x$ a real number. As explained below, this can be written $x \in \mathbb{R}$.

1.4 Set Notation

A set is just a collection of things called elements. Often these are also referred to as points in calculus. For example $\{1, 2, 3, 8\}$ would be a set consisting of the elements 1, 2, 3, and 8. To indicate that 3 is an element of $\{1, 2, 3, 8\}$, it is customary to write $3 \in \{1, 2, 3, 8\}$. $9 \notin \{1, 2, 3, 8\}$ means 9 is not an element of $\{1, 2, 3, 8\}$. Sometimes a rule specifies a set. For example you could specify a set as all integers larger than 2. This would be written as $S = \{x \in \mathbb{Z} : x > 2\}$. This notation says: the set of all integers, $x$, such that $x > 2$.

If $A$ and $B$ are sets with the property that every element of $A$ is an element of $B$, then $A$ is a subset of $B$. For example, $\{1, 2, 3, 8\}$ is a subset of $\{1, 2, 3, 4, 5, 8\}$, in symbols, $\{1, 2, 3, 8\} \subseteq \{1, 2, 3, 4, 5, 8\}$. The same statement about the two sets may also be written as $\{1, 2, 3, 4, 5, 8\} \supseteq \{1, 2, 3, 8\}$.

The union of two sets is the set consisting of everything which is contained in at least one of the sets, $A$ or $B$. As an example of the union of two sets, $\{1, 2, 3, 8\} \cup \{3, 4, 7, 8\} = \{1, 2, 3, 4, 7, 8\}$ because these numbers are those which are in at least one of the two sets. In general

$$A \cup B \equiv \{x : x \in A \text{ or } x \in B\}.$$  

Be sure you understand that something which is in both $A$ and $B$ is in the union. It is not an exclusive or.
The intersection of two sets, \( A \) and \( B \) consists of everything which is in both of the sets. Thus \( \{1, 2, 3, 8\} \cap \{3, 4, 7, 8\} = \{3, 8\} \) because 3 and 8 are those elements the two sets have in common. In general,

\[
A \cap B \equiv \{x : x \in A \text{ and } x \in B\}.
\]

When with real numbers, \([a, b]\) denotes the set of real numbers, \(x\), such that \(a \leq x \leq b\) and \((a, b]\) denotes the set of real numbers such that \(a < x \leq b\). \((a, b)\) consists of the set of real numbers, \(x\) such that \(a < x < b\) and \([-\infty, a]\) means the set of all real numbers which are less than or equal to \(a\). These sorts of sets of real numbers are called intervals. The two points, \(a\) and \(b\) are called endpoints of the interval. Other intervals such as \((-\infty, b]\) are defined by analogy to what was just explained. In general, the curved parenthesis indicates the end point it sits next to is not included while the square parenthesis indicates this end point is included. The reason that there will always be a curved parenthesis next to \(-\infty\) or \(-\infty\) is that these are not real numbers. Therefore, they cannot be included in any set of real numbers.

A special set which needs to be given a name is the empty set also called the null set, denoted by \(\emptyset\). Thus \(\emptyset\) is defined as the set which has no elements in it. Mathematicians like to say the empty set is a subset of every set. The reason they say this is that if it were not so, there would have to exist a set, \(A\), such that \(\emptyset\) has something in it which is not in \(A\). However, \(\emptyset\) has nothing in it and so the least intellectual discomfort is achieved by saying \(\emptyset \subseteq A\).

If \(A\) and \(B\) are two sets, \(A \setminus B\) denotes the set of things which are in \(A\) but not in \(B\). Thus

\[
A \setminus B \equiv \{x \in A : x \notin B\}.
\]

Set notation is used whenever convenient.

To illustrate the use of this notation consider the same three examples of inequalities.

**Example 1.4.1** Solve the inequality \(2x + 4 \leq x - 8\)

This was worked earlier and \(x \leq -12\) was the answer. This is written as \((-\infty, -12]\).

**Example 1.4.2** Solve the inequality \((x + 1) (2x - 3) \geq 0\).

This was worked earlier and \(x \leq -1\) or \(x \geq \frac{3}{2}\) was the answer. In terms of set notation this is denoted by \((-\infty, -1] \cup [\frac{3}{2}, \infty)\).

**Example 1.4.3** Solve the inequality \(x (x + 2) \geq -4\)

Recall this inequality was true for any value of \(x\). It is written as \(\mathbb{R} \) or \((-\infty, \infty)\).

### 1.5 Exercises

1. Solve \((3x + 2) (x - 3) \leq 0\).
2. Solve \((3x + 2) (x - 3) > 0\).
3. Solve \(\frac{x + 2}{3x - 2} < 0\).
4. Solve \(\frac{x + 1}{x + 3} < 1\).
5. Solve \((x - 1) (2x + 1) \leq 2\).
6. Solve \((x - 1) (2x + 1) > 2\).
1.6. Exercises With Answers

7. Solve $x^2 - 2x \leq 0$.

8. Solve $(x + 2)(x - 2)^2 \leq 0$.

9. Solve $\frac{3x - 4}{x + 2} \geq 0$. Hint: $x^2 + 2x + 2 = (x + 1)^2 + 1$.

10. Solve $\frac{3x + 9}{x^2 + 2x + 9} \geq 1$.

11. Solve $\frac{x^2 + 2x + 1}{3x + 1} < 1$. Hint: $x^2 - x - 6 = (x - 3)(x + 2)$.

1.6 Exercises With Answers

1. Solve $(3x + 1)(x - 2) \leq 0$.

This happens when the two factors have different signs. Thus either $3x + 1 \leq 0$ and $x - 2 \geq 0$ in which case $x \leq -\frac{1}{3}$ and $x \geq 2$, a situation which never occurs, or else $3x + 1 \geq 0$ and $x - 2 \leq 0$ so $x \geq \frac{1}{3}$ and $x \leq 2$. Written as $[-\frac{1}{3}, 2]$.

2. Solve $(3x + 1)(x - 2) > 0$.

This is just everything not included in the above problem. Thus the answer would be $(-\infty, -\frac{1}{3}) \cup (2, \infty)$.

3. Solve $\frac{x + 1}{x - 2} < 0$.

Note that $\frac{x + 1}{x - 2}$ is positive if $x > 1$, negative if $x \in (-1, 1)$, and nonnegative if $x \leq -1$. Therefore, the answer is $(-1, 1)$.

To identify the interesting intervals, all that was necessary to do was to look at the two factors, $(x + 1)$ and $(2x - 2)$ and determine where these equal zero.

4. Solve $\frac{3x + 7}{x^2 + 2x + 1} \geq 1$.

On something like this, subtract 1 from both sides to get

$$\frac{6 + x - x^2}{x^2 + 2x + 1} = \frac{(3 - x)(2 + x)}{(x + 1)^2}.$$  

When $x = 3$ or $x = -2$, this equals zero. For $x \in (-2, 3)$ the expression is positive and it is negative if $x > 3$ or if $x < -2$. Therefore, the answer is $[-2, 3]$.

1.7 The Absolute Value

A fundamental idea is the absolute value of a number. This is important because the absolute value defines distance on $\mathbb{R}$. How far away from 0 is the number 3? How about the number $-3$? Look at the number line and observe they are both 3 units away from 0. To describe this algebraically,

**Definition 1.7.1** $|x| \equiv \begin{cases} x \text{ if } x \geq 0, \\ -x \text{ if } x < 0. \end{cases}$

Thus $|x|$ can be thought of as the distance between $x$ and 0. It may be useful to think of this function in terms of its graph if you recall the notion of the graph of a function. This concept is discussed more later but you may have seen it already.

The following is a fundamental theorem about the absolute value.
Theorem 1.7.2 \(|xy| = |x| |y|\).

**Proof:** If both \(x, y \leq 0\), then \(|xy| = xy\) because in this case \(xy \geq 0\) while
\[|x| |y| = (-x) (-y) = (-1) x (-1) y = (-1) (-1) xy = xy.\]
Therefore, in this case the result of the theorem is verified. You should verify the other cases, both \(x, y \geq 0\) and \(x \leq 0\) while \(y \geq 0\).
This theorem is the basis for the following fundamental result which is of major importance in calculus.

Theorem 1.7.3 The following inequalities hold.
\[|x + y| \leq |x| + |y|, \quad ||x| - |y|| \leq |x - y|.\]
Either of these inequalities may be called the triangle inequality.

**Proof:** By Theorem 1.7.2,
\[|x + y|^2 = |(x + y)^2| = (x + y)^2 = x^2 + y^2 + 2xy \leq x^2 + y^2 + 2|x| |y| = |x|^2 + |y|^2 + 2|x| |y| = (|x| + |y|)^2.\]
Now note that if \(0 \leq a \leq b\) then \(0 \leq a^2 \leq ab \leq b^2\) and that if \(a, b \geq 0\) then if \(a^2 \leq b^2\) it follows that \(b^2 \geq ba \geq a^2\) and so \(b \geq a\) (see the above axioms. Multiply by \(a^{-1}\) if \(a \neq 0\).) Applying this observation to the above inequality,
\[|x + y| \leq |x| + |y|.\]
This verifies the first of these inequalities. To obtain the second one, note
\[|x| = |x - y + y| \leq |x - y| + |y|\]
and so
\[|x| - |y| \leq |x - y| \tag{1.1}\]
Now switch the letters to obtain
\[|y| - |x| \leq |y - x| = |x - y| \tag{1.2}\]
Therefore,
\[||x| - |y|| \leq |x - y|\]
because if \(|x| - |y| \geq 0\), then the conclusion follows from (1.1) while if \(|x| - |y| \leq 0\), the conclusion follows from (1.2). This proves the theorem.

Note there is an inequality involved. Consider the following.
\[|3 + (-2)| = |1| = 1\]
while
\[|3| + |(-2)| = 3 + 2 = 5.\]
You observe that \(5 > 1\) and so it is important to remember that the triangle inequality is an inequality.

**Example 1.7.4** Solve the equation \(|x - 1| = 2\)
1.7. THE ABSOLUTE VALUE

This will be true when \( x - 1 = 2 \) or when \( x - 1 = -2 \). Therefore, there are two solutions to this problem, \( x = 3 \) or \( x = -1 \).

**Example 1.7.5** Solve the inequality \(|2x - 1| < 2\)

From the number line, it is necessary to have \( 2x - 1 \) between \(-2 \) and \( 2 \) because the inequality says that the distance from \( 2x - 1 \) to 0 is less than 2. Therefore, \(-2 < 2x - 1 < 2\) and so \(-1/2 < x < 3/2\). In other words, \(-1/2 < x \) and \( x < 3/2\).

**Example 1.7.6** Solve the inequality \(|2x - 1| > 2\).

This happens if \( 2x - 1 > 2 \) or if \( 2x - 1 < -2 \). Thus the solution is \( x > 3/2 \) or \( x < -1/2 \). \((3/2, \infty) \cup (-\infty, -1/2)\).

**Example 1.7.7** Solve \(|x + 1| = |2x - 2|\)

There are two ways this can happen. It could be the case that \( x + 1 = 2x - 2 \) in which case \( x = 3 \) or alternatively, \( x + 1 = 2 - 2x \) in which case \( x = 1/3 \).

**Example 1.7.8** Obtain a number, \( \delta \), such that if \( |x - 2| < \delta \), then \(|x^2 - 4| < 1/10\).

If \(|x - 2| < 1\), then \(||x| - |2|| < 1 \) and so \(|x| < 3\). Therefore, if \(|x - 2| < 1\),

\[ |x^2 - 4| = |x + 2| |x - 2| \leq (|x| + 2) |x - 2| \leq 5 |x - 2|. \]

Therefore, if \(|x - 2| < \frac{1}{50}\), the desired inequality will hold. Note that some of this is arbitrary. For example, if \(|x - 2| < 3\), then \(||x| - |2|| < 3 \) and so \(|x| < 5\). Therefore, for such \(x\),

\[ |x^2 - 4| = |x + 2| |x - 2| \leq (|x| + 2) |x - 2| \leq 7 |x - 2| \]

and so it would also suffice to take \(|x - 2| < \frac{1}{70}\). The example is about the existence of a number which has a certain property, not the question of finding a particular such number. There are infinitely many which will work because if you have found one, then any which is smaller will also work.

**Example 1.7.9** Suppose \( \varepsilon > 0 \) is a given positive number. Obtain a number, \( \delta > 0 \), such that if \(|x - 1| < \delta\), then \(|x^2 - 1| < \varepsilon\).

First of all, note \(|x^2 - 1| = |x - 1||x + 1| \leq (|x| + 1)|x - 1|\). Now if \(|x - 1| < 1\), it follows \(|x| < 2\) and so for \(|x - 1| < 1\),

\[ |x^2 - 1| < 3 |x - 1|. \]

Now let \( \delta = \min \{1, \frac{\varepsilon}{3} \} \). This notation means to take the minimum of the two numbers, 1 and \( \frac{\varepsilon}{3} \). Then if \(|x - 1| < \delta\),

\[ |x^2 - 1| < 3 |x - 1| < 3 \frac{\varepsilon}{3} = \varepsilon. \]
1.8 Exercises

1. Solve $|x + 1| = |2x - 3|$.
2. Solve $[3x + 1] < 8$. Give your answer in terms of intervals on the real line.
3. Sketch on the number line the solution to the inequality $|x - 3| > 2$.
4. Sketch on the number line the solution to the inequality $|x - 3| < 2$.
5. Show $|x| = \sqrt{x^2}$.
6. Solve $|x + 2| < 3x - 3$.
7. Tell when equality holds in the triangle inequality.
8. Solve $x + 2 \leq 8 + |2x - 4|$.
9. Solve $\frac{|x+1|}{|2x-5|} = 1$.
10. Solve $\frac{|x+1|}{|2x-5|} \leq 1$. **Hint:** This one might be a little long. Look at various cases.
11. Verify the axioms for order listed above are reasonable by consideration of the number line. In particular, show that if $x \leq z$ and $y < 0$ then $xy \geq yz$.
12. Solve $(x + 1) (2x - 2) x \geq 0$.
13. Solve $\frac{x+3}{2x+1} > 1$.
14. Solve $\frac{x+2}{4x+1} > 2$.
15. Describe the set of numbers, $a$ such that there is no solution to $|x + 1| = 4 - |x + a|$.
16. Suppose $0 < a < b$. Show $a^{-1} > b^{-1}$.
17. Show that if $|x - 6| < 1$, then $|x| < 7$.
18. Suppose $|x - 8| < 2$. How large can $|x - 5|$ be?
19. Obtain a number, $\delta > 0$, such that if $|x - 1| < \delta$, then $|x^2 - 1| < 1/10$.

1.9 Well Ordering And Archimedian Property

**Definition 1.9.1** A set is well ordered if every nonempty subset $S$, contains a smallest element $z$ having the property that $z \leq x$ for all $x \in S$.

**Axiom 1.9.2** Any set of integers larger than a given number is well ordered.

In particular, the natural numbers defined as 

$$\mathbb{N} \equiv \{1, 2, \cdots\}$$

is well ordered.

The above axiom implies the principle of mathematical induction.

**Theorem 1.9.3** (Mathematical induction) A set $S \subseteq \mathbb{Z}$, having the property that $a \in S$ and $n + 1 \in S$ whenever $n \in S$ contains all integers $x \in \mathbb{Z}$ such that $x \geq a$. 

Proof: Let $T \equiv ([a, \infty) \cap \mathbb{Z}) \setminus S$. Thus $T$ consists of all integers larger than or equal to $a$ which are not in $S$. The theorem will be proved if $T = \emptyset$. If $T \neq \emptyset$ then by the well ordering principle, there would have to exist a smallest element of $T$, denoted as $b$. It must be the case that $b > a$ since by definition, $a \notin T$. Then the integer, $b - 1 \geq a$ and $b - 1 \notin S$ because if $b - 1 \in S$, then $b - 1 + 1 = b \in S$ by the assumed property of $S$. Therefore, $b - 1 \in ([a, \infty) \cap \mathbb{Z}) \setminus S = T$ which contradicts the choice of $b$ as the smallest element of $T$. ($b - 1$ is smaller.) Since a contradiction is obtained by assuming $T \neq \emptyset$, it must be the case that $T = \emptyset$ and this says that everything in $[a, \infty) \cap \mathbb{Z}$ is also in $S$.

Mathematical induction is a very useful device for proving theorems about the integers.

Example 1.9.4 Prove by induction that $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$.

By inspection, if $n = 1$ then the formula is true. The sum yields 1 and so does the formula on the right. Suppose this formula is valid for some $n \geq 1$ where $n$ is an integer. Then

$$\sum_{k=1}^{n+1} k^2 = \sum_{k=1}^{n} k^2 + (n + 1)^2 = \frac{n(n+1)(2n+1)}{6} + (n + 1)^2.$$ 

The step going from the first to the second line is based on the assumption that the formula is true for $n$. This is called the induction hypothesis. Now simplify the expression in the second line,

$$\frac{n(n+1)(2n+1)}{6} + (n + 1)^2.$$ 

This equals

$$(n + 1) \left( \frac{n(2n+1)}{6} + (n + 1) \right)$$ 

and

$$\frac{n(2n+1)}{6} + (n + 1) = \frac{6(n+1)+2n^2+n}{6} = \frac{(n+2)(2n+3)}{6}.$$ 

Therefore,

$$\sum_{k=1}^{n+1} k^2 = \frac{(n+1)(n+2)(2n+3)}{6} = \frac{(n+1)((n+1)+1)(2(n+1)+1)}{6},$$

showing the formula holds for $n + 1$ whenever it holds for $n$. This proves the formula by mathematical induction.

Example 1.9.5 Show that for all $n \in \mathbb{N}$, $\frac{1}{2} \cdot \frac{3}{4} \cdots \frac{2n-1}{2n} < \frac{1}{\sqrt{2n+1}}$.

If $n = 1$ this reduces to the statement that $\frac{1}{2} < \frac{1}{\sqrt{3}}$ which is obviously true. Suppose then that the inequality holds for $n$. Then

$$\frac{1}{2} \cdot \frac{3}{4} \cdots \frac{2n-1}{2n} < \frac{1}{\sqrt{2n+1}} \cdot \frac{2n+1}{2n+2} = \frac{\sqrt{2n+1}}{2n+2}.$$
The theorem will be proved if this last expression is less than \( \frac{1}{\sqrt{2n+3}} \). This happens if and only if
\[
\left( \frac{1}{\sqrt{2n+3}} \right)^2 = \frac{1}{2n+3} > \frac{2n+1}{(2n+2)^2}
\]
which occurs if and only if \((2n+2)^2 > (2n+3)(2n+1)\) and this is clearly true which may be seen from expanding both sides. This proves the inequality.

Let's review the process just used. If \( S \) is the set of integers at least as large as 1 for which the formula holds, the first step was to show 1 \( \in S \) and then that whenever \( n \in S \), it follows \( n+1 \in S \). Therefore, by the principle of mathematical induction, \( S \) contains \( [1, \infty) \cap \mathbb{Z} \), all positive integers. In doing an inductive proof of this sort, the set, \( S \) is normally not mentioned. One just verifies the steps above. First show the thing is true for some \( a \in \mathbb{Z} \) and then verify that whenever it is true for \( m \) it follows it is also true for \( m+1 \). When this has been done, the theorem has been proved for all \( m \geq a \).

**Definition 1.9.6** The Archimedean property states that whenever \( x \in \mathbb{R} \), and \( a > 0 \), there exists \( n \in \mathbb{N} \) such that \( na > x \).

**Axiom 1.9.7** \( \mathbb{R} \) has the Archimedean property.

This is not hard to believe. Just look at the number line. This Archimedean property is quite important because it shows every real number is smaller than some integer. It also can be used to verify a very important property of the rational numbers.

**Theorem 1.9.8** Suppose \( x < y \) and \( y - x > 1 \). Then there exists an integer, \( l \in \mathbb{Z} \), such that \( x < l < y \). If \( x \) is an integer, there is no integer \( y \) satisfying \( x < y < x+1 \).

**Proof:** Let \( x \) be the smallest positive integer. Not surprisingly, \( x = 1 \) but this can be proved. If \( x < 1 \) then \( x^2 < x \) contradicting the assertion that \( x \) is the smallest natural number. Therefore, 1 is the smallest natural number. This shows there is no integer, \( y \), satisfying \( x < y < x+1 \) since otherwise, you could subtract \( x \) and conclude \( 0 < y - x < 1 \) for some integer \( y - x \).

Now suppose \( y - x > 1 \) and let
\[
S = \{ w \in \mathbb{N} : w \geq y \}.
\]
The set \( S \) is nonempty by the Archimedean property. Let \( k \) be the smallest element of \( S \). Therefore, \( k-1 < y \). Either \( k-1 \leq x \) or \( k-1 > x \). If \( k-1 \leq x \), then
\[
y - x \leq y - (k-1) = y - k + 1 \leq 1
\]
contrary to the assumption that \( y - x > 1 \). Therefore, \( x < k-1 < y \) and this proves the theorem with \( l = k-1 \).

It is the next theorem which gives the density of the rational numbers. This means that for any real number, there exists a rational number arbitrarily close to it.

**Theorem 1.9.9** If \( x < y \) then there exists a rational number \( r \) such that \( x < r < y \).

**Proof:** Let \( n \in \mathbb{N} \) be large enough that
\[
n(y - x) > 1.
\]
Thus \( (y - x) \) added to itself \( n \) times is larger than 1. Therefore,
\[
n(y - x) = ny + n(-x) = ny - nx > 1.
\]
It follows from Theorem 1.9.8 there exists \( m \in \mathbb{Z} \) such that
\[
x < m < ny
\]
and so take \( r = m/n. \)

**Definition 1.9.10** A set, \( S \subseteq \mathbb{R} \) is dense in \( \mathbb{R} \) if whenever \( a < b \), \( S \cap (a,b) \neq \emptyset \).

Thus the above theorem says \( \mathbb{Q} \) is “dense” in \( \mathbb{R} \).

You probably saw the process of division in elementary school. Even though you saw it at a young age it is very profound and quite difficult to understand. Suppose you want to do the following problem \( 79 \div 22 \). What did you do? You likely did a process of long division which gave the following result.

\[
\frac{79}{22} = 3 \text{ with remainder } 13.
\]

This meant
\[
79 = 3(22) + 13.
\]

You were given two numbers, 79 and 22 and you wrote the first as some multiple of the second added to a third number which was smaller than the second number. Can this always be done? The answer is in the next theorem and depends here on the Archimedian property of the real numbers.

**Theorem 1.9.11** Suppose \( 0 < a \) and let \( b \geq 0 \). Then there exists a unique integer \( p \) and real number \( r \) such that \( 0 \leq r < a \) and \( b = pa + r. \)

**Proof:** Let \( S \equiv \{ n \in \mathbb{N} : an > b \} \). By the Archimedian property this set is nonempty. Let \( p + 1 \) be the smallest element of \( S \). Then \( pa \leq b \) because \( p + 1 \) is the smallest in \( S \). Therefore,
\[
r = b - pa \geq 0.
\]

If \( r \geq a \) then \( b - pa \geq a \) and so \( b \geq (p + 1)a \) contradicting \( p + 1 \in S \). Therefore, \( r < a \) as desired.

To verify uniqueness of \( p \) and \( r \), suppose \( p_i \) and \( r_i \), \( i = 1, 2 \), both work and \( r_2 > r_1 \). Then a little algebra shows
\[
p_1 - p_2 = \frac{r_2 - r_1}{a} \in (0,1).
\]

Thus \( p_1 - p_2 \) is an integer between 0 and 1, contradicting Theorem 1.9.8. The case that \( r_1 > r_2 \) cannot occur either by similar reasoning. Thus \( r_1 = r_2 \) and it follows that \( p_1 = p_2 \).

This theorem is called the Euclidean algorithm when \( a \) and \( b \) are integers.

### 1.10 Exercises

1. The Archimedian property implies the rational numbers are dense in \( \mathbb{R} \). Now consider the numbers which are of the form \( \frac{k}{m} \) where \( k \in \mathbb{Z} \) and \( m \in \mathbb{N} \). Using the number line, demonstrate that the numbers of this form are also dense in \( \mathbb{R} \).

2. Show there is no smallest number in \((0,1)\). Recall \((0,1)\) means the real numbers which are strictly larger than 0 and smaller than 1.

3. Show there is no smallest number in \( \mathbb{Q} \cap (0,1) \).
4. Show that if \( S \subseteq \mathbb{R} \) and \( S \) is well ordered with respect to the usual order on \( \mathbb{R} \) then \( S \) cannot be dense in \( \mathbb{R} \).

5. Prove by induction that \( \sum_{k=1}^{n} k^3 = \frac{1}{4}n^4 + \frac{1}{2}n^3 + \frac{1}{4}n^2 \).

6. It is a fine thing to be able to prove a theorem by induction but it is even better to be able to come up with a theorem to prove in the first place. Derive a formula for \( \sum_{k=1}^{n} k^3 \) in the following way. Look for a formula in the form \( An^5 + Bn^4 + Cn^3 + Dn^2 + En + F \). Then try to find the constants \( A, B, C, D, E, \) and \( F \) such that things work out right. In doing this, show

\[
(n+1)^4 = \left( A(n+1)^5 + B(n+1)^4 + C(n+1)^3 + D(n+1)^2 + E(n+1) + F \right) - An^5 + Bn^4 + Cn^3 + Dn^2 + En + F
\]

and so some progress can be made by matching the coefficients. When you get your answer, prove it is valid by induction.

7. Prove by induction that whenever \( n \geq 2 \), \( \sum_{k=1}^{n} \frac{1}{\sqrt{k}} > \sqrt{n} \).

8. If \( r \neq 0 \), show by induction that \( \sum_{k=1}^{n} ar^k = a \frac{r^{n+1} - 1}{r-1} - a \frac{r}{r-1} \).

9. Prove by induction that \( \sum_{k=1}^{n} k = \frac{n(n+1)}{2} \).

10. Let \( a \) and \( d \) be real numbers. Find a formula for \( \sum_{k=1}^{n} (a + kd) \) and then prove your result by induction.

11. Consider the geometric series, \( \sum_{k=1}^{n} ar^{k-1} \). Prove by induction that if \( r \neq 1 \), then

\[
\sum_{k=1}^{n} ar^{k-1} = \frac{a - ar^n}{1 - r}.
\]

12. This problem is a continuation of Problem 11. It is discussed more later but you might want to give it a try now. You put money in the bank and it accrues interest at the rate of \( r \) per payment period. These terms need a little explanation. If the payment period is one month, and you started with $100 then the amount at the end of one month would equal \( 100 (1 + r) = 100 + 100r \). In this the second term is the interest and the first is called the principal. Now you have \( 100 (1 + r) \) in the bank. How much will you have at the end of the second month? By analogy to what was just done it would equal

\[
100 (1 + r) + 100 (1 + r) r = 100 (1 + r)^2.
\]

In general, the amount you would have at the end of \( n \) months would be \( 100 (1 + r)^n \).

(When a bank says they offer 6% compounded monthly, this means \( r \), the rate per payment period equals .06/12.) In general, suppose you start with \( P \) and it sits in the bank for \( n \) payment periods. Then at the end of the \( n^{th} \) payment period, you would have \( P (1 + r)^n \) in the bank. In an ordinary annuity, you make payments, \( P \) at the end of each payment period, the first payment at the end of the first payment period. Thus there are \( n \) payments in all. Each accrue interest at the rate of \( r \) per payment period. Using Problem 11, find a formula for the amount you will have in the bank at the end of \( n \) payment periods? This is called the future value of an ordinary annuity. **Hint:** The first payment sits in the bank for \( n-1 \) payment periods and so this payment becomes \( P (1 + r)^{n-1} \). The second sits in the bank for \( n-2 \) payment periods so it grows to \( P (1 + r)^{n-2} \), etc.
13. This topic will also be discussed later. Suppose you want to buy a house by making \( n \) equal monthly payments. Typically, \( n \) is pretty large, 360 for a thirty year loan. Clearly a payment made 10 years from now can’t be considered as valuable to the bank as one made today. This is because the one made today could be invested by the bank and having accrued interest for 10 years would be far larger. So what is a payment made at the end of \( k \) payment periods worth today assuming money is worth \( r \) per payment period? Shouldn’t it be the amount, \( Q \) which when invested at a rate of \( r \) per payment period would yield \( P \) at the end of \( k \) payment periods? Thus from Problem 12 \( Q(1 + r)^k = P \) and so \( Q = P(1 + r)^{-k} \). Thus this payment of \( P \) at the end of \( n \) payment periods, is worth \( P(1 + r)^{-k} \) to the bank right now. It follows the amount of the loan should equal the sum of these “discounted payments”. That is, letting \( A \) be the amount of the loan,

\[
A = \sum_{k=1}^{n} P(1 + r)^{-k}.
\]

Using Problem 11, find a formula for the right side of the above formula. This is called the present value of an ordinary annuity.

14. Suppose the available interest rate is 7% per year and you want to take a loan for $100,000 with the first monthly payment at the end of the first month. If you want to pay off the loan in 20 years, what should the monthly payments be? \textbf{Hint:} The rate per payment period is \( 0.07/12 \). See the formula you got in Problem 13 and solve for \( P \).

15. Consider the first five rows of Pascal’s \(^2\) triangle

\[
\begin{array}{cccc}
1 & 1 & 1 & 1 \\
1 & 2 & 1 & 3 \\
1 & 3 & 3 & 1 \\
1 & 4 & 6 & 4 \\
1 & 5 & 10 & 10 & 5 & 1
\end{array}
\]

What would the sixth row be? Now consider that \((x + y)^1 = 1x + 1y\), \((x + y)^2 = x^2 + 2xy + y^2\), and \((x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3\). Give a conjecture about that \((x + y)^5\) would be.

16. Based on Problem 15 conjecture a formula for \((x + y)^n\) and prove your conjecture by induction. \textbf{Hint:} Letting the numbers of the \( n^{th} \) row of Pascal’s triangle be denoted by \( \binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n} \) in reading from left to right, there is a relation between the numbers on the \((n+1)^{st}\) row and those on the \( n^{th} \) row, the relation being \( \binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \). This is used in the inductive step.

17. Let \( \binom{n}{k} \equiv \frac{n!}{(n-k)!k!} \) where 0! \( \equiv 1 \) and \((n+1)!! \equiv (n+1)! \) for all \( n \geq 0 \). Prove that whenever \( k \geq 1 \) and \( k \leq n \), then \( \binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \). Are these numbers, \( \binom{n}{k} \) the same as those obtained in Pascal’s triangle? Prove your assertion.

18. The binomial theorem states \((a + b)^n = \sum_{k=0}^{n} \binom{n}{k}a^{n-k}b^k\). Prove the binomial theorem by induction. \textbf{Hint:} You might try using the preceding problem. This problem will also be presented later.

19. Prove by induction that for all \( k \geq 4 \), \( 2^k \leq k! \)

\(^2\)Blaise Pascal lived in the 1600’s and is responsible for the beginnings of the study of probability.
20. Use the Problems 19 and 8 to verify for all \( n \in \mathbb{N} \), \((1 + \frac{1}{n})^n \leq 3\).

21. Prove by induction that \(1 + \sum_{i=1}^{n} i (i!) = (n + 1)!\).

22. I can jump off the top of the Empire State Building without suffering any ill effects. Here is the proof by induction. If I jump from a height of one inch, I am unharmed. Furthermore, if I am unharmed from jumping from a height of \( n \) inches, then jumping from a height of \( n + 1 \) inches will also not harm me. This is self evident and provides the induction step. Therefore, I can jump from a height of \( n \) inches for any \( n \). What is the matter with this reasoning?

23. All horses are the same color. Here is the proof by induction. A single horse is the same color as himself. Now suppose the theorem that all horses are the same color is true for \( n \) horses and consider \( n + 1 \) horses. Remove one of the horses and use the induction hypothesis to conclude the remaining \( n \) horses are all the same color. Put the horse which was removed back in and take out another horse. The remaining \( n \) horses are the same color by the induction hypothesis. Therefore, all \( n + 1 \) horses are the same color as the \( n - 1 \) horses which didn’t get moved. This proves the theorem. Is there something wrong with this argument?

24. Using the triangle inequality, verify using math induction that for all \( m \in \mathbb{N} \),

\[
\left| \sum_{k=1}^{m} a_k \right| \leq \sum_{k=1}^{m} |a_k|.
\]

1.11 Completeness of \( \mathbb{R} \)

By Theorem 1.9.9, between any two real numbers, points on the number line, there exists a rational number. This suggests there are a lot of rational numbers, but it is not clear from this Theorem whether the entire real line consists of only rational numbers. Some people might wish this were the case because then each real number could be described, not just as a point on a line but also algebraically, as the quotient of integers. Before 500 B.C., a group of mathematicians, led by Pythagoras believed in this, but they discovered their beliefs were false. It happened roughly like this. They knew they could construct the square root of two as the diagonal of a right triangle in which the two sides have unit length; thus they could regard \( \sqrt{2} \) as a number. Unfortunately, they were also able to show \( \sqrt{2} \) could not be written as the quotient of two integers. This discovery that the rational numbers could not even account for the results of geometric constructions was very upsetting to the Pythagoreans, especially when it became clear there were an endless supply of such “irrational” numbers.

This shows that if it is desired to consider all points on the number line, it is necessary to abandon the attempt to describe arbitrary real numbers in a purely algebraic manner using only the integers. Some might desire to throw out all the irrational numbers, and considering only the rational numbers, confine their attention to algebra, but this is not the approach to be followed here because it will effectively eliminate every major theorem of calculus. In this book real numbers will continue to be the points on the number line, a line which has no holes. This lack of holes is more precisely described in the following way.

**Definition 1.11.1** A non empty set, \( S \subseteq \mathbb{R} \) is bounded above (below) if there exists \( x \in \mathbb{R} \) such that \( x \geq (\leq) s \) for all \( s \in S \). If \( S \) is a nonempty set in \( \mathbb{R} \) which is bounded above, then a number, \( l \) which has the property that \( l \) is an upper bound and that every other upper bound is no smaller than \( l \) is called a least upper bound, \( \text{l.u.b.} (S) \) or often \( \sup (S) \). If \( S \) is a nonempty set bounded below, define the greatest lower bound, \( \text{g.l.b.} (S) \)
1.11. COMPLETENESS OF $\mathbb{R}$

or $\inf (S)$ similarly. Thus $g$ is the greatest lower bound of $S$ means $g$ is a lower bound for $S$ and it is the largest of all lower bounds. If $S$ is a nonempty subset of $\mathbb{R}$ which is not bounded above, this information is expressed by saying $\sup (S) = +\infty$ and if $S$ is not bounded below, $\inf (S) = -\infty$.

Every existence theorem in calculus depends on some form of the completeness axiom.

**Axiom 1.11.2 (completeness)** Every nonempty set of real numbers which is bounded above has a least upper bound and every nonempty set of real numbers which is bounded below has a greatest lower bound.

It is this axiom which distinguishes Calculus from Algebra. A fundamental result about sup and inf is the following.

**Proposition 1.11.3** Let $S$ be a nonempty set and suppose $\sup (S)$ exists. Then for every $\delta > 0$,

$$S \cap (\sup (S) - \delta, \sup (S)] \neq \emptyset.$$  

If $\inf (S)$ exists, then for every $\delta > 0$,

$$S \cap [\inf (S), \inf (S) + \delta) \neq \emptyset.$$  

**Proof:** Consider the first claim. If the indicated set equals $\emptyset$, then $\sup (S) - \delta$ is an upper bound for $S$ which is smaller than $\sup (S)$, contrary to the definition of $\sup (S)$ as the least upper bound. In the second claim, if the indicated set equals $\emptyset$, then $\inf (S) + \delta$ would be a lower bound which is larger than $\inf (S)$ contrary to the definition of $\inf (S)$.

### 1.11.1 Roots

What is $\sqrt[5]{7}$? You probably know this is the number which when multiplied by itself 5 times gives 7. Is there any such number? Why? You can ask for it on your calculator and it will give you a number which multiplied by itself 5 times will yield a number which is close to 7 but it isn’t exactly right. Why should there exists a number which works exactly? Every one you find, appears to be some sort of approximation at best. If you can’t produce one, why should you believe it is even there? The following is an argument that roots exist. You fill in the details of the argument. Basically, roots exist because of completeness of the real line. Here is a lemma.

**Lemma 1.11.4** Suppose $n \in \mathbb{N}$ and $a > 0$. Then if $x^n - a \neq 0$, there exists $\delta > 0$ such that whenever

$$y \in (x - \delta, x + \delta),$$

it follows $y^n - a \neq 0$ and has the same sign as $x^n - a$.

**Proof:** Using the binomial theorem,

$$y^n - a = (y - x + x)^n - a$$

$$= \sum_{k=0}^{n-1} \binom{n}{k} (y-x)^{n-k} x^k + x^n - a$$

Let $|y - x| < 1$. Then using the triangle inequality, see Problem 24 on Page 26, it follows that for $|y - x| < 1$,

$$|y^n - a - (x^n - a)| \leq |y - x| \sum_{k=0}^{n-1} \binom{n}{k} |x|^k \equiv C |x - y|$$
where, as indicated, \( C = \sum_{k=0}^{n-1} \binom{n}{k} |x|^k \). Let

\[
\delta = \min \left( \frac{|x^n - a|}{2C}, 1 \right)
\]

Then if \( y \in (x - \delta, x + \delta) \),

\[
|y^n - a - (x^n - a)| \leq C|x - y| < C\delta \leq \left| \frac{x^n - a}{2} \right|
\]

This says the distance between \( y^n - a \) and \( x^n - a \) is less than the distance between \( \frac{x^n - a}{2} \) and 0. Consequently, \( y^n - a \neq 0 \) and has the same sign as \( x^n - a \). (Draw a picture.)

This proves the lemma.

**Theorem 1.11.5** Let \( a > 0 \) and let \( n > 1 \). Then there exists a unique \( x > 0 \) such that \( x^n = a \).

**Proof:** Let \( S \) denote those numbers \( y \geq 0 \) such that \( t^n - a < 0 \) for all \( t \in [0, y] \). One such number is 0. If \( a \geq 1 \), then a short proof by induction shows \( a^n > a \) and so, in this case, \( S \) is bounded above by \( a \). If \( a < 1 \), then another short argument shows \( (1/a)^n > a \) and so \( S \) is bounded above by \( 1/a \). By completeness, there exists \( x \), the least upper bound of \( S \). Thus for all \( y \leq x \), \( y^n - a < 0 \) since if this is not so, then \( x \) was not a least upper bound to \( S \). If \( x^n - a > 0 \), then by the lemma, \( y^n - a > 0 \) on some interval \( (x - \delta, x + \delta) \). Thus \( x \) fails to be a the least upper bound because an upper bound is \( x - \delta/2 \). If \( x^n - a < 0 \), then by the lemma, \( y^n - a < 0 \) on some interval \( (x - \delta, x + \delta) \) and so \( x \) is not even an upper bound because \( S \) would then contain \( [0, x + \delta) \). Hence the only other possibility is that \( x^n - a = 0 \). That is, \( x \) is an \( n^{th} \) root of \( a \).

This has shown that \( a \) has a positive \( n^{th} \) root. Could it have two? Suppose \( x, z \) both work. If \( z > x \), then by the binomial theorem,

\[
z^n = (x + z - x)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} (z - x)^k
\]

\[
= x^n + \sum_{k=0}^{n-1} \binom{n}{k} x^{n-k} (z - x)^k = a + \sum_{k=0}^{n-1} \binom{n}{k} x^{n-k} (z - x)^k > a.
\]

Turning the argument around, it is also not possible that \( z < x \). Thus the \( n^{th} \) root is also unique. This proves the theorem.

From now on, we will use this fact that \( n^{th} \) roots exist whenever it is convenient to do so.

### 1.12 Exercises

1. Let \( S = [2, 5] \). Find sup \( S \). Now let \( S = [2, 5) \). Find sup \( S \). Is sup \( S \) always a number in \( S \)? Give conditions under which sup \( S \in S \) and then give conditions under which inf \( S \in S \).

2. Show that if \( S \neq \emptyset \) and is bounded above (below) then sup \( S \) (inf \( S \)) is unique. That is, there is only one least upper bound and only one greatest lower bound. If \( S = \emptyset \) can you conclude that 7 is an upper bound? Can you conclude 7 is a lower bound? What about 13.5? What about any other number?
3. Let $S$ be a set which is bounded above and let $-S$ denote the set $\{-x : x \in S\}$. How are $\inf(-S)$ and $\sup(S)$ related? **Hint:** Draw some pictures on a number line. What about $\sup(-S)$ and $\inf S$ where $S$ is a set which is bounded below?

4. Solve the following equations which involve absolute values.
   (a) $|x + 1| = |2x + 3|
   (b) $|x + 1| - |x + 4| = 6$

5. Solve the following inequalities which involve absolute values.
   (a) $|2x - 6| < 4
   (b) |x - 2| < |2x + 2|

6. Which of the field axioms is being abused in the following argument that $0 = 2$?
   Let $x = y = 1$. Then
   
   $0 = x^2 - y^2 = (x - y)(x + y)$
   
   and so
   
   $0 = (x - y)(x + y)$.
   
   Now divide both sides by $x - y$ to obtain
   
   $0 = x + y = 1 + 1 = 2$.

7. Give conditions under which equality holds in the triangle inequality.

8. Show that if it is assumed $\mathbb{R}$ is complete, then the Archimedian property can be proved. **Hint:** Suppose completeness and let $a > 0$. If there exists $x \in \mathbb{R}$ such that $na \leq x$ for all $n \in \mathbb{N}$, then $x/a$ is an upper bound for $\mathbb{N}$. Let $l$ be the least upper bound and argue there exists $n \in \mathbb{N} \cap [l - 1/4, l]$. Now what about $n + 1$?

9. Obtain a number, $\delta > 0$, such that if $|x - 4| < \delta$, then $|\sqrt{x} - 2| < 1/10$.

10. Suppose $\varepsilon > 0$ is a given positive number. Obtain a number, $\delta > 0$, such that if $|x - 1| < \delta$, then $|\sqrt{x} - 1| < \varepsilon$. **Hint:** This $\delta$ will depend in some way on $\varepsilon$. You need to tell how.

1.13 Counting

1.13.1 Combinations

The fundamental problem is to find the number of ways of selecting a subset of $k \leq n$ elements from a set having $n$ elements. For example, consider the set $S = \{1, 2, 3\}$. How many subsets having two elements are there? In this case, you can simply list them. Here they are

$\{1, 2\}, \{1, 3\}, \{2, 3\}$

This seems easy enough, but what if you had a set of 52 things like a deck of cards and you wanted the number of ways of picking a set of 5 things from it. Then it would be a little harder. Here is some standard notation.

**Definition 1.13.1** Let $0 \leq k \leq n$. Then $\binom{n}{k}$ denotes the number of subsets of a set having $n$ elements which have $k$ elements.
Here are some obvious assertions.

\[
\binom{n}{0} = \binom{n}{n} = 1, \quad \binom{n}{1} = n \quad (1.3)
\]

The first says there is one subset which has no elements in it. Of course it is the empty set. The next says there is one subset of a set having \( n \) things which has \( n \) things in it. Of course, this would be the whole set itself. The last says there are \( n \) subsets which have a single element of the set in them. Now to get a formula for \( \binom{n}{k} \), here is a lemma.

**Lemma 1.13.2** Let \( n \) be a positive integer and let \( 1 \leq k \leq n \). Then

\[
\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}
\]

**Proof:** Letting \( 1 \leq k \leq n \), suppose your set of \( n+1 \) things is

\[\{a_1, \ldots, a_n, a_{n+1}\}\]

Here \( a_i \) denotes the \( i \)th element of the set and this is just a list of the elements of the set. Then there are two ways to select a set of \( k \) things from this set depending on whether \( a_{n+1} \) is in the set of \( k \) things. If it is, there are exactly \( \binom{n}{k-1} \) ways to obtain such a set of \( k \) things because it must be the number of ways of selecting the remaining \( k-1 \) elements from the first \( n \) elements in the set. The other case is where all \( k \) elements are selected from the first \( n \) elements of the set. By definition, there are \( \binom{n}{k} \) ways to do this. Thus

\[
\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}
\]

This proves the lemma.

**Definition 1.13.3** Let \( 0! \equiv 1 \) and for \( n \in \mathbb{N}, n! \equiv n(n-1)(n-2)\cdots 1 \). This is called the factorial symbol. We say \( n! \) as \( n \) factorial.

With this definition, it is easy to give a simple description of \( \binom{n}{k} \).

**Theorem 1.13.4** Let \( 0 \leq k \leq n \). Then

\[
\binom{n}{k} = \frac{n!}{k!(n-k)!}
\]

**Proof:** You see easily this is true if \( n = 1 \). In this case, the only possibilities for \( k \) are 0,1 the the formula gives the right answer in either of these cases. Assume the formula holds for \( n \). Then by Lemma 1.13.2 and the induction hypothesis, if \( 1 \leq k \leq n \)

\[
\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}
\]

\[
= \frac{n!}{(k-1)!(n-k+1)!} + \frac{n!}{k!(n-k)!}
\]

\[
= \frac{kn!}{k!(n-k+1)!} + \frac{n!(n-k+1)}{k!(n-k)!(n-k+1)}
\]
and so this proves the formula in the case that \(1 \leq k \leq n\). If \(k = 0\) or \(n + 1\), the definition of the factorial symbol and the obvious observations of 1.3 shows the formula holds in these cases also. This proves the theorem.

Notice that \(\binom{n}{k} = \binom{n}{n-k}\).

### 1.13.2 The Binomial Theorem

The Binomial theorem is one of the most useful and fundamental theorems in algebra. It is easy to prove from the above using induction. Here it is.

**Theorem 1.13.5** Let \(n \in \mathbb{N}\). Then

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

**Proof:** In case \(n = 1\), both sides reduce to \(a + b\) so it works in this case. Suppose now it works for \(n\). Then by induction,

\[
(a + b)^{n+1} = (a + b) \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k
\]

\[
= \sum_{k=0}^{n} \binom{n}{k} a^{n+1-k} b^k + \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^{k+1}
\]

\[
= \sum_{k=0}^{n} \binom{n}{k} a^{n+1-k} b^k + \sum_{k=1}^{n+1} \binom{n}{k-1} a^{n+1-k} b^k
\]

\[
= a^{n+1} + \sum_{k=1}^{n} \binom{n}{k} a^{n+1-k} b^k + \sum_{k=1}^{n} \binom{n}{k-1} a^{n+1-k} b^k + b^{n+1}
\]

\[
= a^{n+1} + \sum_{k=1}^{n} \left( \binom{n}{k} + \binom{n}{k-1} \right) a^{n+1-k} b^k + b^{n+1}
\]

By Lemma 1.13.2 this reduces to

\[
= \sum_{k=1}^{n+1} \binom{n+1}{k} a^{n+1-k} b^k + b^{n+1}
\]

which shows that when the formula holds for \(n\) it also holds for \(n + 1\). This proves the theorem.
1.14 Counting And Basic Probability

You do an experiment $n$ times and there are two possible outcomes to this experiment each time it occurs, a “success” having probability $p$, a positive number less than 1 and a “failure” having probability $(1 - p)$. For example, you could have $p$ be the probability of getting a 4 when you roll a pair of fair dice. What would this probability be? Here is a table of possible outcomes for the pair of dice.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

The first number represents the one on the first die and the second represents the number on the second die. (die is singular for dice) How many ways are there to get a 4? From the table, there are exactly 3 ways, (3, 1), (2, 2), (1, 3). How many possible outcomes are there? There are 36. Thus if every outcome is as likely as any other, the probability of rolling a 4 is $3/36$ or $1/12$.

Now in a succession of rolls of the dice, the probability of a particular outcome on roll $k$ is not affected by what happened on earlier rolls of the dice. Each time the dice are rolled, the probability of rolling a four is $1/12$ and the probability rolling a non four is $11/12$.

What is the probability of rolling a 4 twice in a row? In this case there would be $36^2$ possible outcomes and only $4^2$ of them are favorable to rolling two fours in succession. (Four possibilities for the first roll of the dice and for each of these, four for the second.) Thus the probability of this occuring is

$$\frac{1}{12^2}$$

What about the probability of a four on the first roll and a non four on the second? This probability is

$$\frac{1}{12} \cdot \frac{11}{12} = \frac{11}{144}$$

You can determine this the same way by counting the ways favorable to the desired outcome and dividing this by the number of possible outcomes. Similarly, the probability of rolling a non four followed by a four would be

$$\frac{11}{12} \cdot \frac{1}{12} = \frac{11}{144}$$

More generally, the probability of getting $k$ fours and $n-k$ non fours in a particular order would be

$$\left(\frac{1}{12}\right)^k \left(\frac{11}{12}\right)^{n-k}$$

More generally, you have a situation where the probability of $k$ success with probability $p$ and $(n-k)$ failures happening with probability $q \equiv (1 - p)$ in any particular order is $p^kq^{n-k}$. What is the probability of having $k$ successes in $n$ trials? This is known as the binomial distribution. How many ways can it happen? It can happen exactly the number of ways there are of selecting $k$ of the $n$ trials. There are $\binom{n}{k}$ ways for
this to happen. Therefore, since each of these has the same probability, $p^k q^{n-k}$, the probability of $k$ successes in $n$ trials is

$$\binom{n}{k} p^k q^{n-k}$$

This motivates the following definition of the binomial distribution.

**Definition 1.14.1** Define a “random variable” $X$ to be the number of successes in $n$ trials. Thus $X$ has values $0, 1, \cdots, n$. If the probability that $X$ has value $k$, written $P(X = k)$ is given by

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

then $X$ is said to have a binomial distribution.

Note that the sum

$$\sum_{k=0}^{n} P(X = k)$$

needs to equal 1 because with probability 1 the random variable must achieve one of the numbers $1, 2, \cdots, n$. If this does not happen, there is something wrong. Does it happen? By the binomial theorem,

$$\sum_{k=0}^{n} P(X = k) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p + q)^n = 1^n = 1.$$

### 1.15 Exercises

1. Let $k \leq n$ where $k$ and $n$ are natural numbers. $P(n, k)$, permutations of $n$ things taken $k$ at a time, is defined to be the number of different ways to form an ordered list of $k$ of the numbers, $\{1, 2, \cdots, n\}$. Show

$$P(n, k) = n \cdot (n-1) \cdots (n-k+1) = \frac{n!}{(n-k)!}.$$

2. Now consider the word “mississippi”. By rearranging the letters, how many distinctly different words can you obtain? Note that for each list of these letters the four different s are indistinguishable. There are therefore, $4!$ ways which are not really different.

3. Using Problem 2 show the number of ways of selecting a set of $k$ things from a set of $n$ things is $\binom{n}{k}$.

4. Prove the binomial theorem from Problem 3. **Hint:** When you take $(x + y)^n$, note that the result will be a sum of terms of the form, $a_k x^{n-k} y^k$ and you need to determine what $a_k$ should be. Imagine writing $(x + y)^n = (x + y)(x + y) \cdots (x + y)$ where there are $n$ factors in the product. Now consider what happens when you multiply. Each factor contributes either an $x$ or a $y$ to a typical term.

5. Prove by induction that $n < 2^n$ for all natural numbers, $n \geq 1$.

6. Prove by the binomial theorem and Problem 3 that the number of subsets of a given finite set containing $n$ elements is $2^n$.
7. Show that for \( p \in (0, 1) \), \( \sum_{k=0}^{n} \binom{n}{k} kp^k (1-p)^{n-k} = np \).

8. Using the binomial theorem prove that for all \( n \in \mathbb{N} \),
   \[
   \left(1 + \frac{1}{n}\right)^n \leq \left(1 + \frac{1}{n+1}\right)^{n+1}.
   \]
   \textbf{Hint:} Show first that \( \binom{n}{k} = \frac{n(n-1) \cdots (n-k+1)}{k!n^k} \). By the binomial theorem,
   \[
   \left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{1}{n}\right)^k = \sum_{k=0}^{n} \frac{n(n-1) \cdots (n-k+1)}{k!n^k}.
   \]
   Now consider the term \( \frac{n(n-1) \cdots (n-k+1)}{k!n^k} \) and note that a similar term occurs in the binomial expansion for \( \left(1 + \frac{1}{n+1}\right)^{n+1} \) except that \( n \) is replaced with \( n+1 \) wherever this occurs. Argue the term got bigger and then note that in the binomial expansion for \( \left(1 + \frac{1}{n+1}\right)^{n+1} \), there are more terms.

9. Let \( n \) be a natural number and let \( k_1 + k_2 + \cdots + k_r = n \) where \( k_i \) is a non negative integer. The symbol \( \binom{n}{k_1 k_2 \cdots k_r} \)
   denotes the number of ways of selecting \( r \) subsets of \( \{1, \cdots, n\} \) which contain \( k_1, k_2, \cdots, k_r \) elements in them. Find a formula for this number.

10. Is it ever the case that \( (a + b)^n = a^n + b^n \) for \( a \) and \( b \) positive real numbers?

11. Is it ever the case that \( \sqrt{a^2 + b^2} = a + b \) for \( a \) and \( b \) positive real numbers?

12. Is it ever the case that \( \frac{1}{x+y} = \frac{1}{x} + \frac{1}{y} \) for \( x \) and \( y \) positive real numbers?

13. Derive a formula for the multinomial expansion, \( (\sum a_k)^n \) which is analogous to the binomial expansion. \textbf{Hint:} See Problems 4 and 9.

14. Let \( X \) be a binomial random variable. Thus \( P(X = k) = \binom{n}{k} p^k q^{n-k} \) where \( p \) is the probability of success and \( q = 1 - p \) is the probability of failure. The expected value of \( X \) denoted as \( E(X) \), is defined as
   \[
   \sum_{k=0}^{n} k P(X = k).
   \]
   Show the expected value of \( X \) equals \( np \).

15. The variance of the random variable in the above problem is defined as
   \[
   \sigma^2 = \sum_{k=0}^{n} (k - E(X))^2 P(X = k)
   \]
   Find \( \sigma^2 \). You should get \( npq \).

16. Find the probability of drawing from a shuffled deck of playing cards four hearts. \textbf{Hint:} Use principles of drawing to find the number of ways of drawing four hearts. There are 13 of these. Now how many ways can you pull out four of them? Then note there are 52 cards in all. How many ways can you pull out four cards from these.
17. Find the probability of obtaining 2 clubs and three spades from a shuffled deck of cards.

18. A pond has \( N \) fish and 120 of these are marked fish. What is the probability in terms of \( N \) of catching 10 fish, two of which are marked and 8 of which are unmarked?

19. Show that in general, for \( k \leq m < N \)

\[
\sum_{j=0}^{k} \binom{m}{j} \binom{N-m}{k-j} = 1
\]

If \( X \) is a random variable having values in \( \{0, 1, \cdots, k\} \) such that the probability that \( X = j \) is given by the \( j \)th term of the above sum, then \( X \) is said to have a hypergeometric distribution. Much much more can be said about this topic. Hint: If you pick \( k \) things from \( N \) things \( m \) of which are marked and \( N-m \) unmarked, there are various ways to do it determined by the value of \( j \), the number of marked things out of your sample of \( k \) things.

20. Suppose a pair of dice has one blue and the other one red. What is the probability that when they are rolled the blue die delivers a strictly larger number than the red die? Now what is the probability that either this happen or a 6 is rolled? What is the probability that the blue is greater than the red and a 6 is rolled?

21. Explain why in general, \( P(A \cup B) = P(A) + P(B) - P(A \cap B) \) where \( A, B \) are two events such as in the above problem having the blue be larger than the red die or rolling a 6.

22. Let \( X \) be the random variable which gives the number of heads when you flip a coin 6 times. Which value of \( X \) has the highest probability? What is the expected value of \( X \)? What is the variance of \( X \). For these last parts, see Problem 14 and 15 above.
Functions

2.1 Generalities

The concept of a function is that of something which gives a unique output for a given input.

Definition 2.1.1 Consider two sets, $D$ and $R$ along with a rule which assigns a unique element of $R$ to every element of $D$. This rule is called a function and it is denoted by a letter such as $f$. Given $x \in D$, $f(x)$ is the name of the thing in $R$ which results from doing $f$ to $x$. The symbol, $D(f) = D$ is called the domain of $f$. The set $R$, also written $R(f)$, is called the range of $f$. The set of all elements of $R$ which are of the form $f(x)$ for some $x \in D$ is often denoted by $f(D)$. When $R = f(D)$, the function, $f$, is said to be onto. If whenever $x \neq y$ it follows $f(x) \neq f(y)$ the function is called one to one. It is common notation to write $f : D(f) \to R$ to denote the situation just described in this definition where $f$ is a function defined on $D$ having values in $R$.

Example 2.1.2 Consider the list of numbers, \{1, 2, 3, 4, 5, 6, 7\} $\equiv D$. Define a function which assigns an element of $D$ to $R \equiv \{2, 3, 4, 5, 6, 7, 8\}$ by $f(x) \equiv x+1$ for each $x \in D$. This function is onto because every element of $R$ is the result of doing $f$ to something in $D$. The function is also one to one. This is because if $x + 1 = y + 1$, then it follows $x = y$. Thus different elements of $D$ must go to different elements of $R$.

In this example there was a clearly defined procedure which determined the function. However, sometimes there is no discernible procedure which yields a particular function.

Example 2.1.3 Consider the ordered pairs, (1, 2), (2, −2), (8, 3), (7, 6) and let $D \equiv \{1, 2, 8, 7\}$, the set of first entries in the given set of ordered pairs and let $R \equiv \{2, −2, 3, 6\}$, the set of second entries, and let $f(1) = 2$, $f(2) = −2$, $f(8) = 3$, and $f(7) = 6$.

This specifies a function even though it does not come from a convenient formula. This can get pretty complicated. For example, consider the following function defined on the positive real numbers having the following definition.

Example 2.1.4 For $x \in \mathbb{R}$ define

\[
f(x) = \begin{cases} \frac{1}{n} & \text{if } x = \frac{m}{n} \text{ in lowest terms for } m, n \in \mathbb{Z} \\ 0 & \text{if } x \text{ is not rational} \end{cases}
\]
This is a very interesting function called the Dirichlet function. Note that it is not defined in a simple way from a formula.

**Example 2.1.5** Let $D$ consist of the set of people who have lived on the earth except for Adam and for $d \in D$, let $f(d) \equiv$ the biological father of $d$. Then $f$ is a function.

This function is not the sort of thing studied in calculus but it is a function just the same. Next are some physically defined functions.

**Example 2.1.6** Consider a weight which is suspended at one end of a spring which is attached at the other end to the ceiling. Suppose the weight has extended the spring so that the force exerted by the spring exactly balances the force resulting from the weight on the spring. Measure the displacement of the mass, $x$, from this point with the positive direction being up, and define a function as follows: $x(t)$ will equal the displacement of the spring at time $t$ given knowledge of the velocity of the weight and the displacement of the weight at some particular time.

In calculus or differential equations you learn to find a formula which will describe $x(t)$. However, this is not possible at this time.

**Example 2.1.7** Certain chemicals decay with time. Suppose $A_0$ is the amount of chemical at some given time. Then you could let $A(t)$ denote the amount of the chemical at time $t$.

These last two examples show how physical problems can result in functions.

## 2.2 Real Functions

In this chapter the functions are defined on some subset of $\mathbb{R}$ having values in $\mathbb{R}$. When $D(f)$ is not specified, it is understood to consist of everything for which $f$ makes sense.

The following definition gives several ways to make new functions from old ones.

For example, consider the function

$$f(x) = \sqrt{4-x^2}$$

Then $D(f)$ consists of all real numbers between $-2$ and $2$, the interval $[-2, 2]$. This is because it makes no sense to take the square root of a negative number. As another example, consider

$$f(x) = \frac{1}{1-x}$$

This time $D(f)$ consists of all real numbers not equal to $1$. This is because you can’t divide by $0$. Written in terms of intervals, $D(f) = (-\infty, 1) \cup (1, \infty)$.

**Definition 2.2.1** Let $f, g$ be functions with values in $\mathbb{R}$. Let $a, b$ be points of $\mathbb{R}$. Then $af + bg$ is the name of a function whose domain is $D(f) \cap D(g)$ which is defined as

$$(af + bg)(x) = af(x) + bg(x).$$

The function, $fg$ is the name of a function which is defined on $D(f) \cap D(g)$ given by

$$(fg)(x) = f(x)g(x).$$

Similarly for $k$ an integer, $f^k$ is the name of a function defined as

$$f^k(x) = (f(x))^k$$
The function, \( f/g \), is the name of a function whose domain is
\[ D(f) \cap \{ x \in D(g) : g(x) \neq 0 \} \]
defined as
\[ (f/g)(x) = f(x)/g(x). \]
If \( f : D(f) \to X \) and \( g : D(g) \to Y \), then \( g \circ f \) is the name of a function whose domain
is
\[ \{ x \in D(f) : f(x) \in D(g) \} \]
which is defined as
\[ g \circ f(x) \equiv g(f(x)). \]
This is called the composition of the two functions.

You should note that \( f(x) \) is not a function. It is the value of the function at the
point, \( x \). The name of the function is \( f \). Nevertheless, people often write \( f(x) \) to denote
a function and it doesn’t cause too many problems in beginning courses. When this
is done, the variable, \( x \) should be considered as a generic variable free to be anything
in \( D(f) \). I will use this slightly sloppy abuse of notation whenever convenient. Thus,
\( x^2 + 4 \) may mean the function, \( f \), given by \( f(x) = x^2 + 4 \).

Sometimes people get hung up on formulas and think that the only functions of
importance are those which are given by some simple formula. It is a mistake to think
this way. Functions involve a domain and a range and a function is determined by
what it does to something in its domain, you have told what the function is.

**Example 2.2.2** Let \( f(t) = t \) and \( g(t) = 1 + t \). Then \( fg : \mathbb{R} \to \mathbb{R} \) is given by
\[ fg(t) = t(1 + t) = t + t^2. \]

**Example 2.2.3** Let \( f(t) = 2t + 1 \) and \( g(t) = \sqrt{1 + t} \). Then
\[ g \circ f(t) = \sqrt{1 + (2t + 1)} = \sqrt{2t + 2} \]
for \( t \geq -1 \). If \( t < -1 \) the inside of the square root sign is negative so makes no sense.
Therefore, \( g \circ f : \{ t \in \mathbb{R} : t \geq -1 \} \to \mathbb{R} \).

Note that in this last example, it was necessary to fuss about the domain of \( g \circ f \)
because \( g \) is only defined for certain values of \( t \).

The concept of a one to one function is very important. This is discussed in the
following definition which introduces important notation.

**Definition 2.2.4** For any function, \( f : D(f) \subseteq X \to Y \), define the following set known
as the inverse image of \( y \).
\[ f^{-1}(y) \equiv \{ x \in D(f) : f(x) = y \}. \]
Thus \( f^{-1}(y) \) consists of all the \( x \) in \( D(f) \) such that \( f(x) = y \). There may be many
elements in this set, but when there is always only one element in this set for all \( y \in\)
\( f(D(f)) \), the function \( f \) is one to one sometimes written, \( 1 \to 1 \). Thus \( f \) is one to one,
\( 1 \to 1 \), if whenever \( f(x) = f(x_1) \), then \( x = x_1 \). Equivalently, if \( x \neq x_1 \), then
\( f(x) \neq f(x_1) \). If \( f \) is one to one, the inverse function, \( f^{-1} \) is defined on \( f(D(f)) \) and
\( f^{-1}(y) = x \) where \( f(x) = y \). Thus from the definition, \( f^{-1}(f(x)) = x \) for all \( x \in D(f) \)
and \( f(f^{-1}(y)) = y \) for all \( y \in f(D(f)) \). Defining \( \text{id} \) by \( \text{id}(z) \equiv z \) this says \( f \circ f^{-1} = \text{id} \)
and \( f^{-1} \circ f = \text{id} \).
Example 2.2.5 Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = 3x + 7$. Determine whether $f$ is one to one and if it is, find a formula for its inverse.

If $f(x) = f(y)$, then $3x + 7 = 3y + 7$. Subtracting 7 from both sides and then dividing both sides by 3 yields $x = y$. Therefore, this function is one to one. It is also onto. This is because you can solve the equation

$$3x + 7 = z$$

for $x$. In fact $x = \frac{1}{3}(z - 7)$. Now what about the inverse? The inverse function is defined according to the above definition. Thus $f(f^{-1}(x)) = x$. In other words,

$$3(f^{-1}(x)) + 7 = x$$

Now solve this for $f^{-1}(x)$. This gives

$$f^{-1}(x) = \frac{1}{3}(x - 7).$$

Lets check to see if it works.

$$f(f^{-1}(x)) = 3\left(\frac{1}{3}(x - 7)\right) + 7 = x - 7 + 7 = x$$

so it worked. You should also verify that $f^{-1}(f(x)) = x$.

Note there is nothing sacred about the name of the variable used in describing the function! Remember what the inverse is in terms of what it does and you will have a lot less confusion than if you try to memorize some system for finding it.

Also, for your information, it is sometimes impossible to find the inverse of a given function using algebraic techniques like that just used. By this, I mean that the inverse may exist but you can’t find a formula for it using algebra. For example, you have no way of finding the inverse function for

$$f(x) = x^7 + 2x^3 + x + 1$$

although it can be shown the function has an inverse function. To see this, consider its graph.

![Graph of a polynomial function](image)

It turns out the graph just keeps climbing as you proceed from left to right. Therefore, given $y$ there exists exactly one thing in $f^{-1}(y)$. In other words the function is one to one and so it has an inverse. However, you don’t know how to find a formula for this inverse and neither do I. The function just considered is an example of a polynomial.

Polynomials and rational functions are particularly easy functions to understand because they do come from a simple formula.

Definition 2.2.6 A function $f$ is a **polynomial** if

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$
where the \( a_i \) are real numbers and \( n \) is a nonnegative integer. In this case the degree of the polynomial, \( f(x) \) is \( n \). Thus the degree of a polynomial is the largest exponent appearing on the variable.

* \( f \) is a rational function if \[
f(x) = \frac{h(x)}{g(x)}
\] where \( h \) and \( g \) are polynomials.

For example, \( f(x) = 3x^5 + 9x^2 + 7x + 5 \) is a polynomial of degree 5 and
\[
\frac{3x^5 + 9x^2 + 7x + 5}{x^4 + 3x + x + 1}
\]
is a rational function.

Note that in the case of a rational function, the domain of the function might not be all of \( \mathbb{R} \). For example, if
\[
f(x) = \frac{x^2 + 8}{x + 1},
\]
the domain of \( f \) would be all real numbers not equal to \(-1\).

### 2.3 Cartesian Coordinates And Graphs

There is a general notion called the Cartesian product. It consists of ordered pairs of elements of two sets. More precisely,

**Definition 2.3.1** Given two sets, \( X \) and \( Y \), the Cartesian product of the two sets, written as \( X \times Y \), is assumed to be a set described as follows.
\[
X \times Y = \{(x, y) : x \in X \text{ and } y \in Y\}.
\]

\( \mathbb{R}^2 \) denotes the Cartesian product of \( \mathbb{R} \) with \( \mathbb{R} \). The order matters.

**Example 2.3.2** Suppose \( X \equiv \{1, 2\} \), \( Y \equiv \{a, b, c\} \). Find \( X \times Y \).

By definition this Cartesian produce is the set of all possible ordered pairs. Thus
\[
X \times Y = \{(1, a), (1, b), (1, c), (2, a), (2, b), (2, c)\}
\]

Of more interest is the following example.

**Example 2.3.3** Describe \( \mathbb{R} \times \mathbb{R} \).

This consists of all ordered pairs of the form \((x, y)\) where \( x \in \mathbb{R} \) and \( y \in \mathbb{R} \). Examples are \((1, 2), (3, 4, 7), (-3, 6)\). Note that \((1, 2) \neq (2, 1)\) because the numbers occur in different order.

The notion of a relation is more general than the notion of a function. A relation is simply a set of ordered pairs. In other words, a relation is a subset of the Cartesian product of two sets. The graph of a relation is a way to visualize the relation.

Recall the notion of the Cartesian coordinate system you probably saw earlier. It involved an \( x \) axis, a \( y \) axis, two lines which intersect each other at right angles and one identifies a point by specifying a pair of numbers. For example, the number \((2, 3)\) involves going 2 units to the right on the \( x \) axis and then 3 units directly up on a line perpendicular to the \( x \) axis. For example, consider the following picture.
Because of the simple correspondence between points in the plane and the coordinates of a point in the plane, it is often the case that people are a little sloppy in referring to these things. Thus, it is common to see \((x, y)\) referred to as a point in the plane. I will often indulge in this sloppiness. In terms of relations, if you graph the points as just described, you will have a way of visualizing the relation.

The reader has likely encountered the notion of graphing relations of the form \(y = 2x + 3\) or \(y = x^2 + 5\). The meaning of such an expression in terms of defining a relation is as follows. The relation determined by the equation \(y = 2x + 3\) means the set of all ordered pairs \((x, y)\) which are related by this formula. Thus the relation can be written as

\[
\{(x, y) : y = 2x + 3\}.
\]

The relation determined by \(y = x^2 + 5\) is

\[
\{(x, y) : y = x^2 + 5\}.
\]

Note that these relations are also functions. For the first, you could let \(f(x) = 2x + 3\) and this would tell you a rule which tells what the function does to \(x\). However, some relations are not functions. For example, you could consider \(x^2 + y^2 = 1\). Written more formally, the relation it defines is

\[
\{(x, y) : x^2 + y^2 = 1\}
\]

Now if you give a value for \(x\), there might be two values for \(y\) which are associated with the given value for \(x\). In fact

\[
y = \pm \sqrt{1 - x^2}
\]

Thus this relation would not be a function.

Recall how to graph a relation. You first found lots of ordered pairs which satisfied the relation. For example \((0, 3), (1, 5),\) and \((-1, 1)\) all satisfy \(y = 2x + 3\) which describes a straight line. Then you connected them with a curve. Here are some simple examples which you should see that you understand. First here is the graph of \(y = x^2 + 1\).
Now here is the graph of the relation $y = 2x + 1$ which is a straight line.

Sometimes a relation is defined using different formulas depending on the location of one of the variables. For example, consider

$$y = \begin{cases} 
6 + x & \text{if } x \leq -2 \\
x^2 & \text{if } -2 < x < 3 \\
1 - x & \text{if } x \geq 3 
\end{cases}$$

Then the graph of this relation is sketched below.
Here is a graph of the relation $x^2 + y^2 = 4$. You should find a few points satisfying the relation and verify this is what does result.

A very important type of relation is one of the form $y - y_0 = m(x - x_0)$, where $m, x_0,$ and $y_0$ are numbers. The reason this is important is that if there are two points, $(x_1, y_1)$, and $(x_2, y_2)$ which satisfy this relation, then

$$\frac{y_1 - y_2}{x_1 - x_2} = \frac{(y_1 - y_0) - (y_2 - y_0)}{x_1 - x_2} = m \frac{(x_1 - x_0) - m(x_2 - x_0)}{x_1 - x_2} = m.$$

Remember from high school, the slope of the line segment through two points is always the difference in the $y$ values divided by the difference in the $x$ values, taken in the same order. Sometimes this is referred to as the rise divided by the run. This shows that there is a constant slope, $m$, the slope of the line, between any pair of points satisfying this relation. Such a relation is called a straight line. Also, the point $(x_0, y_0)$ satisfies the relation. This is more often called the equation of the straight line.

Geometrically, this means the graph of the relation is a straight line but this will be clear later. For now it is only the algebraic significance which is featured.

**Example 2.3.4** Find the relation for a straight line which contains the point $(1, 2)$ and has constant slope equal to 3.

From the above discussion, $(y - 2) = 3(x - 1)$.

**Definition 2.3.5** Let $f : D(f) \to R(f)$ be a function. The graph of $f$ consists of the set,

$$\{(x, y) : y = f(x) \text{ for } x \in D(f)\}.$$
Note that knowledge of the graph of a function is equivalent to knowledge of the function. To find \( f(x) \), simply observe the ordered pair which has \( x \) as its first element and the value of \( y \) equals \( f(x) \). The graph of \( f \) can be represented by drawing a picture as mentioned earlier in the section on Cartesian coordinates beginning on Page 41. For example, consider the picture of a part of the graph of the function \( f(x) = 2x - 1 \).

Here is the graph of the function, \( f(x) = x^2 - 2 \)

The following example shows how a graph can help to keep track of what is going on.

**Example 2.3.6** Solve \( |x + 1| \leq |2x - 2| \)

In order to keep track of what is happening, it is a very good idea to graph the two relations, \( y = |x + 1| \) and \( y = |2x - 2| \) on the same set of coordinate axes. This is not a hard job. \( |x + 1| = x + 1 \) when \( x > -1 \) and \( |x + 1| = -1 - x \) when \( x \leq -1 \). Therefore, it is not hard to draw its graph. Similar considerations apply to the other relation. The result is

Equality holds exactly when \( x = 3 \) or \( x = \frac{1}{3} \) as in the preceding example. Consider \( x \) between \( \frac{1}{3} \) and 3. You can see these values of \( x \) do not solve the inequality. For example \( x = 1 \) does not work. Therefore, \( (\frac{1}{3}, 3) \) must be excluded. The values of \( x \) larger than 3 do not produce equality so either \( |x + 1| < |2x - 2| \) for these points or \( |2x - 2| < |x + 1| \) for these points. Checking examples, you see the first of the two cases is the one which holds. Therefore, \( [3, \infty) \) is included. Similar reasoning obtains \( (-\infty, \frac{1}{3}) \). It follows the solution set to this inequality is \( (-\infty, \frac{1}{3}) \cup [3, \infty) \).
2.4 Exercises

1. Sketch the graph of \( y = x^3 + 1 \).

2. Sketch the graph of \( y = x^2 - 2x + 1 \).

3. Sketch the graph of \( y = \frac{x}{x^2 + 1} \).

4. Sketch the graph of \( \frac{1}{1 + x^2} \).

5. Sketch the graph of the straight line which goes through the points \((1, 0)\) and \((2, 3)\) and find the relation which describes this line.

6. Sketch the graph of the straight line which goes through the points \((1, 3)\) and \((2, 3)\) and find the relation which describes this line.

7. Sketch the graph of the straight line which goes through the points \((1, 4)\) and \((1, 3)\) and find the relation which describes this line.

8. Sketch the graph of the straight line which goes through the points \((1, 0)\) and \((1, 3)\) and find the relation which describes this line.

9. Find an equation for the straight line which goes through the point \((2, 3)\) and has slope 5.

10. Find an equation for the straight line which goes through the point \((2, -3)\) and has slope -3.

11. Find an equation for the straight line which goes through the point \((2, 4)\) and has slope 0.

12. Find an equation for the straight line which goes through the point \((-2, -3)\) and has slope -3.

13. Consider the relation \( 2x + 3y = 6 \). Show this is an equation for a straight line, sketch the straight line and determine its slope and a point on the line. **Hint:** You could write the relation in the form

\[
y = \frac{1}{3} \left( 6 - 2x \right) = 2 + \left( \frac{-2}{3} \right)x.
\]

14. Consider the relation \( 3x + 2y = 6 \). Show this is an equation for a straight line, sketch the straight line and determine its slope and a point on the line.

15. Consider the relation \( ax + by = 6 \) where not both \( a \) and \( b \) equal zero. Show this is an equation for a straight line, and determine its slope and a point on the line. What if \( b = 0 \)? Consider the graph of the relation in this case separately.

16. Suppose \( a, b \neq 0 \). Find the equation of the line which goes through the points \((0, a)\) and \((b, 0)\).

17. Two lines are parallel if they have the same slope. Find the equation of the line through the point \((2, 3)\) which is parallel to the line whose equation is \(2x + 3y = 8\).

18. Sketch the graph of the relation defined as

\[
y = \begin{cases} 
1 & \text{if } x \leq -2 \\
1-x & \text{if } -2 < x < 3 \\
1+x & \text{if } x \geq 3
\end{cases}
\]
2.4. EXERCISES

19. Let \( g(t) \equiv \sqrt{2 - t} \) and let \( f(t) = \frac{1}{t} \). Find \( g \circ f \). Include the domain of \( g \circ f \).

20. Give the domains of the following functions.

   (a) \( f(x) = \frac{x + 3}{3x - 2} \)
   (b) \( f(x) = \sqrt{x^2 - 4} \)
   (c) \( f(x) = \sqrt{4 - x^2} \)
   (d) \( f(x) = \sqrt{\frac{x}{3x + 5}} \)
   (e) \( f(x) = \sqrt{\frac{x^2 - 4}{x + 1}} \)

21. Let \( f: \mathbb{R} \to \mathbb{R} \) be defined by \( f(t) \equiv t^3 + 1 \). Is \( f \) one to one? Can you find a formula for \( f^{-1} \)?

22. Solve \( |x + 2| < |3x - 3| \).

23. Tell when equality holds in the triangle inequality. Remember this says \( |x + y| \leq |x| + |y| \).

24. Solve \( |x + 2| \leq 8 + |2x - 4| \).

25. Let \( f: \{ t \in \mathbb{R} : t \neq -1 \} \to \mathbb{R} \) be defined by \( f(t) \equiv \frac{t}{t + 1} \). Find \( f^{-1} \) if possible.

26. A function, \( f: \mathbb{R} \to \mathbb{R} \) is a strictly increasing function if whenever \( x < y \), it follows that \( f(x) < f(y) \). If \( f \) is a strictly increasing function, does \( f^{-1} \) always exist? Explain your answer.

27. Let \( f(t) \) be defined by

\[
 f(t) = \begin{cases} 
 2t + 1 & \text{if } t \leq 1 \\
 t & \text{if } t > 1 
\end{cases}
\]

Find \( f^{-1} \) if possible.

28. Suppose \( f: D(f) \to R(f) \) is one to one, \( R(f) \subseteq D(g) \), and \( g: D(g) \to R(g) \) is one to one. Does it follow that \( g \circ f \) is one to one?

29. If \( f: \mathbb{R} \to \mathbb{R} \) and \( g: \mathbb{R} \to \mathbb{R} \) are two one to one functions, which of the following are necessarily one to one on their domains? Explain why or why not by giving a proof or an example.

   (a) \( f + g \)
   (b) \( fg \)
   (c) \( f^3 \)
   (d) \( f/g \)

30. Draw the graph of the function \( f(x) = x^3 + 1 \).

31. Draw the graph of the function \( f(x) = x^2 + 2x + 2 \).

32. Draw the graph of the function \( f(x) = \frac{x}{x^2} \).
2.5 Quadratic Functions

2.5.1 Maximizing And Minimizing

These are functions which are of the form

\[ f(x) = ax^2 + bx + c \]

Here are the graphs of two of these functions. First here is the graph of \( y = 3x^2 + 5x - 6 \).

\[ -4 \quad -2 \quad 0 \quad 2 \quad 4 \quad -40 \quad -20 \quad 0 \quad 40 \quad 60 \]

You notice it has a point where \( y \) is smallest, a minimum point. Now here is a graph of \( y = -3x^2 + 5x - 6 \).

\[ -2 \quad 0 \quad 2 \quad 4 \quad -20 \quad -40 \]

This graph has a point where \( y \) is largest, a maximum point. How can you find the largest value of \( y \) such that \((x, y)\) is a point on the graph of \( y = -3x^2 + 5x - 6 \) and the smallest value of \( y \) such that \((x, y)\) is a point on the graph of \( y = 3x^2 + 5x - 6 \)?

Example 2.5.1 Find the smallest value of \( 3x^2 + 5x - 6 \) and give the value of \( x \) at which this smallest value occurs.

Here is how you do this.

\[
3x^2 + 5x - 6 = 3 \left( x^2 + \frac{5}{3}x - 2 \right) \tag{2.2}
\
= 3 \left( x^2 + \frac{5}{3}x + \frac{25}{36} - \frac{25}{36} - 2 \right) \tag{2.3}
\
= 3 \left( \left( x + \frac{5}{6} \right)^2 - \frac{25}{36} - 2 \right)
\
= 3 \left( x + \frac{5}{6} \right)^2 - \frac{97}{12}
\
\]

When \( x = -5/6 \), \( 3x^2 + 5x - 6 \) is as small as possible and equals \(-\frac{97}{12}\) because the above shows that if \( x \) is anything else, the result would be something positive added to \(-\frac{97}{12}\).

The tricky part is in going from \(2.2\) to \(2.3\). The trick is to take one half of the number which multiplies the \( x \), square what you get, and add it in and subtract it off. The reason this is a good idea is the following simple observation.
Observation 2.5.2 Note that
\[
\left( x + \frac{b}{2} \right)^2 = \left( x + \frac{b}{2} \right) \left( x + \frac{b}{2} \right) = x \left( x + \frac{b}{2} \right) + \frac{b}{2} \left( x + \frac{b}{2} \right) \\
= x^2 + \frac{b}{2}x + \frac{b^2}{4} + \frac{b}{2} \left( x + \frac{b}{2} \right) \\
= x^2 + bx + \frac{b^2}{4}
\]
and so
\[
x^2 + bx = x^2 + bx + \frac{b^2}{4} - \frac{b^2}{4} = \left( x + \frac{b}{2} \right)^2 - \frac{b^2}{4}
\]

Example 2.5.3 Find the largest value of \(-3x^2 + 5x - 6\) and the value of \(x\) at which this largest value occurs.

Do this one the same way. First factor out the \(-3\). This yields
\[
-3 \left( x^2 - \frac{5}{3}x + 2 \right)
\]
Now do the completing the square trick. This equals
\[
-3 \left( x^2 - \frac{5}{3}x + \frac{25}{36} \right) = -3 \left( x - \frac{5}{6} \right)^2 + \frac{25}{36} + 2
\]
which shows that the largest value of \(-3x^2 + 5x - 6\) equals \(-\frac{47}{12}\) and it occurs when \(x = \frac{5}{6}\). If \(x\) is anything else, \(-3x^2 + 5x - 6\) would equal something negative added to \(-\frac{47}{12}\) which would give something smaller.

These examples illustrate the following procedure known as completing the square.

Procedure 2.5.4 Suppose \(f(x) = ax^2 + bx + c\) where \(a \neq 0\). To complete the square, first write as
\[
a \left( x^2 + \frac{b}{a}x + \frac{c}{a} \right)
\]
Next take half of \(b/a\) and square what you get and add it in and subtract it off as follows.
\[
= a \left( x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} - \frac{b^2}{4a^2} + \frac{c}{a} \right)
\]
Now observe this is of the form
\[
= a \left( \left( x + \frac{b}{2a} \right)^2 - \frac{b^2}{4a^2} + \frac{c}{a} \right)
\]
This completes the process of completing the square. Now here is what you can conclude about the function. If \(a > 0\) then the function has its smallest value when \(x = -b/(2a)\) and this smallest value is
\[
\frac{4ac - b^2}{4a^2}
\]
while if \(a < 0\) then the function has its largest value when \(x = -b/(2a)\) and this value is still given as \(2.4\).
Example 2.5.5 Two numbers add to 8. Find the two numbers if their product is to be a large as possible.

Let the numbers be \( x \) and \( y \). Thus \( x + y = 8 \) and \( y = 8 - x \). Thus it is desired to maximize \( x(8 - x) = -x^2 + 8x \). Doing the completing the square procedure, this equals

\[
-1(x^2 - 8x) = -1(x^2 - 8x + 16 - 16) = -1(x - 4)^2 + 16
\]

Hence the maximum possible value is 16 and it occurs when \( x = 4 \) and \( y = 4 \).

2.5.2 Solving Quadratic Equations

The process of completing the square can also be used to establish an important formula known as the quadratic formula. This wonderful formula will give the solution to any quadratic equation which is one of the form

\[
a x^2 + bx + c = 0, \quad a \neq 0.
\]

I will now derive this formula and then it will be used to solve some specific quadratic equations. First divide by \( a \). Then the equation becomes

\[
x^2 + \frac{b}{a}x + \frac{c}{a} = 0
\]

Next use the completing the square procedure.

\[
x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} - \frac{b^2}{4a^2} + \frac{c}{a} = 0
\]

Thus, taking the last two constant terms to the right side of the equal sign,

\[
\left(x + \frac{b}{2a}\right)^2 = \frac{b^2}{4a^2} - \frac{c}{a} = \frac{b^2 - 4ac}{4a^2}
\]

It follows

\[
x + \frac{b}{2a} = \pm \sqrt{\frac{b^2 - 4ac}{4a^2}} = \pm \frac{\sqrt{b^2 - 4ac}}{2a}
\]

The symbol \( \pm \) means plus or minus. Thus there are two numbers being listed here,

\[
\sqrt{\frac{b^2 - 4ac}{4a^2}}, \quad \text{and} \quad -\sqrt{\frac{b^2 - 4ac}{4a^2}}
\]

Then subtracting the \( b/2a \),

\[
x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}
\]

This is the quadratic formula.\(^1\)

\(^1\)The ancient Babylonians knew how to solve quadratic equations sometime before 1700 B.C. They seem to have used the method of completing the square by about 400 B.C. The Chinese and Indians also studied these equations very early.
**Theorem 2.5.6** Let $a \neq 0$. Then the solutions to the quadratic equation
\[ ax^2 + bx + c = 0 \]
are given by the formula,
\[ x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \]

The expression $b^2 - 4ac$ is called the discriminant. If it is negative, there are no real solutions to the equation. When it is positive, there are exactly two and when it equals 0, there is only one solution. If this discriminant is negative, there are still two complex solutions and the formula gives the right answer. However, complex numbers have not been discussed yet.

**Example 2.5.7** Find the solutions to $x^2 + 5x - 1 = 0$.

By the quadratic formula these are\[ x = \frac{-5 \pm \sqrt{25 + 4}}{2} = -5 + \frac{\sqrt{29}}{2}, -5 - \frac{\sqrt{29}}{2} \]

With the quadratic formula, you can factor any quadratic function.

**Example 2.5.8** Factor $x^2 + 5x - 1$.

This is easy because I have the solutions to the equation $x^2 + 5x - 1 = 0$. Thus this factors as\[ x^2 + 5x - 1 = \left( x - \frac{-5 + \sqrt{29}}{2} \right) \left( x - \frac{-5 - \sqrt{29}}{2} \right) \]

Here is why this works. Using the distributive law, the above product equals
\[
x \left( x - \frac{-5 - \sqrt{29}}{2} \right) - \left( -5 + \frac{\sqrt{29}}{2} \right) \left( x - \frac{-5 - \sqrt{29}}{2} \right)
= x^2 - x \left( -5 - \frac{\sqrt{29}}{2} \right) - x \left( -5 + \frac{\sqrt{29}}{2} \right) + \left( -5 + \frac{\sqrt{29}}{2} \right) \left( -5 - \frac{\sqrt{29}}{2} \right)
\]

Now \[ \left( -5 + \frac{\sqrt{29}}{2} \right) \left( -5 - \frac{\sqrt{29}}{2} \right) = 25 + 5\sqrt{29} - 5\sqrt{29} - 29 = -4 \]
and so the above expression reduces to\[ x^2 + 5x - \frac{4}{4} = x^2 + 5x - 1 \]

**Example 2.5.9** Factor $2x^2 + 3x - 2$.

The solutions to $2x^2 + 3x - 2 = 0$ are\[ x = \frac{-3 \pm \sqrt{9 + 16}}{4} = -\frac{3 \pm 5}{4} = -2, \frac{1}{2}. \]

Divide the original polynomial by 2. This yields\[ x^2 + \frac{3}{2}x - 1 \]
which is easy to factor. It is just
\[(x + 2) \left( x - \frac{1}{2} \right)\]
Thus the original polynomial is just 2 times this.
\[(x + 2) (2x - 1).\]
This illustrates the following procedure which always works to factor any quadratic polynomial.

**Procedure 2.5.10** To factor the quadratic polynomial
\[ax^2 + bx + c,\]
it is always
\[a \left( x - \frac{-b - \sqrt{b^2 - 4ac}}{2a} \right) \left( x - \frac{-b + \sqrt{b^2 - 4ac}}{2a} \right)\]

### 2.6 Exercises

1. Suppose you have 1200 feet of fencing and you want to enclose a rectangular garden. What are the dimensions of the largest such garden which can be enclosed?

2. In the context of the above problem, suppose you can use a straight river as one of the sides of the garden. Then what are the dimensions of the largest garden plot?

3. Find the rectangle of largest area which can be drawn inside a circle of radius \(r\). **Hint:** If you are trying to maximize a positive function \(f(x)\), it suffices to maximize \(f^2(x)\).

4. A projectile’s height is given in feet by \(64t - 16t^2 + 10\) where \(t\) is given in seconds. Find the time at which the projectile reaches its maximum altitude. Also find how high it is at this time.

5. Find the point on the line \(y = 2x + 1\) which is closest to the point \((1, 2)\).

6. Suppose \(f(x) = 3x^2 + 7x - 17\). Find the value of \(x\) at which \(f(x)\) is smallest. Also determine \(f(\mathbb{R})\) and sketch the graph of \(f\).

7. Suppose \(f(x) = -5x^2 + 8x - 7\). Find \(f(\mathbb{R})\). In particular, find the largest value of \(f(x)\) and the value of \(x\) at which it occurs.

8. Show the maximum value of \(x (x + b)\) always occurs when \(x = -b/2\).

9. Using the previous problem, show the maximum (minimum) of \(ax^2 + bx + c, a \neq 0\), always occurs when \(x = -b / (2a)\).

10. A function of the two variables \(x, y\) is of the form
\[2x^2 + 3y^2 - 2xy + 3y = f(x, y)\]
Show that \(f(x, y)\) has a minimum value and find it along with the appropriate value of \(x\) and \(y\) which cause the minimum to be achieved. **Hint:** Follow the completing the square procedure for \(2x^2 - 2xy\) treating \(y\) as a constant. Next complete the square on the expression involving the \(y\).
11. Suppose
\[ f(x, y) = ax^2 + by^2 + 2cxy \]
and suppose \( a + b > 0 \) and \( ab - c^2 > 0 \). Show that under these conditions, the above function is nonnegative. This is a function of two variables. You can pick \( x, y \) to be any numbers and the above expression under the given conditions is always nonnegative.

12. Start with
\[ x = \frac{-b \pm \sqrt{b^2 - 4ca}}{2a} \]
and show directly that \( x \) given by either of the two numbers in the above formula satisfies the quadratic equation \( ax^2 + bx + c = 0 \).

13. Verify the assertion of Procedure 2.5.10. That is, multiply the polynomials and show the given product really does equal the polynomial \( ax^2 + bx + c \).

14. Suppose there are two real solutions to \( ax^2 + bx + c, x_1, x_2 \). Show the maximum (minimum), depending on sign of \( a \), of \( f(x) = ax^2 + bx + c \) occurs at \( (x_1 + x_2)/2 \).

15. Find all real values of \( x \) which satisfy the inequality
\[ x^2 + 3x - 1 \leq 0. \]

16. Find all real values of \( x \) which satisfies the inequality
\[ x^2 + 2x + 2 \geq 0. \]

17. Find all solutions of the inequality
\[ \frac{2x^2 + 5x + 3}{x^2 + 2x + 1} \leq 1 \]

18. Find the domain of the function
\[ f(x) = \sqrt{x^2 - 3x + 2}. \]

19. Solve the inequality
\[ \frac{(x + 1)(2x + 1)}{(x - 1)} \geq 1. \]

20. Solve if possible, \( \frac{1}{x+2} + \frac{2}{x^2-3} = 0 \).

21. Solve the inequality if possible
\[ \frac{1}{x+2} + \frac{2}{x-3} \leq 1 \]

22. Find all solutions of
\[ |x + 1| = |x^2 - 4| \]
2.7 Asymptotes

These are straight lines with the property that the line closely resembles the graph of the function either for large $x$ or for $x$ close to some number. Here is an example which illustrates these ideas. It is the graph of the function $y = \frac{2x + 1}{x - 1}$.

The graph of the function is the curved part. The vertical line is a vertical asymptote for this function. Note how the graph of the function looks like the graph of the vertical line $x = 1$ as $x$ gets close to 1. As $|x|$ becomes large the graph of the function resembles the graph of the line $y = 2$. Consider why these things are so.

As $x$ gets close to 1 but is larger than 1, the top (numerator) of the rational function

$$f(x) = \frac{2x + 1}{x - 1}$$

is close to 3 while the bottom (denominator) is a very small positive number. Thus the quotient must be a large positive number. This explains why for $x - 1$ small and positive the value of the function is large and positive. For $x - 1$ small and negative the top is still close to 3 while the bottom is now a small negative number which requires $f(x)$ to be large and negative. This explains the behavior of the graph, how it “approaches” the line $x = 1$.

Now consider the case that $|x|$ is large. Divide the top and the bottom of the rational function by $x$.

$$f(x) = \frac{2x + 1}{x - 1} = \frac{2 + \frac{1}{x}}{1 - \frac{1}{x}}$$

For $|x|$ large, the expressions $2/x$ and $1/x$ are very small and so the function’s value ought to be close to 2. More precisely,

$$f(x) - 2 = \frac{2 + \frac{1}{x}}{1 - \frac{1}{x}} - 2 = \frac{3}{x - 1}$$

and for $|x|$ very large, this becomes very small. Thus for large $|x|$, the function is close to 2 and this explains the behavior observed in the graph which shows this happening. The explanation of the algebra in the last step is as follows.

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

This illustrates two kinds of asymptotes, the vertical and the horizontal asymptote.
The following is a procedure for finding horizontal and vertical asymptotes for rational functions (quotients of polynomials).

**Procedure 2.7.1** The line \( x = c \) is a vertical asymptote for the rational function \( \frac{p(x)}{q(x)} \) if \( q(c) = 0 \) and \( p(c) \neq 0 \). If

\[
p(x) \equiv a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0
\]

and

\[
q(x) \equiv b_n x^n + b_{n-1} x^{n-1} + \cdots + b_1 x + b_0
\]

where \( b_n \neq 0 \) then

\[
\frac{p(x)}{q(x)}
\]

has the horizontal asymptote \( y = \frac{a_n}{b_n} \).

The reason this works for the horizontal asymptote is that

\[
\frac{p(x)}{q(x)} = \frac{x^{-n} \left( a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \right)}{x^{-n} \left( b_n x^n + b_{n-1} x^{n-1} + \cdots + b_1 x + b_0 \right)}
\]

\[
= \frac{a_n + \frac{a_{n-1}}{x} + \cdots + \frac{1}{x^{n-1}} a_1 + \frac{1}{x^n} a_0}{b_n + \frac{b_{n-1}}{x} + \cdots + \frac{1}{x^{n-1}} b_1 + \frac{1}{x^n} b_0}
\]

Now when \( |x| \) is large, all those terms in the numerator and denominator which consist of a constant divided by a power of \( x \) must be very small and so the result must be close to \( \frac{a_n}{b_n} \).

**Example 2.7.2** Find horizontal and vertical asymptotes for the function

\[
f(x) = \frac{x^2}{(x + 1)^2}
\]

By the procedure, it follows \( y = 1 \) is a horizontal asymptote while \( x = -1 \) is a vertical asymptote. Here is the graph of this function with the horizontal and vertical asymptotes shown.

**Example 2.7.3** Find the asymptotes of

\[
f(x) = \frac{x^3}{(1 + x^2)^2}
\]
In this case there are no vertical asymptotes because the denominator is never equal to 0. However, there is a horizontal asymptote \( y = 0 \) because the rational function is of the form
\[
\frac{0x^4 + x^3 + 0x^2 + 0x + 0}{x^4 + 2x^2 + 1}
\]
Here is the graph of this function.

**Example 2.7.4** Find the asymptotes of

\[
f(x) = \frac{(x - 1)^2}{(x^2 - 1)}
\]

Candidates for the vertical asymptotes are \( x = 1, -1 \). It is not clear whether \( x = 1 \) is an asymptote. This is because both the numerator and the denominator equal 0 at \( x = 1 \). However, \( x = -1 \) is a vertical asymptote because the numerator is positive there and the denominator vanishes there. In fact, \( x = 1 \) is not an asymptote. This is because for \( x \) near but not equal to 1, the function reduces to

\[
f(x) = \frac{x - 1}{x + 1}
\]

The only thing which happens at \( x = 1 \) is that the original function is not defined. Here is the graph of this function which results from doing the indicated cancelation. The graph of the original function is just like this except there should be a hole in the graph at the point where \( x = 1 \).

**2.8 Exercises**

1. Find the horizontal and vertical asymptotes of the rational function

\[
f(x) = \frac{2x^2 + 1}{(x + 2)(x - 1)}
\]

Also sketch its graph.
2. Find the horizontal and vertical asymptotes of the rational function

\[ f(x) = \frac{x^2 + 1}{(x^2 + 2)(x - 1)} \]

Also sketch its graph.

3. Find the horizontal and vertical asymptotes of the rational function

\[ f(x) = \frac{x^2 - 1}{(x + 2)(x - 1)} \]

Also sketch its graph.

4. Find the horizontal and vertical asymptotes of the rational function

\[ f(x) = \frac{x^2 + 3x + 2}{(x + 2)(x + 1)} \]

Also sketch its graph.

5. Find the horizontal and vertical asymptotes of the rational function

\[ f(x) = \frac{x^2 + 3x + 2}{(x + 2)(x^2 + 1)} \]

Also sketch its graph.

6. Find the horizontal and vertical asymptotes of the rational function

\[ f(x) = \frac{x^3 + 3x + 2}{(x + 2)x^2} \]

Also sketch its graph.

7. Here is a rational function.

\[ f(x) = \frac{x^3 + x^2}{x^2 + 2x + 1} \]

Find its vertical asymptotes if any. Show that the function has no horizontal asymptotes but that for \(|x|\) large the function is approximately equal to \(x + 1\). The line \(y = x + 1\) is sometimes called an oblique asymptote.
Division

3.1 Division And Numbers

First recall Theorem 1.9.11, the Euclidean algorithm.

**Theorem 3.1.1** Suppose $0 < a$ and let $b \geq 0$. Then there exists a unique integer $p$ and real number $r$ such that $0 \leq r < a$ and $b = pa + r$.

The following definition describes what is meant by a prime number and also what is meant by the word “divides”.

**Definition 3.1.2** The number, $a$ divides the number, $b$ if in Theorem 1.9.11, $r = 0$. That is there is zero remainder. The notation for this is $a|b$, read $a$ divides $b$ and $a$ is called a factor of $b$. A prime number is one which has the property that the only numbers which divide it are itself and 1.

Two integers are relatively prime if their greatest common divisor is one. The greatest common divisor of $m$ and $n$ is denoted as $(m, n)$.

There is a phenomenal and amazing theorem which relates the greatest common divisor to the smallest number in a certain set. Suppose $m, n$ are two positive integers. Then if $x, y$ are integers, so is $xm + yn$. Consider all integers which are of this form. Some are positive such as $1m + 1n$ and some are not. The set $S$ in the following theorem consists of exactly those integers of this form which are positive. Then the greatest common divisor of $m$ and $n$ will be the smallest number in $S$. This is what the following theorem says.

**Theorem 3.1.3** Let $m, n$ be two positive integers and define

$$S \equiv \{xm + yn \in \mathbb{N} : x, y \in \mathbb{Z} \}.$$

Then the smallest number in $S$ is the greatest common divisor, denoted by $(m, n)$.

**Proof:** First note that both $m$ and $n$ are in $S$ so it is a nonempty set of positive integers. By well ordering, there is a smallest element of $S$, called $p = x_0m + y_0n$. Either $p$ divides $m$ or it does not. If $p$ does not divide $m$, then by Theorem 1.9.11

$$m = pq + r$$

where $0 < r < p$. Thus $m = (x_0m + y_0n)q + r$ and so, solving for $r$,

$$r = m(1 - x_0) + (-y_0q) n \in S.$$
However, this is a contradiction because $p$ was the smallest element of $S$. Thus $p|m$. Similarly $p|n$.

Now suppose $q$ divides both $m$ and $n$. Then $m = qx$ and $n = qy$ for integers, $x$ and $y$. Therefore,

$$p = mx_0 + ny_0 = x_0qx + y_0qy = q(x_0x + y_0y)$$

showing $q|p$. Therefore, $p = (m, n)$. This proves the theorem.

There is a relatively simple algorithm for finding $(m, n)$ which will be discussed now. Suppose $0 < m < n$ where $m, n$ are integers. Also suppose the greatest common divisor is $(m, n) = d$. Then by the Euclidean algorithm, there exist integers $q, r$ such that

$$n = qm + r, \ r < m$$

(3.1)

Now $d$ divides $n$ and $m$ so there are numbers $k, l$ such that $dk = m, dl = n$. From the above equation,

$$r = n - qm = dl - qdk = d(l - qk)$$

Thus $d$ divides both $m$ and $r$. If $k$ divides both $m$ and $r$, then from the equation of (3.1) it follows $k$ also divides $n$. Therefore, $k$ divides $d$ by the definition of the greatest common divisor. Thus $d$ is the greatest common divisor of $m$ and $n$ but $m + r < m + n$. This yields another pair of positive integers for which $d$ is still the greatest common divisor but the sum of these integers is strictly smaller than the sum of the first two. Now you can do the same thing to these integers. Eventually the process must end because the sum gets strictly smaller each time it is done. It ends when there are not two positive integers produced. That is, one is a multiple of the other. At this point, the greatest common divisor is the smaller of the two numbers.

**Procedure 3.1.4** To find the greatest common divisor of $m, n$ where $0 < m < n$, replace the pair \{(m, n)\} with \{(m, r)\} where $n = qm + r$ for $r < m$. This new pair of numbers has the same greatest common divisor. Do the process to this pair and continue doing this till you obtain a pair of numbers where one is a multiple of the other. Then the smaller is the sought for greatest common divisor.

**Example 3.1.5** Find the greatest common divisor of 165 and 385.

Use the Euclidean algorithm to write

$$385 = 2(165) + 55$$

Thus the next two numbers are 55 and 165. Then

$$165 = 3 \times 55$$

and so the greatest common divisor of the first two numbers is 55.

**Example 3.1.6** Find the greatest common divisor of 1237 and 4322.

Use the Euclidean algorithm

$$4322 = 3(1237) + 611$$

Now the two new numbers are 1237, 611. Then

$$1237 = 2(611) + 15$$

The two new numbers are 15, 611. Then

$$611 = 40(15) + 11$$
The two new numbers are 15, 11. Then
\[ 15 = 1(11) + 4 \]
The two new numbers are 11, 4
\[ 2(4) + 3 \]
The two new numbers are 4, 3. Then
\[ 4 = 1(3) + 1 \]
The two new numbers are 3, 1. Then
\[ 3 = 3 \times 1 \]
and so 1 is the greatest common divisor. Of course you could see this right away when the two new numbers were 15 and 11. Recall the process delivers numbers which have the same greatest common divisor.

This amazing theorem will now be used to prove a fundamental property of prime numbers which leads to the fundamental theorem of arithmetic, the major theorem which says every integer can be factored as a product of primes.

**Theorem 3.1.7** If \( p \) is a prime and \( p \mid ab \) then either \( p \mid a \) or \( p \mid b \).

**Proof:** Suppose \( p \) does not divide \( a \). Then since \( p \) is prime, the only factors of \( p \) are 1 and \( p \) so follows \((p, a) = 1\) and therefore, there exists integers, \( x \) and \( y \) such that
\[ 1 = ax + yp. \]
Multiplying this equation by \( b \) yields
\[ b = abx + ybp. \]
Since \( p \mid ab \), \( ab = pz \) for some integer \( z \). Therefore,
\[ b = abx + ybp = pxz + ybp = p(xz + yb) \]
and this shows \( p \) divides \( b \).

**Theorem 3.1.8** (Fundamental theorem of arithmetic) Let \( a \in \mathbb{N} \setminus \{1\} \). Then \( a = \prod_{i=1}^{n} p_i \) where \( p_i \) are all prime numbers. Furthermore, this prime factorization is unique except for order of the factors.

**Proof:** If \( a \) equals a prime number, the prime factorization clearly exists. In particular the prime factorization exists for the prime number 2. Assume this theorem is true for all \( a \leq n - 1 \). If \( n \) is a prime, then it has a prime factorization. On the other hand, if \( n \) is not a prime, then there exist two integers \( k \) and \( m \) such that \( n = km \) where each of \( k \) and \( m \) are less than \( n \). Therefore, each of these is no larger than \( n - 1 \) and consequently, each has a prime factorization. Thus so does \( n \). It remains to argue the prime factorization is unique except for order of the factors.

Suppose
\[ \prod_{i=1}^{n} p_i = \prod_{j=1}^{m} q_j \]
where the \( p_i \) and \( q_j \) are all prime, there is no way to reorder the \( q_k \) such that \( m = n \) and \( p_i = q_i \) for all \( i \) and \( n + m \) is the smallest positive integer such that this happens. Then by Theorem 3.1.7, \( p_i \mid q_j \) for some \( j \). Since these are prime numbers this requires
If $p_1 = q_j$. Reordering if necessary it can be assumed that $q_j = q_1$. Then dividing both sides by $p_1 = q_1$,

$$\prod_{i=1}^{n-1} p_{i+1} = \prod_{j=1}^{m-1} q_{j+1}.$$  

Since $n + m$ was as small as possible for the theorem to fail, it follows that $n-1 = m-1$ and the prime numbers, $q_2, \cdots, q_m$ can be reordered in such a way that $p_k = q_k$ for all $k = 2, \cdots, n$. Hence $p_i = q_i$ for all $i$ because it was already argued that $p_1 = q_1$, and this results in a contradiction, proving the theorem.

### 3.2 Exercises

1. Euclid\(^{1}\) showed there were infinitely many prime numbers using a very simple argument. He assumed there were only finitely many, \{\(p_1, \cdots, p_n\}\} and then considered the number $p_1 \cdots p_n + 1$ consisting of the product of all the primes plus 1. Then this number can’t be prime because it is larger than every prime number. Therefore, some prime number, $p_k$ from the above list must divide it. Now obtain a terrible contradiction.

2. Explain in your own words why ($p, n$) = 1 whenever $p$ is a prime number.

3. Find the greatest common divisor of 231 and 457.

4. Find the greatest common divisor of 3795 and 1155.

5. Provide an algorithm which will find the greatest common divisor of three positive integers. Use to find the greatest common divisor of 3003, 4485, 1443.

6. If \(a, b\) are integers, \([a, b]\) will denote their least common multiple. This is the smallest number which has both \(a\) and \(b\) as factors. Show \([a, b] = ab/(a, b)\). \textbf{Hint:}\ Show \([a, b]\) must divide \(ab\). Here is how you might proceed. If not, \(ab = [a, b]q + r\) where 0 < $r < [a, b]$. Then verify \(r\) is a common multiple of \(a\) and \(b\) contradicting that \([a, b]\) is the \textbf{least} common multiple. Hence \(r = 0\). Therefore, \([a, b] = ab/q\) for some \(q\) an integer. Since \([a, b]\) is a common multiple of \(a\) and \(b\), argue that \(q\) must divide both \(a\) and \(b\). Now what is the largest such \(q\)? This would yield the smallest \(ab/q\). You fill in the details.

7. Suppose \((m, n) = 1\). Show there exist integers \(x, y\) such that \(xm + yn = 17\).

8. Let \(n\) be any positive integer and let \(p\) be any nonnegative integer. Explain why at least one of the integers $p + 1, \cdots, p + n$ must be a multiple of $n$. Is this also true for $p$ any integer? \textbf{Hint:}\ You might consider using induction starting with $p = 0$.

9. Let $a_1, \cdots, a_p$ be positive integers. Show there exist integers $x_i$ such that the greatest common divisor of these integers $m$ may be written as

$$m = \sum_{i=1}^{p} x_i a_i$$

The greatest common divisor divides each of the $a_i$ and if $l$ also divides each $a_i$, then $l$ divides $m$.

\(^1\)He lived about 300 B.C.
3.3. RATIONAL ROOT THEOREM

10. Let \( a_n = 2^{2n} + 1 \) for \( n = 1, 2, \cdots \). Show that if \( n \neq m \), then \( a_n \) and \( a_m \) are relatively prime. Either \( a_n \) is prime or it is not. If it is not, then all the numbers dividing it other than 1 fail to divide \( a_m \) for all \( m < n \). Explain why this shows there must be infinitely many primes. This argument about infinitely many primes is due to Polya. It gives more information than the argument of Euclid. The numbers, \( 2^{(2^n)} + 1 \) are prime numbers for several values of \( n \) but Euler\(^2\) showed that when \( n = 5 \), the number is not prime\(^3\). When numbers of this form are prime, they are called Fermat\(^4\) primes. At this time it is unknown whether there are infinitely many Fermat primes. For more information on these matters, you should see the book by Chahal,\(^4\).

Hint: To verify \( a_n \) and \( a_m \) are relatively prime for \( m > n \), suppose they are not and that for some number, \( p \neq 1 \), \( a_n = pk_1 \) while \( a_m = pk_2 \). Then letting \( m = n + r \), explain why

\[
pk_2 = a_m = (2^{2n})^{2^r} + 1 = (pk_1 - 1)^{2^r} + 1 = p \text{ (integer)} + 2.
\]

Consequently, \( p \text{ (integer)} = 2 \). What does this say about \( p \)? How does \( pk_1 = 2^{2n} + 1 \) yield a contradiction?

3.3 Rational Root Theorem

The next theorem is a very nice high school theorem which characterizes all possible rational zeros for polynomials having integer coefficients. A polynomial is something of the form

\[ a_n x^n + \cdots + a_1 x + a_0 \]

where the \( a_k \) are numbers in some field and \( x \) is called a variable. Often people want to find values of \( x \) such that

\[ a_n x^n + \cdots + a_1 x + a_0 = 0. \]

These values of \( x \) are called roots or zeros of the polynomial. For example, find \( x \) such that \( x^2 + 5x - 7 = 0 \). In general, this is a very hard problem and cannot be solved using algebra but sometimes you can find a rational solution to such an equation. The rational root theorem gives a systematic way to look for these rational roots.

Theorem 3.3.1 (rational root theorem) Let

\[ a_n x^n + \cdots + a_1 x + a_0 = 0 \]

where each \( a_i \) is an integer and \( a_n, a_0 \neq 0 \). Then \( \text{IF} \) the equation has any rational solutions, then they are of the form

\[ \pm \frac{\text{factor of } a_0}{\text{factor of } a_n}. \]

---

\(^2\)Leonhard Euler, born in Switzerland, lived from 1707 to 1783. He was the most prolific mathematician ever to live. He made major contributions to number theory, analysis, algebra, mechanics, and differential equations. He and Lagrange invented the branch of mathematics known as calculus of variations. His collected papers take up more shelf space than a typical encyclopedia. His memory was prodigious and he could do unbelievable feats of computation in his head. He had 13 children.

\(^3\)The number in this case is 4, 294,967,297.

\(^4\)Fermat lived from 1601 to 1665. He is generally regarded as the founder of number theory. His most famous conjecture was that there is no solution to the equation \( x^n + y^n = z^n \) if \( n \geq 3 \). That is there is no analog to pythagorean triples with higher exponents than 2. This was finally proved in the 1990's by Andrew Wiles.
Proof: Let \( \frac{p}{q} \) be a rational solution. Dividing \( p \) and \( q \) by \((p, q)\) if necessary, the fraction may be reduced to lowest terms such that \((p, q) = 1\). Substituting into the equation,

\[ a_n \left( \frac{p}{q} \right)^n + \cdots + a_1 \left( \frac{p}{q} \right) + a_0 = 0 \]

Now multiply both sides by \( q^n \). Thus

\[ a_n p^n + a_{n-1} p^{n-1} q + \cdots + a_1 p q^{n-1} + a_0 q^n = 0 \]

Hence

\[ a_n p^n = -(a_{n-1} p^{n-1} q + \cdots + a_0 q^n) \]

and \( q \) divides the right side of the equation and therefore, \( q \) must divide the left side also. However, \((q, p^n) = 1\) and so by Theorem 3.1.7 \( q | a_n \) because it does not divide \( p^n \) due to the fact that \( p^n \) and \( q \) have no prime factors in common.

Similarly,

\[ a_0 q^n = -(a_n p^n + \cdots + a_1 q^{n-1}) \]

and so \( p | a_0 q^n \) but \((p, q^n) = 1\). By Theorem 3.1.7 again, \( p | a_0 \) and this proves the theorem.

Example 3.3.2 An irrational number is one which is not rational. Show \( \sqrt{2} \) is irrational if it exists.

\( \sqrt{2} \) is the solution of the equation \( x^2 - 2 = 0 \). However, from Theorem 3.3.1 the only possible rational zeros to this equation are \( \pm 2 \) and \( \pm 1 \) and none of these work. Therefore, \( \sqrt{2} \) must be irrational.

3.4 Division And Polynomials

That which you can do for integers often can be modified and done to polynomials because polynomials behave a lot like integers. You add and multiply polynomials using the distributive law and with the understanding that the variable represents a number and so it is treated as one. Thus

\[ (2x^2 + 5x - 7) + (3x^3 + x^2 + 6) = 3x^2 + 5x - 1 + 3x^3 \]

and

\[ (2x^2 + 5x - 7) (3x^3 + x^2 + 6) = 6x^5 + 17x^4 + 5x^2 - 16x^3 + 30x - 42 \]

The second assertion is established as follows. From the distributive law,

\[ (2x^2 + 5x - 7) (3x^3 + x^2 + 6) \]

\[ = (2x^2 + 5x - 7)3x^3 + (2x^2 + 5x - 7)x^2 + (2x^2 + 5x - 7)6 \]

\[ = 2x^2 (3x^3) + 5x (3x^3) - 7 (3x^3) + 2x^2 (x^2) + 5x (x^2) - 7 (x^2) + 12x^2 + 30x - 42 \]

which simplifies to the claimed result. Note that \( x^2 x^3 = x^5 \) because the left side simply says to multiply \( x \) by itself 5 times. Other axioms satisfied by the integers are also satisfied by polynomials and like integers, polynomials typically don’t have multiplicative inverses which are polynomials. In this section the polynomials have coefficients which come from a field. This field is usually \( \mathbb{R} \) in calculus but it doesn’t have to be.

First is the Euclidean algorithm for polynomials. This is a lot like the Euclidean algorithm for numbers, Theorem 1.9.11. First here is the definition of the degree of a polynomial.
Definition 3.4.1 Let $a_n x^n + \cdots + a_1 x + a_0$ be a polynomial. The degree of this polynomial is $n$ if $a_n \neq 0$. The degree of a polynomial is the largest exponent on $x$ provided the polynomial does not have all the $a_i = 0$. If each $a_i = 0$, we don’t speak of the degree because it is not defined. In writing this, it is only assumed that the coefficients $a_i$ are in some field such as the real numbers or the rational numbers.

Theorem 3.4.2 Let $f(x)$ and $g(x)$ be polynomials with coefficients in a some field. Then there exists a polynomial, $q(x)$ such that

$$f(x) = q(x) g(x) + r(x)$$

where the degree of $r(x)$ is less than the degree of $g(x)$ or $r(x) = 0$. All these polynomials have coefficients in the same field.

Proof: Consider the polynomials of the form $f(x) - g(x) l(x)$ and out of all these polynomials, pick one which has the smallest degree. This can be done because of the well ordering of the natural numbers. Let this take place when $l(x) = q_1(x)$ and let

$$r(x) = f(x) - g(x) q_1(x).$$

It is required to show degree of $r(x) < \text{degree of } g(x)$ or else $r(x) = 0$.

Suppose $f(x) - g(x) l(x)$ is never equal to zero for any $l(x)$. Then $r(x) \neq 0$. It is required to show the degree of $r(x)$ is smaller than the degree of $g(x)$. If this doesn’t happen, then the degree of $r \geq$ the degree of $g$. Let

$$r(x) = b_m x^m + \cdots + b_1 x + b_0$$
$$g(x) = a_n x^n + \cdots + a_1 x + a_0$$

where $m \geq n$ and $b_m$ and $a_n$ are nonzero. Then let $r_1(x)$ be given by

$$r_1(x) = r(x) - \frac{x^{m-n} b_m}{a_n} g(x)$$

$$= (b_m x^m + \cdots + b_1 x + b_0) - \frac{x^{m-n} b_m}{a_n} (a_n x^n + \cdots + a_1 x + a_0)$$

which has smaller degree smaller than $m$, the degree of $r(x)$. But

$$r_1(x) = \frac{r(x)}{f(x) - g(x) q_1(x)} - \frac{x^{m-n} b_m}{a_n} g(x)$$

$$= \frac{r(x)}{f(x) - g(x) q_1(x)} - \frac{x^{m-n} b_m}{a_n},$$

and this is not zero by the assumption that $f(x) - g(x) l(x)$ is never equal to zero for any $l(x)$ yet has smaller degree than $r(x)$ which is a contradiction to the choice of $r(x)$. This proves the Theorem.

Now with this theorem, here is another one which is very fundamental. It is like Theorem 3.1.3. First here is a definition. A polynomial is monic means it is of the form

$$x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0.$$ 

That is, the leading coefficient, that which goes with the highest power of $x$ is 1.
Definition 3.4.3 A polynomial $f$ is said to divide a polynomial $g$ if $g(x) = f(x)\ r(x)$ for some polynomial $r(x)$. Let $\{\phi_i(x)\}$ be a finite set of polynomials. The greatest common divisor will be the monic polynomial $q$ such that $q(x)$ divides each $\phi_i(x)$ and if $p(x)$ divides each $\phi_i(x)$, then $p(x)$ divides $q(x)$. The finite set of polynomials $\{\phi_i\}$ is said to be relatively prime if their greatest common divisor is 1. A polynomial $f(x)$ is irreducible if there is no polynomial with coefficients in $\mathbb{F}$, the field from which the coefficients are taken, which divides it except nonzero multiples of $f(x)$ and constants.

Theorem 3.4.4 Let $\psi(x)$ be the greatest common divisor of $\{\phi_i(x)\}$. Then there exist polynomials $r_i(x)$ such that

$$\psi(x) = \sum_{i=1}^{p} r_i(x) \phi_i(x).$$

Furthermore, $\psi(x)$ is the monic polynomial of smallest degree which can be written in the above form.

Proof: Let $S$ denote the set of monic polynomials which are of the form

$$\sum_{i=1}^{p} r_i(x) \phi_i(x)$$

where $r_i(x)$ is a polynomial. Then let the $r_i$ be chosen such that the degree of the expression $\sum_{i=1}^{p} r_i(x) \phi_i(x)$ is as small as possible. Letting $\psi(x)$ equal this sum, it remains to verify it is the greatest common divisor. First, does it divide each $\phi_i(x)$? Suppose it fails to divide $\phi_1(x)$. Then by Lemma 3.4.2

$$\phi_1(x) = \psi(x) l(x) + r(x)$$

where degree of $r(x)$ is less than that of $\psi(x)$. Then

$$r(x) = \phi_1(x) - \psi(x) l(x).$$

Divide by the leading coefficient if necessary and denoting the result on the left by $\psi_1(x)$, it follows the degree of $\psi_1(x)$ is less than the degree of $\psi(x)$ and $\psi_1(x)$ is of the form

$$\psi_1(x) = (\phi_1(x) - \psi(x) l(x)) a$$

for some number $a$. Then this equals

$$= \left( \phi_1(x) - \sum_{i=1}^{p} r_i(x) \phi_i(x) l(x) \right) a$$

$$= \left( 1 - r_1(x) \right) \phi_1(x) a + \sum_{i=2}^{p} (-r_i(x) l(x)) \phi_i(x) a$$

for a suitable $a \in \mathbb{F}$, the field from which the coefficients come. This is one of the polynomials in $S$. Therefore, $\psi(x)$ does not have the smallest degree after all because the degree of $\psi_1(x)$ is smaller. This is a contradiction. Therefore, $\psi(x)$ divides $\phi_1(x)$. Similar reasoning shows it divides all the other $\phi_i(x)$.

If $p(x)$ divides all the $\phi_i(x)$, then it divides $\psi(x)$ because of the formula for $\psi(x)$ which equals $\sum_{i=1}^{p} r_i(x) \phi_i(x)$. This proves the theorem.

\footnote{Think $\mathbb{R}$ if you wish a specific example. It all works for any field and there are many others besides the usual one. At some point, this is important to realize.}
3.5 Factoring Polynomials

The idea is to start with a polynomial and then to write it as the product of polynomials of smaller degree. In general, you can’t do it unless you are lucky. You can prove the factors exist but finding them is another matter. Therefore, it is ironic to see how often what is taught is made to depend unnecessarily on skill in factoring polynomials. However, this is the way it is. Whole courses are based substantially on one’s technique in factoring polynomials. For example the techniques used for the solution of differential equations in undergraduate differential equations courses are usually based on factoring polynomials although this is not the way such equations are solved in practice. Therefore, to survive you need to acquire some skill at doing this.

Let \( a \) be a number and \( p(x) \) a polynomial. Then by the Euclidean algorithm, Theorem 3.4.2, there exist polynomials \( q(x) \) and \( r(x) \) with the degree of \( r(x) \) < 1 or else \( r(x) = 0 \) such that

\[
p(x) = (x - a)q(x) + r(x)
\]

Thus \( r(x) = r \), a constant. Consider the problem of finding \( r \) as well as the polynomial \( q(x) \). Say

\[
p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0
\]

and

\[
q(x) = b_{n-1} x^{n-1} + b_{n-2} x^{n-2} + \cdots + b_1 x + b_0
\]

You are given what the \( a_k \) are and you want to find the \( b_j \). You have

\[
a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = (x - a) (b_{n-1} x^{n-1} + b_{n-2} x^{n-2} + \cdots + b_1 x + b_0) + r
\]

Comparing the coefficients of \( x^n \) on both sides, you must have \( b_{n-1} = a_n \). Now consider the coefficient of \( x^{n-1} \) on both sides. This must be the same and so

\[
a_{n-1} = -ab_{n-1} + b_{n-2}
\]

and so you must have

\[
b_{n-2} = a_{n-1} + ab_{n-1}
\]

You could have done this with the coefficient of \( x^k \) for any \( k \leq n \) obtaining

\[
a_k = -ab_k + b_{k-1}, \quad b_{k-1} = a_k + ab_k.
\]

Summarizing this, for each \( 1 \leq k < n \),

\[
b_{n-1} = a_n, \quad b_{k-1} = a_k + ab_k
\]

Thus this will give all the \( b_k \). What about the \( r \)? The constant term in both sides of (3.2) must be the same and so

\[
r - ab_0 = a_0
\]

and so

\[
r = ab_0 + a_0,
\]

which follows the same pattern.

A simple algorithm which will compute these \( b_k \) and \( r \) using the above considerations is the method of synthetic division. I will outline it in general terms and then show how to implement it for specific examples. First write the coefficients of the polynomial across the top line, the coefficients of the highest powers of \( x \) on the left and falling toward the lowest powers on the right. Then follow the procedure indicated in the
following, starting at the left and working toward the right. Place $a_n$ in the bottom of
the second column. Then place $aa_{n-1}$ right below the $a_{n-1}$ and add $a_{n-1}$ to
this which you place in the bottom of the third column. You take this thing, which by
the above discussion equals $b_{n-2}$, multiply it by $a$ and place in the fourth column right
under the $a_{n-2}$ and then add to $a_{n-2}$. This gives $b_{n-3}$. Continue doing this till you
get to the end and the thing in the bottom right position is the value of the polynomial
evaluated at $a$.

<table>
<thead>
<tr>
<th></th>
<th>$a_n$</th>
<th>$a_{n-1}$</th>
<th>$ab_{n-1}$</th>
<th>$b_{n-2}$</th>
<th>$ab_{n-2}$</th>
<th>$b_{n-3}$</th>
<th>$ab_{n-3}$</th>
<th>$\cdots$</th>
<th>$a_0$</th>
<th>$ab_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_n = b_{n-1}$</td>
<td>$a_{n-1}$</td>
<td>$ab_{n-1}$</td>
<td>$b_{n-2}$</td>
<td>$ab_{n-2}$</td>
<td>$b_{n-3}$</td>
<td>$ab_{n-3}$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$a_0$</td>
<td>$ab_0$</td>
</tr>
</tbody>
</table>

**Example 3.5.1** Let $p(x) = x^4 - 9x^3 + 12x^2 - 3x + 7$. Find $p(8)$.

Of course you could just plug in 8 and see what happens.

$$(8)^4 - 9(8)^3 + 12(8)^2 - 3(8) + 7 = 239$$

Now lets do it the other way.

<table>
<thead>
<tr>
<th></th>
<th>8</th>
<th>-9</th>
<th>12</th>
<th>-3</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>-8</td>
<td>32</td>
<td>8 x 29</td>
<td></td>
</tr>
</tbody>
</table>

Note that the computations are much easier than doing things like $8^4$. This is an easier
way to evaluate a polynomial at various values than simply plugging it in. Also I have
found additional information.

$$x^4 - 9x^3 + 12x^2 - 3x + 7 = (x^3 - x^2 + 4x + 29)(x - 8) + 239$$

The technique of synthetic division and the rational root theorem can be used to
factor polynomials sometimes. The rational root theorem can be used to identify pos-
sible zeros which are rational numbers. Thus if $a$ is one of those zeros, the $r$ in 3.2
must equal 0. Therefore, the above algorithm will identify a polynomial $q(x)$ such
that $p(x) = (x - a)q(x)$ and you will have made at least a first step in factoring the
polynomial.

**Example 3.5.2** Let $p(x) = 3x^4 + 8x^3 - 18x^2 + 60$. Find $p(-5)$.

This time I will just use synthetic division.

<table>
<thead>
<tr>
<th></th>
<th>-5</th>
<th>3</th>
<th>8</th>
<th>-18</th>
<th>0</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>-7</td>
<td>17</td>
<td>-85</td>
<td>485</td>
<td></td>
</tr>
</tbody>
</table>

and so $p(-5) = 485$. Thus it is also true that

$$p(x) = (x + 5) (3x^3 - 7x^2 + 17x - 85) + 485$$

**Example 3.5.3** Factor the polynomial $p(x) = x^3 - 4x^2 + 5x - 2$. 
3.5. FACTORING POLYNOMIALS

First identify the possible rational roots. These are ±1, ±2. I don’t know which of these are really roots and if I did, I would also need to find the polynomial \( q(x) \) such that

\[
x^3 - 4x^2 + 5x - 2 = (x - a)q(x)
\]

Let’s see if \(-1\) is a root first. Using the above algorithm,

<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
<th>-4</th>
<th>5</th>
<th>-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>5</td>
<td>-10</td>
</tr>
<tr>
<td>1</td>
<td>-5</td>
<td>10</td>
<td>-12</td>
<td></td>
</tr>
</tbody>
</table>

Thus \(-1\) is not a root and \((x + 1)\) is not a factor. In fact, if you plug in \(-1\) to the equation you will get \(-12\).

\[
(-1)^3 - 4(-1)^2 + 5(-1) - 2 = -12
\]

which is what I got using synthetic division. That one didn’t work so let’s look at another possible rational root. Let’s try \(1\)

<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
<th>-4</th>
<th>5</th>
<th>-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-3</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

This worked. Therefore, the above algorithm also gives

\[
p(x) = (x - 1) (x^2 - 3x + 2)
\]

At this point, you could either use the quadratic formula to factor the quadratic polynomial or you could use another application of the rational root theorem. This is what I will do. The possible rational roots are the same so I will try \(1\) again.

<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
<th>-3</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

and so it worked. Thus \(x^2 - 3x + 2 = (x - 1)(x - 2)\) and so

\[
p(x) = (x - 1)^2 (x - 2) .
\]

**Example 3.5.4** Factor the polynomial \( p(x) = 2x^4 + 7x^3 + x^2 - 7x - 3 \) if possible.

The possible rational roots are ±1, ±\(\frac{1}{2}\), ±3, ±\(\frac{1}{4}\). I think I have found all possibilities. Now it is necessary to try them. Let’s try \(-1/2\) first.

<table>
<thead>
<tr>
<th>-1/2</th>
<th>2</th>
<th>7</th>
<th>1</th>
<th>-7</th>
<th>-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>6</td>
<td>-2</td>
<td>-6</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

and so \(-1/2\) is a root. Furthermore, from the above

\[
2x^4 + 7x^3 + x^2 - 7x - 3
\]

\[
= \left(x + \frac{1}{2}\right) \left(2x^3 + 6x^2 - 2x - 6\right)
\]
Now I can use the rational root theorem to identify possible rational roots for the second of the above terms in the product. It looks like the possible ones include the original possibilities and those multiplied by 2. I shall try 1.

\[
\begin{array}{cccc}
1 & 2 & 6 & -2 \\
2 & 8 & 6 & 0 \\
\end{array}
\]

It was a good choice. Thus

\[p(x) = (x + \frac{1}{2}) (x - 1) (2x^2 + 8x + 6)\]

At this point, I will factor out the 2 and write it as follows.

\[p(x) = (2x + 1) (x - 1) (x^2 + 4x + 3)\]

That last one is easy to factor but if not, I could use the quadratic formula to factor it. Thus

\[p(x) = (2x + 1) (x - 1) (x + 3) (x + 1)\]

**Example 3.5.5** Factor \(x^4 - 1\) as far as possible.

The possible rational roots are \(\pm 1\). Let's try 1

\[
\begin{array}{ccccccc}
1 & 1 & 0 & 0 & 0 & -1 \\
1 & 1 & 1 & 1 & 1 & 0 \\
\end{array}
\]

and so \(x^4 - 1 = (x - 1) (x^3 + x^2 + x + 1)\). At this point, you could observe that

\[x^3 + x^2 + x + 1 = (x + 1) (x^2 + 1)\]

and so \(x^4 - 1 = (x - 1) (x + 1) (x^2 + 1)\) and this is as far as you can go if you want the polynomials to have real coefficients because \(x^2 + 1\) has no real roots. You can see this from the quadratic formula. If you didn’t see the last step, you could just do synthetic division and again. Try -1. The only possibilities are -1 or 1.

\[
\begin{array}{ccccccc}
-1 & 1 & 1 & 1 & 1 & -1 \\
1 & 0 & 1 & 0 & 1 \\
\end{array}
\]

Therefore, this reduces to \((x - 1) (x + 1) (x^2 + 1)\) as above.

**3.6 Exercises**

1. Using Theorem 3.3.1 show \(\sqrt{6}, \sqrt{7}, \sqrt{5}\) are all irrational numbers. This means they are not rational.

2. Consider \(\sqrt{n}\) where \(n\) and \(m\) are positive integers. Show that unless \(\sqrt{n}\) is an integer, it must be irrational.

3. Using the fact that \(\sqrt{2}\) is irrational, (not rational) show that numbers of the form \(r\sqrt{2}\) where \(r \in \mathbb{Q}\) are dense in \(\mathbb{R}\). Then verify these numbers are irrational.

4. Find the solutions to \(x^4 - 5x^3 + 5x^2 + 5x - 6 = 0\).
5. Find the solutions to $x^4 - 5x^3 - 3x^2 + 13x + 10 = 0$.

6. Find the solutions to $2x^3 - x^2 - 5x - 2 = 0$.

7. Suppose $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ is a polynomial and that it has a root $x_0$. Show that there exists a polynomial $q(x)$ such that $p(x) = (x - x_0) q(x)$.

8. If possible, factor $3x^4 + 4x^3 - 14x^2 + 4x + 3$.

9. If possible, factor $x^4 - 3x^2 + 2$.

10. If possible, factor $x^4 - 2x^2 + 4x^3 - 20x - 15$.

11. Let $p(x) = x^5 - 4x^4 + 3x^3 - 9x + 3$. Find $p(5)$.

3.7 Partial Fractions

A useful technique in calculus is the method of partial fractions. First, here is some simple notation.

**Definition 3.7.1** The symbol $\prod_{i=1}^n q_i(x)$ means to take the product $q_1(x)q_2(x)\cdots q_m(x)$.

Similarly the symbol $\prod_{j \neq i} q_j(x)$ means the product of all the polynomials except the $i$th, $q_1(x)q_2(x)\cdots q_{i-1}(x)q_{i+1}(x)\cdots q_m(x)$ continuing this way $\prod_{j \neq \{i_1, \cdots, i_s\}} q_j(x)$ means to multiply all the $q_j(x)$ together except leave out $q_{i_1}(x), \cdots, q_{i_s}(x)$ from the product.

3.7.1 Polynomials With Coefficients In A Field

The polynomials will have coefficients which come from some field. This field could be $\mathbb{R}$ but it could also be the rational numbers or some other field.

The method pertains to quotients of polynomials, called rational functions. The situation is as follows. You are given

$$R(x) = \frac{p(x)}{q_1(x)q_2(x)\cdots q_m(x)},$$

where the polynomials $\{q_1(x), q_2(x), \cdots, q_m(x)\}$ are relatively prime. Also assume the degree of $p(x)$ is less than the degree of the product $q_1(x)q_2(x)\cdots q_m(x)$. If this is not so, you use Theorem 3.4.2 to write

$$R(x) = a(x) + \frac{b(x)}{q_1(x)q_2(x)\cdots q_m(x)}.$$
where \( a(x) \) is a polynomial and the degree of \( b(x) \) is indeed less than the degree of \( q_1(x) q_2(x) \cdots q_m(x) \). Then you work with this last expression. This case is discussed later with examples. (In calculus, polynomials are good, rational functions are bad. Thus the polynomial \( a(x) \) is innocuous and now the rational function is at least of the right form.)

By Theorem 3.4.4, there exist polynomials \( a_i(x) \) such that

\[
1 = \sum_{i=1}^{m} a_i(x) q_i(x).
\]

Then multiply the above rational function by this. Thus

\[
R(x) = \frac{p(x)}{q_1(x) q_2(x) \cdots q_m(x)} \cdot 1 = \frac{p(x) \sum_{i=1}^{m} a_i(x) q_i(x)}{q_1(x) q_2(x) \cdots q_m(x)} = \sum_{i=1}^{m} a_i(x) p(x) = \sum_{i=1}^{m} a_i(x) p(x) = \prod_{j \neq i} q_j(x)
\]

Thus each rational function in the above sum involves a denominator which has the product of only \( m - 1 \) of the \( q_i(x) \) of the original \( m \). Now you can do the same thing to each rational function in the above sum, replacing each one by a sum of rational functions in which the denominator is a product of \( m - 2 \) of the \( q_i(x) \). Continuing this way, you finally end up with the existence of polynomials \( b_1(x), \ldots, b_m(x) \) such that

\[
R(x) = \frac{p(x)}{q_1(x) q_2(x) \cdots q_m(x)} = \sum_{i=1}^{m} b_i(x) q_i(x)
\]

By Euclidean algorithm, Theorem 3.4.2, there exist polynomials \( m_i(x) \) and \( r_i(x) \) with the degree of \( r_i(x) \) either equal to 0 or having degree smaller than the degree of \( q_i(x) \) such that

\[
b_i(x) = m_i(x) q_i(x) + r_i(x).
\]

Substituting this in to the above sum in [3.3] it follows

\[
R(x) = \frac{p(x)}{q_1(x) q_2(x) \cdots q_m(x)} = \sum_{i=1}^{m} r_i(x) q_i(x)
\]

where \( M(x) \) is the sum of the polynomials, \( m_i(x) \). Now \( M(x) \) must equal 0 because if it didn’t equal zero, you could multiply both sides by the product of the \( q_i(x) \) and obtain

\[
p(x) = M(x) q_1(x) q_2(x) \cdots q_m(x) + \sum_{i=1}^{m} r_i(x) \prod_{j \neq i} q_j(x).
\]

The polynomial on the right has degree at least as large as the degree of \( q_1(x) q_2(x) \cdots q_m(x) \) while the polynomial on the left was given to have degree smaller than this. It follows the rational function \( R(x) \) can be written in the form

\[
R(x) = \frac{p(x)}{q_1(x) q_2(x) \cdots q_m(x)} = \sum_{i=1}^{m} \frac{r_i(x)}{q_i(x)}
\]

where the degree of \( r_i(x) \) is less than the degree of \( q_i(x) \) or else \( r_i(x) = 0 \).

Of course this last case cannot occur if the original rational function was reduced to lowest terms. You can show that if \( r_i(x) = 0 \), then \( q_i(x) \) divides \( p(x) \).

In the case where \( q_i(x) = k(x)^{m_i} \), you can go further in the above expansion. This is the message of the following proposition.
Proposition 3.7.2 Let $n$ be a positive integer and let the degree of $r(x)$ be less than the degree of $k(x)^n$. Then there exist polynomials $a_j(x)$ with the degree of $a_j(x)$ less than the degree of $k(x)$ such that

$$\frac{r(x)}{k(x)^n} = \sum_{j=1}^{n} \frac{a_j(x)}{k(x)^j}$$

Proof: This is clearly true if $n = 1$. Suppose it is true for $n - 1$ where $n - 1 \geq 1$ and consider $\frac{r(x)}{k(x)^n}$. Then

$$\frac{r(x)}{k(x)^n} = \frac{1}{k(x)} \left( \frac{r(x)}{k(x)^{n-1}} \right)$$

If the degree of $r(x)$ is less than the degree of $k(x)^{n-1}$ then apply the induction hypothesis and obtain

$$\frac{r(x)}{k(x)^n} = \frac{1}{k(x)} \left( \sum_{j=1}^{n-1} \frac{a_j(x)}{k(x)^j} \right)$$

where the degree of $a_j(x)$ is less than the degree of $k(x)$ and this implies the desired result.

If the degree of $r(x)$ is larger than the degree of $k(x)^{n-1}$, then by the Euclidean algorithm, Theorem 3.4.2,

$$\frac{r(x)}{k(x)^{n-1}} = \frac{1}{k(x)} \left( \frac{b(x)}{k(x)^{n-1}} \right)$$

where $s(x)$ is a polynomial and the degree of $b(x)$ is less than the degree of $k(x)^{n-1}$. Thus

$$\frac{r(x)}{k(x)^n} = \frac{s(x)}{k(x)} + \frac{b(x)}{k(x)^{n-1}}$$

It must be the case that the degree of $s(x)$ is less than the degree of $k(x)$ since otherwise, multiply by $k(x)^n$ and conclude

$$r(x) = s(x) k(x)^{n-1} + b(x)$$

and the degree of the polynomial on the right would be larger than the degree of the polynomial on the left. Now this reduces the problem to the first case. Use the induction hypothesis and get the result from (3.5). This proves the proposition.

3.7.2 Real Polynomials

In calculus, the polynomials have coefficients which come from $\mathbb{R}$.

Later I shall discuss the fundamental theorem of algebra. This fundamental result can be used to show that if you have any polynomial $p(x)$ with real coefficients, there always exist degree 1 polynomials ($a_i x + b_i$) and quadratic polynomials

$$A_i x^2 + B_i x + C_i$$

They may exist but you typically won’t be able to find them. This is often the case in mathematics. There will be a big theorem which says something exists but it gives no hint on how to go about finding the thing.
along with positive integers $m_i$ and $n_i$ such that
\[
p(x) = \prod_{i=1}^{s} (a_i x + b_i)^{m_i} \prod_{j=1}^{p} (A_j x^2 + B_j x + C_j)^{n_j}.
\]
and the polynomials
\[
(a_i x + b_i)^{m_i}, (A_j x^2 + B_j x + C_j)^{n_j}, i = 1 \cdots s, j = 1 \cdots p
\] are relatively prime. Also $A_i x^2 + B_i x + C_i$ is irreducible meaning that no degree 1 polynomial with real coefficients divides it.

Then this gives the following fundamental result for quotients of real polynomials (those with real coefficients).

**Theorem 3.7.3** Let $R(x) = \frac{p(x)}{q(x)}$ be a rational function, the quotient of polynomials with real coefficients. Then there exist linear and irreducible quadratic factors of the sort in (3.6) such that
\[
q(x) = \prod_{i=1}^{m} (a_i x + b_i)^{m_i} \prod_{j=1}^{n} (A_j x^2 + B_j x + C_j)^{n_j}
\]
and there also exist constants $c_{si}$ and linear polynomials
\[
m_{pj} x + t_{pj}
\]
such that
\[
R(x) = \sum_{i=1}^{m} \sum_{s=1}^{m_i} \frac{c_{si}}{(a_i x + b_i)^s} + \sum_{j=1}^{n} \sum_{p=1}^{n_j} m_{pj} x + t_{pj}
\]

It can also be shown the partial fractions expansion is unique but this is not of concern here. Its existence is the important thing in the applications. Nevertheless, I shall use the definite article in referring to it. Now it is necessary to consider some examples to illustrate the technique.

### 3.7.3 Examples

**Example 3.7.4** Find the partial fractions expansion for
\[
x^4 + 2x^2 + 2x + 1 = \frac{x^4 + 2x^2 + 2x + 1}{(x + 1)(x^2 + 1)} = \frac{x^4 + 2x^2 + 2x + 1}{x^3 + x + x^2 + 1}
\]

First of all, you note that the degree of the top, 4 is larger than the degree of the bottom 3. By the Euclidean algorithm,
\[
x^4 + 2x^2 + 2x + 1 = a(x)(x^3 + x + x^2 + 1) + r(x)
\]
where $r(x)$ has degree less than 3. It only remains to find $a(x)$ and $r(x)$. Clearly $a(x)$ must have degree 1 since otherwise the degree of the right side would be larger than the degree of the left. Say $a(x) = ax + b$. Thus
\[
x^4 + 2x^2 + 2x + 1 = a(x)(x^3 + x + x^2 + 1) + r(x)
\]
\[
= ax^4 + ax^3 + ax + bx^3 + bx + bx^2 + b + r(x)
\]
\[
= ax^4 + (a + b)x^3 + (a + b)x^2 + (a + b)x + b + r(x)
\]
It follows right away that \( a = 1 \) and \( a + b = 0 \) so \( b = -1 \). Then also
\[
2x^2 + 2x + 1 = -1 + r(x)
\]
and so \( r(x) = 2(x^2 + x + 1) \). Thus it reduces to
\[
\frac{x^4 + 2x^2 + 2x + 1}{(x + 1)(x^2 + 1)} = x - 1 + \frac{2(x^2 + x + 1)}{(x + 1)(x^2 + 1)}
\] (3.7)
The two polynomials in the bottom are obviously relatively prime because if a polynomial divides both of them, then it must have degree 1 and so if it divides \( x^2 + 1 \), then \( x^2 + 1 \) would have a real zero which it doesn’t. From the partial fractions theorem, the form of this last rational function is known to be
\[
\frac{a}{x + 1} + \frac{bx + c}{x^2 + 1} = \frac{2(x^2 + x + 1)}{(x + 1)(x^2 + 1)}
\]
and it only remains to find \( a, b, \) and \( c \). This is easy. Multiply both sides by \( (x + 1)(x^2 + 1) \) and get
\[
a(x^2 + 1) + (bx + c)(x + 1) = 2(x^2 + x + 1)
\] (3.8)
Now plug in \( x = -1 \) to both sides. This gives
\[
2a = 2
\]
so \( a = 1 \). Now place this value of \( a \) into the equation of (3.8) and obtain
\[
(bx + c)(x + 1) = x^2 + 2x + 1
\]
Then multiplying the left side gives \( b = 1, c = 1 \) and so
\[
\frac{x^4 + 2x^2 + 2x + 1}{(x + 1)(x^2 + 1)} = x - 1 + \frac{1}{(x + 1)} + \frac{x + 1}{x^2 + 1}
\]
In the first step, you can use the traditional algorithm which used to be taught in fourth grade for finding the quotient.
\[
\begin{array}{c|ccccc}
& x^3 & + x^2 & + x & + 1 \\
\hline
x^4 & + 0x^3 & + 2x^2 & + 2x & + 1 \\
\hline
& -x^3 & + x^2 & + x & + 1 \\
& -x^3 & - x^2 & - x & - 1 \\
\hline
& 2x^2 & + 2x & + 2
\end{array}
\]
and so, as in fourth grade,
\[
x^4 + 2x^2 + 2x + 1 = (x - 1)(x^3 + x^2 + x + 1) + 2(x^2 + x + 1)
\]
which is what was obtained above.

Now here is an exotic example.

**Example 3.7.5** Find the partial fractions expansion for
\[
\frac{x + 1}{x(x^2 - 3)}
\]
in terms of sums of quotients of polynomials which have all rational coefficients.
The degree of the top is less than the degree of the bottom. The second order polynomial in the bottom is irreducible in the sense that no polynomial of degree 1 which has rational coefficients can divide this. This is because is $ax + b$ divided it, then
\[ x^2 - 3 = (ax + b) \cdot a(x) \]
and so $x = -b/a$ would be a zero of $x^2 - 3$. If both $a, b$ are rational, then this quotient is also rational and so $x^2 - 3$ would have a rational zero which it doesn’t. A short computation involving the rational root theorem shows this is the case. Therefore, it is indeed irreducible. In the first part of the above argument for the partial fractions decomposition theorem, it referred to polynomials and the coefficients were in a field. Here I am taking the field to be the rational numbers. Then the partial fractions theorem implies there exist rational numbers, $a, b, c$ such that
\[ \frac{x + 1}{x(x^2 - 3)} = \frac{a}{x} + \frac{bx + c}{x^2 - 3} \]
Multiply both sides by $x(x^2 - 3)$. Then
\[ x + 1 = a(x^2 - 3) + (bx + c)x \]
Thus letting $x = 0$, it follows $a = -1/3$. Next plug this in to the above equation to use what you just found out.
\[ x + 1 = \left( \frac{-1}{3} \right) (x^2 - 3) + (bx + c)x \]
and so
\[ x + 1 + \left( \frac{1}{3} \right) (x^2 - 3) = bx^2 + cx. \]
Thus, equating coefficients,
\[ b = \frac{1}{3}, \quad c = 1 \]
It follows
\[ \frac{x + 1}{x(x^2 - 3)} = \frac{-1/3}{x} + \frac{(1/3)x + 1}{x^2 - 3} \]
Of course the polynomial $x^2 - 3$ is not irreducible if the field of scalars is $\mathbb{R}$. Indeed, it equals $(x - \sqrt{3})(x + \sqrt{3})$. In terms of real coefficients,
\[ \frac{x + 1}{x(x^2 - 3)} = -\frac{1}{3x} + \frac{1 + \sqrt{3}}{6x - \sqrt{3}} - \frac{1 - \sqrt{3}}{6x + \sqrt{3}} \]
**Example 3.7.6** Find the partial fractions expansion for
\[ \frac{x^4 + 4x^3 + 8x^2 + 8x + 3}{(2x + 1)(x^2 + x + 1)^2} \]
The degree of the top is less than the degree of the bottom and so I don’t have to first divide the top by the bottom first. Then the partial fractions theorem says there is a partial fractions expansion of the form
\[ \frac{x^4 + 4x^3 + 8x^2 + 8x + 3}{(2x + 1)(x^2 + x + 1)^2} = \frac{a}{2x + 1} + \frac{bx + c}{x^2 + x + 1} + \frac{cx + d}{(x^2 + x + 1)^2} \]
This is the case because the polynomial \( x^2 + x + 1 \) is irreducible in the sense it cannot be divided by any degree one polynomial having real coefficients. This follows from the quadratic formula. Using this formula, you see that this polynomial has no real zeros. However, you can say even more using the rational root theorem. This polynomial has no rational zeros and so it is irreducible with respect to polynomials having rational coefficients. Thus it will end up being the case that each of \( a, b, c, \) and \( d \) will be rational numbers. Multiply both sides by the denominator of the left side. Then

\[
x^4 + 4x^3 + 8x^2 + 8x + 3 = a (x^2 + x + 1)^2 + (bx + c) (x^2 + x + 1) (2x + 1) + (cx + d) (2x + 1)
\] (3.9)

First let \( x = -1/2 \). Then

\[
\left( -\frac{1}{2} \right)^4 + 4 \left( -\frac{1}{2} \right)^3 + 8 \left( -\frac{1}{2} \right)^2 + 8 \left( -\frac{1}{2} \right) + 3 = a \left( \left( -\frac{1}{2} \right)^2 + \left( -\frac{1}{2} \right) + 1 \right)^2
\]

and so

\[
\frac{9}{16} = a \frac{9}{16}
\]

which requires \( a = 1 \). Now use this information in the equation of (3.9). This yields

\[
x^4 + 4x^3 + 8x^2 + 8x + 3 = \left( x^2 + x + 1 \right)^2 + (bx + c) \left( x^2 + x + 1 \right) (2x + 1) + (cx + d) (2x + 1)
\]

Then subtracting the \( \left( x^2 + x + 1 \right)^2 \) from both sides

\[
2x^3 + 5x^2 + 6x + 2 = (bx + c) \left( x^2 + x + 1 \right) (2x + 1) + (cx + d) (2x + 1)
\]

At this point, you notice that \( 2x + 1 \) divides the right side. Therefore, it divides the left side also. Using the fourth grade algorithm, divide both sides by it.

\[
x^2 + 2x + 2 = (bx + c) \left( x^2 + x + 1 \right) + (cx + d)
\]

\[
= bx^3 + (b + c) x^2 + (b + 2c) x + c + d
\]

It follows \( b = 0 \) since there are 0 \( x^3 \) on the left and \( b \) on the right. Then \( c = 1 \) and now \( 1 + d = 2 \) so \( d = 1 \). It follows

\[
\frac{x^4 + 4x^3 + 8x^2 + 8x + 3}{(2x + 1) \left( x^2 + x + 1 \right)^2} =
\]

\[
\frac{1}{2x + 1} + \frac{1}{x^2 + x + 1} + \frac{x + 1}{(x^2 + x + 1)^2}
\]

When you do one of these, you ought to check your work because it is easy to make mistakes. Let's check this one. The alleged partial fractions expansion is of the form

\[
\frac{(x^2 + x + 1)^2}{(2x + 1) \left( x^2 + x + 1 \right)^2} + \frac{(x^2 + x + 1) (2x + 1)}{(x^2 + x + 1)^2 (2x + 1)} + \frac{(x + 1) (2x + 1)}{(x^2 + x + 1)^2 (2x + 1)}
\]
After lots of algebra and much sorrow, this reduces to

\[
\frac{(x^2 + x + 1)^2 + (x^2 + x + 1) (2x + 1) + (x + 1) (2x + 1)}{(2x + 1) (x^2 + x + 1)^2} = \frac{x^4 + 4x^3 + 8x^2 + 8x + 3}{(2x + 1) (x^2 + x + 1)^2}
\]

so it appears I got it right.

This is the way you do partial fractions. The thing to be very careful about is that you look for a partial fractions expansion which is in the correct form. Otherwise, you might be looking for something which is not there. When you do this, you won’t find it. Math is like that. Attitude and positive thinking and hard work are useless if you are trying to do something which cannot be done and there are such things. For example, if I had tried to find a partial fractions expansion of the form

\[
\frac{a}{2x + 1} + \frac{cx + d}{(x^2 + x + 1)^2},
\]

I couldn’t have done it no matter how hard I tried. The other aspect of this is to be tricky when you find the coefficients. Never work harder than you have to. Look for ways to find some of the constants easily and then use what you find to make it easier to find the others. This policy will result in less sorrow when finding partial fractions expansions.

**Example 3.7.7** Find the partial fractions expansion for

\[
\frac{6x^3 + 16x^2 + 6 + 17x}{x^3 + 3x^3 + 2x^2 - 2x - 4}
\]

First it is necessary to factor the bottom. From the rational root theorem, there are many possibilities for rational roots, among them ±1 which is always a possibility. Let’s start with one of these.

<table>
<thead>
<tr>
<th>1</th>
<th>3</th>
<th>2</th>
<th>-2</th>
<th>-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

and so the bottom equals \((x - 1) (x^3 + 4x^2 + 6x + 4)\). Now I will do it again. This time, I will try \(-2\) which is going to work because I made up the problem. However, when you do it in practice, you need to try things till you find one which does work.

<table>
<thead>
<tr>
<th>-2</th>
<th>1</th>
<th>4</th>
<th>-2</th>
<th>6</th>
<th>4</th>
<th>-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and so the rational function is of the form

\[
\frac{6x^3 + 16x^2 + 6 + 17x}{(x - 1) (x + 2) (x^2 + 2x + 2)}
\]

That quadratic polynomial is irreducible as can be seen from the quadratic formula which will show it has no real roots. Then the form of the partial fractions expansion is

\[
\frac{a}{x - 1} + \frac{b}{x + 2} + \frac{cx + d}{x^2 + 2x + 2}
\]

Thus

\[
6x^3 + 16x^2 + 6 + 17x
\]
3.8 Field Extensions

When you have a polynomial like \( x^2 - 3 \) which has no rational roots, it turns out you can enlarge the field of rational numbers to obtain a larger field such that this polynomial does have roots in this larger field. I am going to discuss a systematic way to do this. It will turn out that for any polynomial with coefficients in any field, there always exists a possibly larger field such that the polynomial has roots in this larger field. First it is necessary to discuss the concept of an equivalence relation.

There are many ways to compare elements of a set other than to say two elements are equal or the same. For example, in the set of people let two people be equivalent if they have the same weight. This would not be saying they were the same person, just that they weighed the same. Often such relations involve considering one characteristic of the elements of a set and then saying the two elements are equivalent if they are the same as far as the given characteristic is concerned.

**Definition 3.8.1** Let \( S \) be a set. \( \sim \) is an equivalence relation on \( S \) if it satisfies the following axioms.

1. \( x \sim x \) for all \( x \in S \). (Reflexive)
2. If \( x \sim y \) then \( y \sim x \). (Symmetric)
3. If \( x \sim y \) and \( y \sim z \), then \( x \sim z \). (Transitive)

**Definition 3.8.2** \( [x] \) denotes the set of all elements of \( S \) which are equivalent to \( x \) and \( [x] \) is called the equivalence class determined by \( x \) or just the equivalence class of \( x \).

With the above definition one can prove the following simple theorem.

**Theorem 3.8.3** Let \( \sim \) be an equivalence class defined on a set, \( S \) and let \( \mathcal{H} \) denote the set of equivalence classes. Then if \( [x] \) and \( [y] \) are two of these equivalence classes, either \( x \sim y \) and \( [x] = [y] \) or it is not true that \( x \sim y \) and \( [x] \cap [y] = \emptyset \).
Let \([x] \cap [y] \neq \emptyset\). Then there exists \(z \in [x] \cap [y]\) and so \(x \sim z\) and \(z \sim y\). By the transitive property \(x \sim y\) and so \([x] \subseteq [y]\) and \([y] \subseteq [x]\) so the two are the same. Thus any two equivalence classes, \([x]\), \([y]\) either have empty intersection when \(x\) is not similar to \(y\) or they coincide and \(x\) is similar to \(y\). This proves the theorem.

**Definition 3.8.4** Let \(\mathbb{F}\) be a field, for example the rational numbers, and denote by \(\mathbb{F}(x)\) the polynomials having coefficients in \(\mathbb{F}\). Suppose \(p(x)\) is a polynomial. Let \(a(x) \sim b(x)\) \((a(x)\) is similar to \(b(x))\) when

\[
a(x) - b(x) = k(x)p(x)
\]

for some polynomial \(k(x)\).

**Proposition 3.8.5** In the above definition, \(\sim\) is an equivalence relation.

**Proof:** First of all, note that \(a(x) \sim a(x)\) because there difference equals \(0p(x)\). If \(a(x) \sim b(x)\), then \(a(x) - b(x) = k(x)p(x)\) for some \(k(x)\). But then \(b(x) - a(x) = -k(x)p(x)\) and so \(b(x) \sim a(x)\). Next suppose \(a(x) \sim b(x)\) and \(b(x) \sim c(x)\). Then \(a(x) - b(x) = k(x)p(x)\) for some polynomial \(k(x)\) and also \(b(x) - c(x) = l(x)p(x)\) for some polynomial \(l(x)\). Then

\[
a(x) - c(x) = a(x) - b(x) + b(x) - c(x)
\]

\[
= k(x)p(x) + l(x)p(x) = (l(x) + k(x))p(x)
\]

and so \(a(x) \sim c(x)\) and this shows the transitive law. This proves the proposition.

With this proposition, here is another definition which essentially describes the elements of the new field. It will be necessary to assume the polynomial \(p(x)\) in the above definition is irreducible so I will begin assuming this.

**Definition 3.8.6** Let \(\mathbb{F}\) be a field and let \(p(x) \in \mathbb{F}(x)\) be irreducible. This means there is no polynomial which divides \(p(x)\) except for itself and constants. For the similarity relation defined in Definition [3.8.4], define

\[
[a(x)] + [b(x)] \equiv [a(x) + b(x)]
\]

\[
[a(x)][b(x)] \equiv [a(x)b(x)]
\]

**Proposition 3.8.7** In the situation of Definition [3.8.6], \(p(x)\) and \(q(x)\) are relatively prime for any \(q(x) \in \mathbb{F}(x)\) which is not a multiple of \(p(x)\). Also the definitions of addition and multiplication are well defined. In addition, if \(a, b \in \mathbb{F}\) and \([a] = [b]\), then \(a = b\).

**Proof:** First consider the claim about \(p(x), q(x)\) being relatively prime. If \(\psi(x)\) is the greatest common divisor, it follows \(\psi(x)\) is either equal to \(p(x)\) or 1. If it is \(p(x)\), then \(q(x)\) is a multiple of \(p(x)\). If it is 1, then by definition, the two polynomials are relatively prime.

To show the operations are well defined, suppose

\[
[a(x)] = [a'(x)], [b(x)] = [b'(x)]
\]

It is necessary to show

\[
[a(x) + b(x)] = [a'(x) + b'(x)]
\]

\[
[a(x)b(x)] = [a'(x)b'(x)]
\]
Consider the second of the two.

\[ a' (x) b' (x) - a (x) b (x) \]
\[ = a' (x) b' (x) - a (x) b' (x) + a (x) b' (x) - a (x) b (x) \]
\[ = b' (x) (a' (x) - a (x)) + a (x) (b' (x) - b (x)) \]

Now by assumption \((a' (x) - a (x))\) is a multiple of \(p (x)\) and so is \((b' (x) - b (x))\) so the above is a multiple of \(p (x)\) and by definition this shows \([a (x) b (x)] = [a' (x) b' (x)]\). The case for addition is similar.

Now suppose \([a] = [b]\). This means \(a - b = k (x)p (x)\) for some polynomial \(k (x)\). Then \(k (x)\) must equal 0 since otherwise the two polynomials \(a - b\) and \(k (x)p (x)\) could not be equal. This proves the proposition.

Note that from this proposition and math induction, if each \(a_i \in \mathbb{F}\),

\[ [a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0] \]
\[ = [a_n] [x]^n + [a_{n-1}] [x]^{n-1} + \cdots [a_1] [x] + [a_0] \quad (3.10) \]

With the above preparation, here is a definition of a field in which the irreducible polynomial \(p (x)\) has a root.

**Definition 3.8.8** Let \(p (x) \in \mathbb{F}\) be irreducible and let \(a (x) \sim b (x)\) when \(a (x) - b (x)\) is a multiple of \(p (x)\). Let \(\mathcal{G}\) denote the set of equivalence relations as described above with the operations also described in Definition 3.8.6.

**Theorem 3.8.9** The set of all equivalence relations \(\mathcal{G}\) described above with the multiplicative identity given by \([1]\) and the additive identity given by \([0]\) along with the operations of Definition 3.8.6 is a field and \(p ([x]) = [0]\).

**Proof:** Everything is obvious except for the existence of the multiplicative inverse and the assertion that \(p ([x]) = 0\). Suppose then that \([a (x)] \neq [0]\). That is, \(a (x)\) is not a multiple of \(p (x)\). Why does \([a (x)]^{-1}\) exist? By Proposition 3.8.7 \(a (x), p (x)\) are relatively prime and so there exist polynomials \(\psi (x), \phi (x)\) such that

\[ 1 = \psi (x) p (x) + a (x) \phi (x) \]

and so

\[ [1 - a (x) \phi (x)] = [a (x) p (x)] \]

Thus \(1 \sim a (x) \phi (x)\). It follows from the definition that

\[ [a (x) p (x)] = [1] \]

and so \([\phi (x)] = [a (x)]^{-1}\). This shows \(\mathcal{G}\) is a field.

Now if \(p (x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0, p ([x]) = 0\) by 3.10 and the definition which says \([p (x)] = [0]\). This proves the theorem.

Usually, people simply write \(b\) rather than \([b]\) if \(b \in \mathbb{F}\). Then with this convention,

\[ [b \phi (x)] = [b] [\phi (x)] = b [\phi (x)] \]

This shows how to enlarge a field to get a new one in which the polynomial has a root. By using a succession of such enlargements, called field extensions there will exist a field in which the given polynomial can be factored into a product of polynomials having degree one. See Problem 21 below.
Example 3.8.10  The polynomial \( x^2 + 1 \) is irreducible in \( \mathbb{R}(x) \), polynomials having real coefficients. To see this is the case, suppose \( \psi(x) \) divides \( x^2 + 1 \). Then
\[
x^2 + 1 = \psi(x) q(x)
\]
where the degree of \( r(x) \) is less than the degree of \( \psi(x) \). If the degree of \( \psi(x) \) is 2, then it must be \( x^2 + 1 \) and \( q(x) \) must be 1 since otherwise the two sides could not be equal. If the degree of \( \psi(x) \) is less than 2, then it must be either a constant or of the form \( ax + b \). In the latter case, \( -b/a \) must be a zero of the right side, hence of the left but \( x^2 + 1 \) has no real zeros. Therefore, this shows it is irreducible. Find the inverse of \( [x^2 + x + 1] \).

You can solve this with partial fractions.

\[
\frac{1}{(x^2 + 1)(x^2 + x + 1)} = \frac{x}{x^2 + 1} + \frac{x + 1}{x^2 + x + 1}
\]

and so
\[
1 = (-x)(x^2 + x + 1) + (x + 1)(x^2 + 1)
\]

which implies
\[
1 \sim (-x)(x^2 + x + 1)
\]

and so the inverse is \([-x]\).

The following proposition is interesting. It gives more information on the things in \( \mathcal{G} \) described above.

**Proposition 3.8.11**  Suppose \( p(x) \in \mathbb{F}(x) \) is irreducible and has degree \( n \). Then every element of \( \mathcal{G} \) is of the form \([0]\) or \([r(x)]\) where the degree of \( r(x) \) is less than \( n \).

**Proof:** This follows right away from the Euclidean algorithm for polynomials. If \( k(x) \) has degree larger than \( n - 1 \), then
\[
k(x) = q(x)p(x) + r(x)
\]

where \( r(x) \) is either equal to 0 or has degree less than \( n \). Hence
\[
[k(x)] = [r(x)].
\]

This proves the proposition.

**Example 3.8.12**  In the situation of the above example, find \( [ax + b]^{-1} \) assuming \( a^2 + b^2 \neq 0 \). Note this includes all cases of interest thanks to the above proposition.

You can do it with partial fractions as above.

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

and so
\[
1 = \frac{1}{a^2 + b^2} (b - ax)(ax + b) + \frac{a^2}{(a^2 + b^2)} (x^2 + 1)
\]

Thus
\[
\frac{1}{a^2 + b^2} (b - ax)(ax + b) \sim 1
\]

and so
\[
[ax + b]^{-1} = \frac{1}{a^2 + b^2} [(b - ax)]
\]
3.9 Exercises

1. Show that if
   \[ \frac{p(x)}{q_1(x)q_2(x)\cdots q_m(x)} = \sum_{i=1}^{m} r_i(x) \]
   and \( r_i(x) = 0 \) for some \( i \), then \( q_i(x) \) divides \( p(x) \).

2. To get an idea why the partial fractions expansion is unique, suppose
   \[ \frac{a}{x + 1} + \frac{bx + c}{x^2 + 1} = \frac{a'}{x + 1} + \frac{b'x + c'}{x^2 + 1}. \]
   Show this implies \( a = a', b = b', c = c' \).

3. Find the partial fractions expansion of
   \[ \frac{2x^4 + 17x^3 + 55x^2 + 79x + 43}{(x + 1)^2 (x + 3)(x + 2)^2} \]

4. Find the partial fractions expansion for
   \[ \frac{5x^4 + 10x^2 + 3 + 4x^3 + 5x}{(2x + 1)(x^2 + 1)(x^2 + 2)} \]

5. Find the partial fractions expansion for
   \[ \frac{2x^3 - 4 + 3x^2 + x}{(x + 3)(x^3 - 7)} \]
   in terms rational functions which have all rational coefficients. **Hint:** The polynomial \( x^3 - 7 \) is irreducible over the rationals. You can verify this by using the rational root theorem.

6. Find the partial fractions expansion for
   \[ \frac{2x^3 - 5 + x^2}{x^4 - 5x + x^3 - 5} \]
   in terms rational functions which have all rational coefficients.

7. Find the partial fractions expansion for
   \[ \frac{2x^2 + x}{x^3 + 2x^2 + 2x + 1} \]

8. Find the partial fractions expansion of
   \[ \frac{2x^5 + 5x^4 + 4x^3 + 4x^2 + 7x + 3}{(x + 1)^2(2x + 1)} \]
   **Hint:** Alas, the degree of the top is larger than the degree of the bottom. Recall, this means you have to first do a division.

9. Show using the technique of partial fractions that
   \[ \sum_{k=1}^{n} \frac{1}{k(k+1)} = 1 - \frac{1}{n+1} \]
10. Show using the technique of partial fractions that
\[ \sum_{k=1}^{n} \frac{1}{k(k+2)} = \frac{3}{4} \left( \frac{1}{n+2} + \frac{1}{2(n+1)(n+2)} \right) \]

11. Using the synthetic division algorithm, show that if \( p(x) \) is the polynomial
\[ p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \]
and each \( a_i \) is larger than 0, then this polynomial can’t have any positive real zeros. There is a famous rule called Descartes’ rule of signs which allows you to estimate the number of positive and then negative roots of a polynomial written in the above form. It says the number of positive roots equals the number of variations in the sign of the nonzero coefficients or an even integer less than that number. The number of negative roots equals the number of variations in the signs of the coefficients of \( p(-x) \) written as
\[ (-1)^n a_n x^n + (-1)^{n-1} a_{n-1} x^{n-1} + \cdots + (-1) a_1 x + a_0 \]
or is an even integer less than this.

12. Using Descartes’ rule of signs, estimate the number of positive and negative roots of \( 2x^4 - 5x^3 - 5x^2 + 5x + 3 \). Next find the roots using synthetic division and the rational root theorem.

13. As explained earlier, \( \sqrt{2} \) is not rational. Consider numbers which are of the form \( a + b\sqrt{2} \) where \( a, b \) are rational. Show such numbers form a field. **Hint:** The only hard part is to show \( (a + b\sqrt{2})^{-1} \) exists.

14. In the above problem, explain why the numbers of the form described there constitute the smallest field containing the rational numbers which also contains \( \sqrt{2} \). You started with a field, the rational numbers, and you enlarged it just enough to get a field which also contains \( \sqrt{2} \).

15. There are lots of fields. Examples are the rational numbers, the real numbers and the complex numbers. In this problem, I will give another example of a finite field. Let \( \mathbb{Z} \) denote the set of integers. Thus \( \mathbb{Z} = \{ \cdots, -3, -2, -1, 0, 1, 2, 3, \cdots \} \). Also let \( p \) be a prime number. We will say that two integers, \( a, b \), are equivalent and write \( a \sim b \) if \( a - b \) is divisible by \( p \). Thus they are equivalent if \( a - b = px \) for some integer \( x \). First show that \( a \sim a \). Next show that if \( a \sim b \) then \( b \sim a \). Finally show that if \( a \sim b \) and \( b \sim c \) then \( a \sim c \). For \( a \) an integer, denote by \([a]\) the set of all integers which is equivalent to \( a \), the equivalence class of \( a \). Show first that is suffices to consider only \([a]\) for \( a = 0, 1, 2, \cdots, p-1 \) and that for \( 0 \leq a < b \leq p-1, [a] \neq [b] \). That is, \([a] = [r]\) where \( r \in \{0, 1, 2, \cdots, p-1\} \). Thus there are exactly \( p \) of these equivalence classes. **Hint:** Recall the Euclidean algorithm. For \( a > 0 \), \( a = mp+r \) where \( r < p \). Next define the following operations.

\[
[a] + [b] \equiv [a + b] \\
[a][b] \equiv [ab]
\]

Show these operations are well defined. That is, if \([a] = [a']\) and \([b] = [b']\), then \([a] + [b] = [a'] + [b']\) with a similar conclusion holding for multiplication. Thus for addition you need to verify \([a + b] = [a' + b']\) and for multiplication you need to verify \([ab] = [a'b']\). For example, if \( p = 5 \) you have \([3] = [8]\) and \([2] = [7]\). Is
[2 \times 3] = [8 \times 7]? Is [2 + 3] = [8 + 7]? Clearly so in this example because when you subtract, the result is divisible by 5. So why is this so in general? Now verify that \{[0], [1], \ldots, [p - 1]\} with these operations is a Field. This is called the integers modulo a prime and is written \mathbb{Z}_p. Since there are infinitely many primes \(p\), it follows there are infinitely many of these finite fields. **Hint:** Most of the axioms are easy once you have shown the operations are well defined. The only two which are tricky are the ones which give the existence of the additive inverse and the multiplicative inverse. Of these, the first is not hard. \(-[x] = [-x].\) For the second one, you should use Theorem 3.3.3. Since \(p\) is prime, there exists integers \(x, y\) such that \(1 = px + ky\) and so \(1 - ky = px\) which says \(1 \sim ky\) and so \([1] = [ky]\). Now you finish the argument. What is the multiplicative identity in this collection of equivalence classes?

16. Consider the polynomials which have coefficients in a field of scalars. For example, you could consider the polynomials which have rational coefficients. Show that with the usual conventions of multiplication and addition, these polynomials satisfy all the field axioms except for the one which says 1 is the multiplicative identity. Also show that if \(a \in \mathbb{Q}\) is a polynomial and \(q(x) \in I\), it follows \(a(x)q(x)\) is also in \(I\). Such a set is called an ideal.

**Definition 3.9.1** An ideal \(I\) in the set of polynomials is a set which has the properties that it satisfies the field axioms except for having the multiplicative identity 1 and nonzero elements having multiplicative inverses and if \(b \in I\) and \(a\) is any polynomial, then \(ab \in I\).

17. Consider the polynomials which have coefficients in a field. Now let \(p(x)\) be a polynomial and let \(I\) consist of all polynomials of the form \(a(x)p(x)\) where \(a(x)\) is a polynomial. Show that \(I\) also satisfies all the field axioms except for the one about multiplicative inverses and having a multiplicative identity. Also show that whenever \(a(x)\) is a polynomial and \(q(x) \in I\), it follows \(a(x)q(x)\) is also in \(I\). Such a set is called an ideal.

18. Now let \(p(x)\) be an irreducible polynomial as discussed above. That is, no polynomial divides \(p(x)\) except itself and nonzero constants. Let \(I\) be the ideal defined above as all polynomials of the form \(a(x)p(x)\) for \(a(x)\) a polynomial. Show that if \(J\) is any ideal which contains \(I\), then \(J\) must consist of all polynomials. Such an ideal is called a maximal ideal. **Hint:** Take \(b(x) \in J\) but \(b(x) \notin I\). Argue \(b(x)\) and \(p(x)\) are relatively prime. Then use Theorem 3.3.4 to verify \(1 \in J\). Since \(J\) is an ideal, show this implies all polynomials are in \(J\).

19. The polynomial \(x^3 + x + 1\) is irreducible in \(\mathbb{Q}(x)\), the polynomials having rational coefficients. Find \([x^2 + x + 3]^{-1}\). **Hint:** To save you trouble,

\[
\frac{1}{(x^3 + x + 1)(x^2 + x + 1)} = \frac{1}{3} \left( \frac{2x^2 + x + 1}{x^3 + x + 1} + \frac{1 - x}{x^2 + x + 1} \right)
\]

20. Explain why \(x^2 + x + 1\) is irreducible in \(\mathbb{Q}(x)\), the polynomials having rational coefficients. Find \([x + 2]^{-1}\). Would the result have been any different if I had used the fact this polynomial is irreducible in \(\mathbb{R}(x)\)?

21. Let \(p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0\) be a polynomial with coefficients in a field of scalars \(F\). Show there exists a larger field \(G\) such that there exist
\{z_1, \cdots, z_n\} listed according to multiplicity such that

\[ p(x) = \prod_{i=1}^{n} (x - z_i) \]

This is called the splitting field. **Hint:** Use the preceding problem to argue there exists \( F_1 \) at least as large as \( F \) such that \( p(x) = (x - z_1) q_1(x) \) where \( q(x) \) is polynomial with coefficients in \( F_1 \). If \( q(x) \) is not irreducible, then \( q_1(x) = (x - z_2) q_2(x) \) for some \( z_2 \in F_2 \). If it is irreducible, you can use the result of the preceding problem again and obtain a larger field \( F_2 \) such that \( q_1(x) = (x - z_2) q_2(x) \) where \( q_2(x) \) has coefficients in \( F_2 \). Continue this way to obtain the desired result.
Sequences And Series

4.1 Basic Concepts

A sequence is just a kind of a function whose domain is an infinite set of integers. More precisely, here is a definition.

Definition 4.1.1 A function whose domain is defined as a set of the form
\[ \{k, k + 1, k + 2, \cdots \} \]
for \( k \) an integer is known as a sequence. Thus you can consider \( f(k), f(k + 1), f(k + 2), \) etc. Usually the domain of the sequence is either \( \mathbb{N} \), the natural numbers consisting of \( \{1, 2, 3, \cdots \} \) or the nonnegative integers, \( \{0, 1, 2, 3, \cdots \} \). Also, it is traditional to write \( f_1, f_2, \) etc. instead of \( f(1), f(2), f(3) \) etc. when referring to sequences. In the above context, \( f_k \) is called the first term, \( f_{k+1} \) the second and so forth. It is also common to write the sequence, not as \( f \) but as \( \{f_i\} \) or just \( \{f_i\} \) for short.

Example 4.1.2 Let \( \{a_k\}_{k=1}^\infty \) be defined by \( a_k \equiv k^2 + 1 \).

This gives a sequence. In fact, \( a_7 = a(7) = 7^2 + 1 = 50 \) just from using the formula for the \( k^{th} \) term of the sequence.

It is nice when sequences come in this way from a formula for the \( k^{th} \) term. However, this is often not the case. Sometimes sequences are defined recursively. This happens, when the first several terms of the sequence are given and then a rule is specified which determines \( a_{n+1} \) from knowledge of \( a_1, \cdots, a_n \). This rule which specifies \( a_{n+1} \) from knowledge of \( a_k \) for \( k \leq n \) is known as a recurrence relation.

Example 4.1.3 Let \( a_1 = 1 \) and \( a_2 = 1 \). Assuming \( a_1, \cdots, a_{n+1} \) are known, \( a_{n+2} \equiv a_n + a_{n+1} \).

Thus the first several terms of this sequence, listed in order, are \( 1, 1, 2, 3, 5, 8, \cdots \). This particular sequence is called the Fibonacci sequence and is important in the study of reproducing rabbits. Note this defines a function without giving a formula for it. Such sequences occur naturally in the solution of differential equations using power series methods and in many other situations of great importance.

For sequences, it is very important to consider something called a subsequence.

Definition 4.1.4 Let \( \{a_n\} \) be a sequence and let \( n_1 < n_2 < n_3, \cdots \) be any strictly increasing list of integers such that \( n_1 \) is at least as large as the first number in the domain of the function. Then if \( b_k \equiv a_{n_k} \), \( \{b_k\} \) is called a subsequence of \( \{a_n\} \).
For example, suppose \( a_n = (n^2 + 1) \). Thus \( a_1 = 2, a_3 = 10 \), etc. If
\[ n_1 = 1, n_2 = 3, n_3 = 5, \ldots, n_k = 2k - 1, \]
then letting \( b_k = a_{n_k} \), it follows
\[ b_k = \left( (2k - 1)^2 + 1 \right) = 4k^2 - 4k + 2. \]
A subsequence is obtained by restricting the domain of the function (sequence) in the manner described above.

### 4.2 Finding A Formula

As mentioned above, sometimes sequences come in terms of a formula like
\[ a_n = n^2 + 1 \]
and sometimes they are defined recursively like
\[ a_{n+2} = -2a_n + 3a_{n+1}, \quad a_1 = 1, \quad a_2 = 2 \tag{4.1} \]
Can you find a formula which will deliver these values and also \( a_4, a_5, a_6, \) etc.? It turns out that sometimes you can do this.

**Example 4.2.1** Find a formula for \( a_n \) given in the above.

In this case, you have \( 1, 1, 4, \ldots \). Consider the above recursively defined sequence. Look for a solution to the recursion relation in the form
\[ a_n = r^n \]
and try to find \( r \) such that things will work out. From (4.1), you must have
\[ r^{n+2} = -2r^n + 3r^{n+1} \]
Divide by \( r^n \). Then you obtain
\[ r^2 = -2 + 3r \]
Writing on the same side of the equal sign,
\[ r^2 - 3r + 2 = 0. \]
At this point, you could use the quadratic formula of Theorem 2.5.6 on Page 51 or simply notice the left side of the above equation factors to give
\[ (r - 1)(r - 2) = 0. \]
Thus there are two possible values of \( r \). Now if \( a_n \) and \( b_n \) both satisfy the recursion relation, of (4.1), consider \( Ca_n + Db_n \). You can verify this also satisfies the recursion relation. Therefore, it is reasonable to look for a solution to (4.1) which is in the form
\[ C + 2D^n = a_n. \]
You need to satisfy the condition \( a_1 = 1, a_2 = 2 \). This requires
\[ C + 2D = 1 \]
\[ C + 4D = 2 \]
Later, a general procedure is presented for solving such a system of equations but right now, you can do the following. Subtract the equations to obtain

\[ 2D = 1 \]

so \( D = 1/2. \) Then also \( C = 0. \) Therefore, the desired formula is

\[ a_n = \frac{1}{2} 2^n = 2^{n-1}. \]

The recursion relation above is sometimes called a second order difference equation. The above technique will often work to find a formula for \( a_n. \) Here is another example.

**Example 4.2.2** Suppose \( a_n = 3a_{n-1} + 2, a_0 = 1. \) Find a formula for \( a_n. \) Note this has a constant added on at the end.

You could reduce this to something like the above example as follows.

\[ a_{n+1} = 3a_n + 2, \quad a_n = 3a_{n-1} + 2 \]

Subtract these and obtain

\[ a_{n+1} - a_n = 3(a_n - a_{n-1}) \]

and so

\[ a_{n+1} = 4a_n - 3a_{n-1} \]

Now proceed as before, look for \( r^n \) satisfying the above recurrence relation. Thus

\[ r^{n+1} = 4r^n - 3r^{n-1} \]

and dividing by \( r^{n-1} \) you get

\[ r^2 - 4r + 3 = 0 \]

which has the solution \( r = 1, 3. \) Therefore, as before, \( a_n \) is of the form

\[ a_n = A1^n + B3^n = A + B3^n \]

where \( A, B \) need to be chosen to satisfy the initial condition \( a_0 = 1. \) Letting \( n = 0, \) this shows

\[ A + B = a_0 = 1 \]

Thus \( A = (1 - B). \) Then the solution must be

\[ a_n = (1 - B) + B3^n. \]

However, from the original problem, \( a_1 = 5. \) Therefore

\[ 5 = (1 - B) + 3B = 1 + 2B \]

which shows \( B = 2. \) Therefore,

\[ a_n = -1 + 2(3^n) \]

Now you could check this to see if it works. Is

\[ -1 + 2(3^n) = 3 \left(-1 + 2 \left(3^{n-1}\right)\right) + 2? \]

Yes, this is so.

There is much more which can be said about recurrence relations like this but this is all which will be attempted here. It is not always possible to find a neat formula for \( a_n \) given by a recurrence relation. In following sections, special recurrence relations will be considered which lead to the geometric and arithmetic sequences.
4.3 Exercises

1. Suppose \( a_n = \frac{1}{n} \) and let \( n_k = 2^k \). Find \( b_k \) where \( b_k = a_{n_k} \).

2. If \( X_i \) are sets and for some \( j \), \( X_j = \emptyset \), the empty set. Verify carefully that \( \prod_{i=1}^{n} X_i = \emptyset \).

3. Suppose \( f(x) + f\left(\frac{1}{x}\right) = 7x \) Find \( D(f) \) and find \( f \).

4. Does there exist a function \( f \), satisfying \( f(x) - f\left(\frac{1}{x}\right) = 3x \) which has both \( x \) and \( \frac{1}{x} \) in the domain of \( f \)? What is \( D(f) \)?

5. In Example 4.2.1 keep the same recursion relation but change the “initial condition” to \( a_1 = 3, a_2 = 5 \).

6. In Example 4.2.1 keep the same recursion relation but change the “initial condition” to \( a_1 = -1, a_2 = 7 \).

7. In Example 4.2.1 keep the same recursion relation but change the “initial condition” to \( a_1 = A, a_2 = B \).

8. In the situation of the Fibonacci sequence show that the formula for the \( n^{th} \) term can be found and is given by

\[
a_n = \frac{\sqrt{5}}{5} \left( \frac{1 + \sqrt{5}}{2} \right)^n - \frac{\sqrt{5}}{5} \left( \frac{1 - \sqrt{5}}{2} \right)^n.
\]

**Hint:** You might be able to do this by induction but a better way would be to look for a solution to the recurrence relation, \( a_{n+2} = a_n + a_{n+1} \) of the form \( r^n \). You will be able to show that there are two values of \( r \) which work, one of which is \( r = \frac{1 + \sqrt{5}}{2} \). Next you can observe that if \( r_1^n \) and \( r_2^n \) both satisfy the recurrence relation then so does \( cr_1^n + dr_2^n \) for any choice of constants \( c, d \). Then you try to pick \( c \) and \( d \) such that the conditions, \( a_1 = 1 \) and \( a_2 = 1 \) both hold.

9. Suppose \( a_1 = 1, a_2 = 3, \) and \( a_3 = -1. \) Suppose also that for \( n \geq 4 \) it is known that \( a_n = a_{n-1} + 2a_{n-2} - 3a_{n-3} \). Find \( a_7 \). Are you able to obtain a formula for the \( k^{th} \) term of this sequence?

10. In an ordinary annuity, you make constant payments, \( P \) at the beginning of each payment period. These accrue interest at the rate of \( r \) per payment period. This means at the start of the first payment period, there is the payment \( P \equiv A_1 \). Then this produces an amount \( rP \) in interest so at the beginning of the second payment period, you would have \( rP + P \equiv A_2 \). Thus \( A_2 = A_1 (1 + r) + P \). Then at the beginning of the third payment period you would have \( A_2 (1 + r) + P \equiv A_3 \). Continuing in this way, you see that the amount in at the beginning of the \( n^{th} \) payment period would be \( A_n \) given by \( A_n = A_{n-1} (1 + r) + P \) and \( A_1 = P \). Thus \( A \) is a function defined on the positive integers given recursively as just described and \( A_n \) is the amount at the beginning of the \( n^{th} \) payment period. Now if you wanted to find out \( A_n \) for large \( n \), how would you do it? One way would be to use the recurrence relation \( n \) times. A better way would be to find a formula for \( A_n \). Look for one in the form \( A_n = Cz^n + s \) where \( C, z \) and \( s \) are to be determined. Show that \( C = \frac{B}{r}, \) \( z = (1 + r), \) and \( s = \frac{P}{r} \).

11. A well known puzzle consists of three pegs and several disks each of a different diameter, each having a hole in the center which allows it to be slid down each of the pegs. These disks are piled one on top of the other on one of the pegs,
in order of decreasing diameter, the larger disks always being below the smaller disks. The problem is to move the whole pile of disks to another peg such that you never place a disk on a smaller disk. If you have \( n \) disks, how many moves will it take? Of course this depends on \( n \). If \( n = 1 \), you can do it in one move. If \( n = 2 \), you would need 3. Let \( A_n \) be the number required for \( n \) disks. Then in solving the puzzle, you must first obtain the top \( n - 1 \) disks arranged in order on another peg before you can move the bottom disk of the original pile. This takes \( A_{n-1} \) moves. Explain why \( A_n = 2A_{n-1} + 1 \), \( A_1 = 1 \) and give a formula for \( A_n \). Look for one in the form \( A_n = Cr^n + s \). This puzzle is called the Tower of Hanoi. When you have found a formula for \( A_n \), explain why it is not possible to do this puzzle if \( n \) is very large.

4.4 Arithmetic And Geometric Sequences

4.4.1 The \( n \)th Term

These special sequences turn out to be very important in applications, especially the geometric sequences. Here is how they are defined.

**Definition 4.4.1** *An arithmetic sequence \( \{a_n\} \) is one which satisfies the recurrence relation*

\[
a_n = a_{n-1} + d
\]

*The constant \( d \) is called the common difference. A sequence \( \{a_n\} \) is called a geometric sequence if *

\[
a_n = ra_{n-1}
\]

*Here \( r \) is called the common ratio.*

Now I will give formulas for the \( n \)th term. In the case of the arithmetic sequence,

\[
a_1 = d + a_0, a_2 = d + a_1 = a_0 + 2d, \\
a_3 = a_2 + d = a_0 + 2d + d = a_0 + 3d
\]

You see the pattern. The formula is just

\[
a_n = a_0 + nd.
\]

If you like, you could give a proof of this by induction. In the case where the first term is \( a_1 \), you would have \( a_n = a_1 + d(n-1) \).

Now consider the geometric sequence.

\[
a_1 = a_0r, a_2 = a_1r = a_0rr = a_0r^2, etc.
\]

The pattern gives

\[
a_n = a_0r^n.
\]

To prove this by induction, note it is true if \( n = 0, 1 \). Suppose it is true for \( n \). Then

\[
a_{n+1} = a_nr = a_0r^n + 1
\]

and so it is true for \( n + 1 \).

Sometimes the first term is considered to be \( a_1 \). Then similarly,

\[
a_2 = a_1r, \cdots a_n = a_1r^{n-1}.
\]
4.4.2 The Sum

It is useful to have a formula for the sum of arithmetic and geometric sequences. First consider the arithmetic sequence. I want to sum from $a_1$ up to $a_n$, the first $n$ terms. First consider the problem of summing the first $n$ positive integers. Let this sum be denoted by $S_n$. Then

\[
S_n = 1 + 2 + 3 + \cdots + (n - 2) + (n - 1) + n
\]

Then it follows

\[
2S_n = n \times \overbrace{(1 + n) + (1 + n) + \cdots + (1 + n)}^{n \text{ times}}
\]

Therefore,

\[
S_n = \frac{(n + 1) n}{2}
\]

In other words, you take the average of the first term and the last term and multiply by the number of terms to get the sum. Now consider the more general case of summing the first $n$ terms in an arithmetic sequence. Again, write this sum as $S_n$. Then using the formula for the $k^{th}$ term given above,

\[
S_n = \sum_{k=1}^{n} a_1 + (k - 1)d
\]

\[
= na_1 + d \sum_{k=1}^{n} (k - 1) = na_1 - dn + d \sum_{k=1}^{n} k
\]

\[
na_1 - dn + d \frac{(n + 1) n}{2} = na_1 + \frac{(n + 1) dn - 2dn}{2}
\]

\[
= na_1 + \frac{d(n - 1) n}{2} = n \cdot a_1 + (a_1 + d(n - 1))
\]

Thus you do the same thing for the general case. You take the average of the first and the last term and multiply by the number of terms to get the sum.

**Example 4.4.2** Find the sum of the odd integers from 1 to 101.

From the above, this is really easy if I can find the number of terms. However, the first term is 1 and the common difference is 2. Therefore, it suffices to solve the following for $n$

\[
1 + (n - 1) \times 2 = 101
\]

Solving this yields $n = 51$. The sum is then

\[
\left( \frac{1 + 101}{2} \right) \times 51 = 2601
\]

Summarizing, here is how you find the sum of an arithmetic sequence.

**Procedure 4.4.3** To find the sum of the first $n$ terms of an arithmetic sequence, take the average of the first and last terms and multiply by the number of terms which can be determined by the formula given above for the $n^{th}$ term.
Next consider the formula for finding the sum of the first $n$ terms of a geometric sequence. From the above recall the $n^{th}$ term is of the form

$$a_1 r^{n-1} = a_n.$$ 

Let $S_n$ denote the sum of the first $n$ terms of this sequence. This will be done only for the case where $r \neq 1$. If $r = 1$, it is too easy. In this case you are just adding $a_1$ to itself $n$ times and so the answer is $na_1$. Thus letting $S_n$ denote the sum of the first $n$ terms,

$$S_n = \sum_{k=1}^{n} a_1 r^{k-1} = \sum_{k=0}^{n-1} a_1 r^k$$

$$rS_n = \sum_{k=1}^{n} a_1 r^k$$

Subtracting these yields

$$S_n (1 - r) = \sum_{k=0}^{n-1} a_1 r^k - \sum_{k=1}^{n} a_1 r^k$$

$$= a_1 + \sum_{k=1}^{n-1} a_1 r^k - \left(\sum_{k=1}^{n-1} a_1 r^k + a_1 r^n\right)$$

$$= a_1 - a_1 r^n.$$ 

It follows

$$S_n = \frac{a_1 - a_1 r^n}{1 - r}.$$ 

This establishes the following procedure for finding the sum of the first $n$ terms of a geometric sequence.

**Procedure 4.4.4** To find the sum $S_n$ of the first $n$ terms of a geometric sequence whose $n^{th}$ term is given by $a_1 r^{n-1}$, you use the following formula

$$S_n = \frac{a_1 - a_1 r^n}{1 - r}.$$ 

In words, you subtract the $(n + 1)^{st}$ term from the first term and then divide what you get by $1 - r$.

Note that the sum of the first $n$ terms $S_n$, is itself a sequence. It is called the sequence of partial sums.

### 4.4.3 Compound Interest

One of the most important applications of the ideas of geometric sequences has to do with compound interest. With compound interest, you have an initial amount, often called the principal, $P$ and it sits in the bank for a period of time during which time the bank pays interest on it. The amount of interest received is $rP$ where $r$ is the interest rate.

This is a very old idea and was known to the ancient Babylonians earlier than 1700 B.C., before most of the events recorded in the Old Testament. Abraham probably knew about this. They proposed problems about how long it would take a given amount to double from compound interest. It is also called usury and is condemned in the Koran and in the Bible, at least for fellow Israelites. The fact that it is mentioned in Exodus shows that the idea was in existence when this book was written which also demonstrates its antiquity. Luther and other religious leaders regarded it as a sin. In ancient Babylon, the interest rate is thought to have been 20%.
rate per “payment period”. This payment period may be a month, a year, a day, etc. However, interest rates are usually stated in terms of a year. If a bank gives 5% interest, they mean 5% per year. However, they might compound the interest quarterly. Thus 5% compounded quarterly refers to 4 payment periods per year and \((5/4)\% = 1.25\%\) being the interest rate for each of these payment periods. Also recall the meaning of \%. In words, we say per cent. The cent refers to 100. Thus 5\% is really 5 per 100, the fraction \(5/100\), written in decimals as .05 and the interest rate corresponding to 5\% to be used in computations is .05.

Now at the end of a payment period, you have principal plus interest, \(P + rP = (1 + r)P\) in the bank. This is the new amount the bank will use for the next payment period. Thus if \(A_n\) is the amount at the end of the \(n^{th}\) payment period, \(A_{n+1} = A_n (1 + r)\), \(A_0 = P\).

Thus \(\{A_n\}\) is a geometric sequence and so the amount at the end of period \(n\) is \(A_n = P (1 + r)^n\)

where \(P\) is the initial amount or principal placed in the bank at time 0, the beginning of the first payment period. This is called the future value of \(P\).

**Example 4.4.5** The bank offers an interest rate of 5\% compounded quarterly. Find the amount in the bank at the end of 6 years if the initial amount placed in the bank is \$1000. How much would there be at the end of 15 years?

According to the above, the interest rate per payment period is 1.25\% or .0125. There are 24 payment periods. Therefore, the amount present in the bank at the end of these 24 payment periods is

\[1,000 (1.0125)^{24} = \$1,347.40\]

You made $347 in 6 years by just letting it sit in the bank. If it sits there for 12 years, you would have

\[1,000 (1.0125)^{15 \times 4} = \$2,107.20\]

I am doing the computations with the computer, not by hand and you should do the same. Use a calculator or a computer.

However, if you were on a desert island and desired to do such computations, you could do it with the aid of the binomial theorem, taking a few terms. For example,

\[(1.0125)^{60} = 1 + 60 (.0125) + \frac{60 \times 59}{2} (.0125)^2 + \frac{60 \times 59 \times 58}{6} (.0125)^3 + \cdots\]

and so if you used just the first four terms you would be pretty close. This would yield $2,093.40 for the above problem of leaving it in the bank for 15 years.

The effective annual rate is the rate which would give the same interest for one year as the given rate compounded more often. To illustrate, here is an example.

**Example 4.4.6** A bank offers an interest rate of 5\% compounded daily. What is the effective rate?

The rate per payment period is 
\(.05/365 = 1.3699 \times 10^{-4}\). Then letting \(r\) be the effective rate you need

\[P (1 + r) = P \left(1 + \frac{.05}{365}\right)^{365} = P1.0513 = P (1 + .0513)\]

Now cancel the \(P\) on both sides and you see the effective rate is 5.13\%.
4.4.4 Annuities

Annuities deal with many equally spaced payments rather than with just one. This is where the use of the sum of a geometric sequence becomes most important and things get really interesting. Consider the following diagram which illustrates payments equal to \( P \) made at equally spaced intervals of time.

\[
\begin{array}{cccccccccc}
0 & P & P & P & P & P & P & P & P & P \\
\end{array}
\]

The 0 at the far left indicates this is at time equal to 0. The payments are made the ends of equally spaced intervals intervals of time as suggested by the picture. This situation is known as an ordinary annuity.

**Definition 4.4.7** For an ordinary annuity in which the payments are made to a bank giving an interest rate of \( r \) per payment period, the future value of the annuity after \( n \) payment periods is the amount in the bank at the end of \( n \) payment periods.

**Problem 4.4.8** Find a formula for the future value of an annuity.

Let \( A_n \) be the amount at the end of \( n \) payment periods. Then

\[
A_n = A_{n-1} (1 + r) + P
\]

This is because the amount at the end of period \( n - 1 \) sits in the bank for one payment period and grows to \( A_{n-1} (1 + r) \). Then you have to add in the payment \( P \). For an ordinary annuity, \( A_0 = 0, A_1 = P \). This is one of those difference equations discussed earlier. As explained earlier, you can find a solution in the form \( A_n = C \alpha^n + D \). Plugging this in to the recurrence relation,

\[
C \alpha^n + D = (1 + r) \left( C \alpha^{n-1} + D \right) + P
\]

and so

\[
C \alpha^n + D = C \alpha^{n-1} (1 + r) + D (1 + r) + P
\]

from the initial condition,

\[
C \alpha + D = P.
\]

You can see a solution to 4.2, 4.3 is obtained by letting

\[
\alpha = (1 + r), rD + P = 0, C = (P - D) / (1 + r)
\]

so \( D = -P/r \). Therefore,

\[
A_n = \frac{(P + P/r)}{1 + r} (1 + r)^n - \frac{P}{r}
\]

\[
= P \frac{(1 + r)^n - 1}{r}
\]

Alternatively, you could compute this by adding the following geometric series which gives the sum of the amounts in the bank corresponding to each of the payments, the sum of the future values of each payment.

\[
P (1 + r)^{n-1} + P (1 + r)^{n-2} + \ldots + P
\]

By the formula for the geometric series this is

\[
\frac{P - P (1 + r)^n}{-r} = P \frac{(1 + r)^n - 1}{r}
\]
Example 4.4.9 To save for her daughter’s education a mother places $1000 in an account which pays 5% per year compounded monthly. Thereafter, she places $100 at the end of every month in this account. How much will be in the account at the end of 20 years?

It will be the amount resulting from the ordinary annuity which comes from the $100 payments added to the amount which results from the initial payment. Thus this amount will be

\[
1000 \left(1 + \frac{.05}{12}\right)^{20 \times 12} + 100 \left(\frac{\left(1 + \frac{.05}{12}\right)^{20 \times 12} - 1}{\frac{.05}{12}}\right)
\]

The computer can compute this easily and it is

$43,816$

The following procedure for finding the future value of an annuity summarizes the above.

Procedure 4.4.10 To find the future value of an ordinary annuity as described above, consisting of \(n\) equal payments of \(P\) for money which is worth \(r\) per payment period, you compute

\[
P\left(\frac{1+r}{r}\right)^n - 1
\]

The other important concept for an ordinary annuity is its present value. This is just the sum of the present values of the payments. So what is the present value of a payment? Assuming money is worth \(r\) per payment period, how much is a payment of \(P\) worth right now if it is not obtained till \(k\) payment periods have transpired. Letting \(V\) denote this value, it should be such that

\[
P = (1 + r)^k V
\]

because this says that if you started off with \(V\) right now, you would have \(P\) at the end of \(k\) payment periods. Therefore, the present value of a payment obtained after \(k\) payment periods is given by

\[
P (1 + r)^{-k}
\]

This motivates the following definition.

Definition 4.4.11 The present value of a payment \(P\) made at the end of \(k\) payment periods where the interest rate is \(r\) per payment period is

\[
P (1 + r)^{-k}
\]

The present value of an ordinary annuity is the sum of the present values of the payments.

Thus the present value of an ordinary annuity consisting of \(n\) equal payments where the value of money is \(r\) per payment period is

\[
\sum_{k=1}^{n} P (1 + r)^{-k} = \sum_{k=1}^{n} \frac{P}{(1 + r)^k}
\]

This is again the sum of a geometric series and in this case the first term is \(P/(1 + r)\). Therefore, this sum equals

\[
\frac{P}{1+r} - \frac{P}{1+r} \left(\frac{1}{1+r}\right)^n = \frac{P - P \left(\frac{1}{1+r}\right)^n}{r}
\]
This establishes the following procedure for finding the present value of an ordinary annuity.

**Procedure 4.4.12** To find the present value of an ordinary annuity, consisting of \( n \) equal payments where the money is worth \( r \) per payment period, do the following computation.

\[
P \left( \frac{1 - (1 + r)^{-n}}{r} \right)
\]

**Example 4.4.13** A person has won \( \$10 \times 10^6 \) because of a lucky purchase of a lottery ticket. Does he have enough to place in an account paying 5% per year compounded quarterly such that he will receive \( \$10,000 \) every month for the next 30 years?

This is asking for the present value of the annuity just described in which the payments are \( \$10,000 \) because it needs to equal what those payments made in the future are worth now. What is \( n \)? It is \( 30 \times 12 \) because there are thirty years and 12 payment periods in each year. Since it is happening monthly, the interest rate is \( .05/12 \). Therefore, the present value of the annuity which gives the amount needed to obtain this income for 30 years is

\[
10000 \left( \frac{1 - (1 + \frac{.05}{12})^{-(30 \times 12)}}{\frac{.05}{12}} \right)
\]

Using the computer or calculator to compute this, you find the present value is

\[1.8628 \times 10^6,\]

less than two million dollars. Thus he can easily afford the new Bugatti he wants to buy along with a yacht.

The next example is more realistic.

**Example 4.4.14** He wants to buy a car which costs \( \$25,000 \). If money is worth 3% compounded monthly, what are the payments if he has a \( \$5,000 \) downpayment and wishes to pay off the loan in 5 years? How much interest will he pay?

After the \( \$5000 \) down payment, he has to finance \( \$20,000 \). This needs to be the present value of the annuity resulting from the monthly payments. There are \( 5 \times 12 = 60 \) payment periods. The interest rate per payment period is \( .03/12 \). Therefore, the following equation needs to hold.

\[
20000 = P \left( \frac{1 - (1 + (0.03/12))^{-60}}{(0.03/12)} \right)
\]

and you see the only thing which is not known is the payment \( P \). Thus

\[
P = \frac{20000}{\left( \frac{1-(1+(0.03/12))^{-60}}{(0.03/12)} \right)} = \$359.37
\]

The amount of interest payed must be

\[
\frac{\text{what he paid}}{\text{what he owed}} - \frac{\text{what he paid}}{\text{what he owed}} = $1,562.20
\]
4.5 Exercises

1. Let \( a_n = 2 + 3n \). Find \( \sum_{n=1}^{23} (2 + 3n) \).

2. Let \( a_k = 3 \left( \frac{-1}{2} \right)^{k+3} \). Find \( \sum_{k=1}^{n} 3 \left( \frac{-1}{2} \right)^{k+3} \).

3. They take out a 30 year fixed rate loan of 6% per year compounded monthly for an amount of $100,000. Find the payments.

4. You have found a used car which costs $7000. If you make a down payment of $1000, what are your monthly payments if you have a 3% loan and you wish to pay for the car in 3 years? How much interest will you pay during the three years?

5. A diesel car costs $27000 but gets a combined milage of 45 miles per gallon. A conventional engine in the same car gets only gets 35 miles per gallon but it only costs $22000. Suppose the car will be driven 15000 miles per year and fuel costs $3.00 per gallon for both gasoline and diesel. (Of course this is not the case because of taxes placed on diesel fuel.) Assuming there are no differences in other expenses for the two cars, about how many years will it take for the diesel car to be a better buy assuming money is worth 3%? Hint: Consider the $5000 discrepancy as the present value of an annuity, the payments being the monthly discrepancy in the fuel costs. Now vary \( n \), the number of payments, till the present value of the payments is larger than $5000. This is a calculator problem.

6. At 20% interest, how long does it take the amount of money to double? This is an old Babylonian problem, about 3700 years old, found in ancient records.

7. He takes out a loan for a house worth $Q and proposes to make payments equal to $rQ. What is the matter with this plan? What happens to the bank making the loan when the price of the house drops significantly and the man loses his job and can no longer make the payment? What happens to the bank if the price of the house stays high and the man keeps right on making payments of $rQ? Would a bank which makes such a loan be guilty of risky behavior? Hint: How long will it take to pay off the loan?

8. Suppose you expect to live 30 years after retirement and you want to have an income of $30,000 every year. How much money should you place every month in an account paying 3% per year compounded monthly for 35 years previous to retirement so that when you retire, you can withdraw $30,000 every year and not run out of money before you die? Hint: First find the present value of the $30,000 payments and then adjust the monthly payments so their future value equals the present value you just computed.

9. I can’t walk the length of a football field. Here is why. It takes me one minute to go half way, then one half a minute to go half of what is left, then 1/4 minute to go half of what is left and then 1/8 minute to go half of what is left etc. Thus the time it takes me to go the whole distance is \( 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots \), an infinite sum of positive numbers. However, reasoning more simply, if I take one minute to go half way, I will have gone the whole distance in two minutes. What is wrong? If you can’t figure out this paradox, wait till the material on limits of sequences.

10. Consider a repeating decimal. Here is an example .33333 where the bar on the last 3 indicates it repeats forever. What does this mean? Consider letting the
4.6. **THE LIMIT OF A SEQUENCE**

Consider the sequence \( a_n = \frac{1}{n} \). Then

\[
\lim_{n \to \infty} a_n = 0.
\]

This is because \( a_n \) is very close to 0 whenever \( n \) is sufficiently large. A way of saying this is that for any measure of closeness, \( a_n \) is that close to 0 for all \( n \) sufficiently large. The precise definition follows. It was first defined by Bolzano. Here it is.

---

Bernhard Bolzano lived from 1781 to 1848. He was a Catholic priest and held a position in philosophy at the University of Prague. He had strong views about the absurdity of war, educational reform, and the need for individual conscience. His convictions got him in trouble with Emperor Franz I of Austria and when he refused to recant, was forced out of the university. He understood the need for absolute rigor in mathematics. He also did work on physics.
Definition 4.6.1 A sequence \( \{a_n\}_{n=1}^{\infty} \) converges to \( a \),
\[
\lim_{n \to \infty} a_n = a \quad \text{or} \quad a_n \to a
\]
if and only if for every \( \varepsilon > 0 \) there exists \( n_\varepsilon \) such that whenever \( n \geq n_\varepsilon \),
\[
|a_n - a| < \varepsilon.
\]

In words the definition says that given any measure of closeness, \( \varepsilon \), the terms of the sequence are eventually all this close to \( a \). The above definition is always the definition of what is meant by the limit of a sequence in whatever context it occurs.

The next theorem says you can speak of the limit.

Theorem 4.6.2 If \( \lim_{n \to \infty} a_n = a \) and \( \lim_{n \to \infty} a_n = a_1 \) then \( a_1 = a \).

Proof: Suppose \( a_1 \neq a \). Then let \( 0 < \varepsilon < |a_1 - a|/2 \) in the definition of the limit. It follows there exists \( n_\varepsilon \) such that if \( n \geq n_\varepsilon \), then \( |a_n - a| < \varepsilon \) and \( |a_n - a_1| < \varepsilon \). Therefore, using the triangle inequality, Theorem 1.7.3, it follows that for such \( n \),
\[
|a_1 - a| \leq |a_1 - a_n| + |a_n - a| < \varepsilon + \varepsilon < |a_1 - a|/2 + |a_1 - a|/2 = |a_1 - a|,
\]
a contradiction.

Example 4.6.3 Let \( a_n = \frac{1}{n^2 + 1} \).

Then it seems clear that
\[
\lim_{n \to \infty} \frac{1}{n^2 + 1} = 0.
\]
In fact, this is true from the definition. Let \( \varepsilon > 0 \) be given. Let \( n_\varepsilon \geq \sqrt{\varepsilon - 1} \). Then if
\[
n > n_\varepsilon \geq \sqrt{\varepsilon - 1},
\]
it follows that \( n^2 + 1 > \varepsilon - 1 \) and so
\[
0 < \frac{1}{n^2 + 1} = a_n < \varepsilon.
\]
Thus \( |a_n - 0| < \varepsilon \) whenever \( n \) is this large.

Note the definition was of no use in finding a candidate for the limit. This had to be produced based on other considerations. The definition is for verifying beyond any doubt that something is the limit. It is also what must be referred to in establishing theorems which are good for finding limits.

Example 4.6.4 Let \( a_n = n^2 \).

Then in this case \( \lim_{n \to \infty} a_n \) does not exist. Sometimes this situation is also referred to by saying \( \lim_{n \to \infty} a_n = \infty \).

Example 4.6.5 Let \( a_n = (-1)^n \).

In this case, \( \lim_{n \to \infty} (-1)^n \) does not exist. This follows from the definition. Let \( \varepsilon = 1/2 \). If there exists a limit, \( l \), then eventually, for all \( n \) large enough, \( |a_n - l| < 1/2 \). However, \( |a_n - a_{n+1}| = 2 \) and so,
\[
2 = |a_n - a_{n+1}| \leq |a_n - l| + |l - a_{n+1}| < 1/2 + 1/2 = 1
\]
which cannot hold. Therefore, there can be no limit for this sequence.

The following is another of those theorems which is intuitively fairly clear. However, I will give a proof for those who cannot believe that which is intuitively obvious. For example, if \( a_n \) is getting close to \( a \) and \( b_n \) is getting close to \( b \) then it seems like it should be the case that \( a_n b_n \) is getting close to \( ab \). However, as you progress in math and encounter enough examples, you begin to realize that intuition is not adequate to be sure you have it right. That is why there is a proof given.

**Theorem 4.6.6** Suppose \( \{a_n\} \) and \( \{b_n\} \) are sequences and that

\[
\lim_{n \to \infty} a_n = a \quad \text{and} \quad \lim_{n \to \infty} b_n = b.
\]

Also suppose \( x \) and \( y \) are real numbers. Then

\[
\lim_{n \to \infty} xa_n + yb_n = xa + yb \tag{4.4}
\]

\[
\lim_{n \to \infty} a_n b_n = ab \tag{4.5}
\]

If \( b \neq 0 \),

\[
\lim_{n \to \infty} \frac{a_n}{b_n} = \frac{a}{b} \tag{4.6}
\]

If \( c \geq a_n \) for all \( n \) and \( \lim_{n \to \infty} a_n = a \), then \( a \leq c \). If \( c \leq a_n \) for all \( n \) and \( \lim_{n \to \infty} a_n = a \), then \( c \leq a \).

**Proof:** The first of these claims is left for you to do. To do the second, let \( \varepsilon > 0 \) be given and choose \( n_1 \) such that if \( n \geq n_1 \) then

\[
|a_n - a| < 1.
\]

Then for such \( n \), the triangle inequality implies

\[
|a_n b_n - ab| \leq |a_n b_n - a_n b| + |a_n b - ab| \\
\leq |a_n| |b_n - b| + |b| |a_n - a| \\
\leq (|a| + 1) |b_n - b| + |b| |a_n - a|.
\]

Now let \( n_2 \) be large enough that for \( n \geq n_2 \),

\[
|b_n - b| < \frac{\varepsilon}{2(|a| + 1)}, \text{ and } |a_n - a| < \frac{\varepsilon}{2(|b| + 1)}.
\]

Such a number exists because of the definition of limit. Therefore, let

\[
n_\varepsilon > \max(n_1, n_2).
\]

For \( n \geq n_\varepsilon \),

\[
|a_n b_n - ab| \leq (|a| + 1) |b_n - b| + |b| |a_n - a| \\
< (|a| + 1) \frac{\varepsilon}{2(|a| + 1)} + |b| \frac{\varepsilon}{2(|b| + 1)} \leq \varepsilon.
\]

This proves \( 4.5 \). Next consider \( 4.6 \).

Let \( \varepsilon > 0 \) be given and let \( n_1 \) be so large that whenever \( n \geq n_1 \),

\[
|b_n - b| < \frac{|b|}{2}.
\]
Thus for such $n$,

$$\left| \frac{a_n}{b_n} - \frac{a}{b} \right| = \left| \frac{a_n b - ab_n}{b b_n} \right| \leq \frac{2}{|b|^2} \left| a_n b - ab \right| + \frac{2 |a|}{|b|^2} |b_n - b|$$

$$\leq \frac{2}{|b|} |a_n - a| + \frac{2 |a|}{|b|^2} |b_n - b|.$$

Now choose $n_2$ so large that if $n \geq n_2$, then

$$|a_n - a| < \varepsilon \frac{|b|}{4} \quad \text{and} \quad |b_n - b| < \varepsilon \frac{|b|^2}{4 (|a| + 1)}.$$

Letting $n_\varepsilon = \max (n_1, n_2)$, it follows that for $n \geq n_\varepsilon$,

$$\left| \frac{a_n}{b_n} - \frac{a}{b} \right| \leq \frac{2}{|b|} |a_n - a| + \frac{2 |a|}{|b|^2} |b_n - b|$$

$$< \frac{2 \varepsilon |b|}{|b|} \left[ 1 + \frac{2 |a|}{|b|^2} \frac{\varepsilon |b|^2}{4 (|a| + 1)} \right] < \varepsilon.$$

It remains to prove the last claim. Suppose it is not true. Then $a_n \leq c$ but $a > c$. Then there exists $\varepsilon > 0$ small enough that $c < a - \varepsilon$. Then by the definition of the limit, for $n$ large enough,

$$a_n \in (a - \varepsilon, a + \varepsilon)$$

which is a contradiction. In case $a_n \geq c$ for all $n$ the argument is similar and is left for you. This proves the theorem.

Another very useful theorem for finding limits is the squeezing theorem.

**Theorem 4.6.7** Suppose $\lim_{n \to \infty} a_n = a = \lim_{n \to \infty} b_n$ and $a_n \leq c_n \leq b_n$ for all $n$ large enough. Then $\lim_{n \to \infty} c_n = a$.

**Proof:** Let $\varepsilon > 0$ be given and let $n_1$ be large enough that if $n \geq n_1$,

$$|a_n - a| < \varepsilon / 2 \quad \text{and} \quad |b_n - a| < \varepsilon / 2.$$

Then for such $n$,

$$|c_n - a| \leq |a_n - a| + |b_n - a| < \varepsilon.$$

The reason for this is that if $c_n \geq a$, then

$$|c_n - a| = c_n - a \leq b_n - a \leq |a_n - a| + |b_n - a|$$

because $b_n \geq c_n$. On the other hand, if $c_n \leq a$, then

$$|c_n - a| = a - c_n \leq a - a_n \leq |a - a_n| + |b - b_n|.$$

This proves the theorem.

As an example, consider the following.

**Example 4.6.8** Let

$$c_n \equiv (-1)^n \frac{1}{n}$$

and let $b_n = \frac{1}{n}$, and $a_n = -\frac{1}{n}$. Then you may easily show that

$$\lim_{n \to \infty} a_n = \lim_{n \to \infty} b_n = 0.$$

Since $a_n \leq c_n \leq b_n$, it follows $\lim_{n \to \infty} c_n = 0$ also.
4.6. THE LIMIT OF A SEQUENCE

Theorem 4.6.9 \( \lim_{n \to \infty} r^n = 0 \) whenever \( |r| < 1 \).

**Proof:** If \( 0 < r < 1 \) if follows \( r^{-1} > 1 \). Why? Letting \( 0 < \alpha = \frac{1}{2} - 1 \), it follows

\[
r = \frac{1}{1 + \alpha}.
\]

By the binomial theorem

\[
(1 + \alpha)^n = 1 + n\alpha + \frac{n(n-1)}{2}\alpha^2 + \cdots + \alpha^n
\]

\[
\geq 1 + n\alpha
\]

Therefore,

\[
0 < r^n = \frac{1}{(1 + \alpha)^n} \leq \frac{1}{1 + \alpha^n}.
\]

It follows, \( \lim_{n \to \infty} r^n = 0 \) if \( 0 < r < 1 \). Now in general, if \( |r| < 1 \), \( |r^n| = |r|^n \to 0 \) by the first part. This proves the theorem.

An important theorem is the one which states that if a sequence converges, so does every subsequence. You should review Definition 4.1.4 on Page 87 at this point. Recall that if \( \{x_n\} \) is a sequence and \( n_1 < n_2 < \cdots \) is an increasing sequence of integers with \( n_1 \geq 1 \), then \( \{x_{n_k}\}_{k=1}^{\infty} \) is called a subsequence.

**Theorem 4.6.10** Let \( \{x_n\} \) be a sequence with \( \lim_{n \to \infty} x_n = x \) and let \( \{x_{n_k}\} \) be a subsequence. Then \( \lim_{k \to \infty} x_{n_k} = x \).

**Proof:** Let \( \varepsilon > 0 \) be given. Then there exists \( n_\varepsilon \) such that if \( n > n_\varepsilon \), then \( |x_n - x| < \varepsilon \). Suppose \( k > n_\varepsilon \). Then \( n_k \geq k > n_\varepsilon \) and so

\[
|x_{n_k} - x| < \varepsilon
\]

showing \( \lim_{k \to \infty} x_{n_k} = x \) as claimed.

4.6.1 Sequences And Completeness

You recall the definition of completeness which stated that every nonempty set of real numbers which is bounded above has a least upper bound and that every nonempty set of real numbers which is bounded below has a greatest lower bound and this is a property of the real line known as the completeness axiom. Geometrically, this involved filling in the holes. There is another way of describing completeness in terms of sequences which I believe is more useful than the least upper bound and greatest lower bound property.

**Definition 4.6.11** \( \{a_n\} \) is a Cauchy sequence if for all \( \varepsilon > 0 \), there exists \( n_\varepsilon \) such that whenever \( n, m \geq n_\varepsilon \),

\[
|a_n - a_m| < \varepsilon.
\]

A sequence is Cauchy means the terms are “bunching up to each other” as \( m, n \) get large.

**Theorem 4.6.12** The set of terms in a Cauchy sequence in \( \mathbb{R} \) is bounded above and below.

**Proof:** Let \( \varepsilon = 1 \) in the definition of a Cauchy sequence and let \( n > n_1 \). Then from the definition,

\[
|a_n - a_{n_1}| < 1.
\]
It follows that for all \( n > n_1 \),
\[
|a_n| < 1 + |a_{n_1}|
\]
Therefore, for all \( n \),
\[
|a_n| \leq 1 + |a_{n_1}| + \sum_{k=1}^{n_1} |a_k|.
\]
This proves the theorem.

**Theorem 4.6.13** If a sequence \( \{a_n\} \) in \( \mathbb{R} \) converges, then the sequence is a Cauchy sequence.

**Proof:** Let \( \varepsilon > 0 \) be given and suppose \( a_n \to a \). Then from the definition of convergence, there exists \( n_\varepsilon \) such that if \( n > n_\varepsilon \), it follows that
\[
|a_n - a| < \frac{\varepsilon}{2}
\]
Therefore, if \( m, n \geq n_\varepsilon + 1 \), it follows that
\[
|a_n - a_m| \leq |a_n - a| + |a - a_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon
\]
showing that, since \( \varepsilon > 0 \) is arbitrary, \( \{a_n\} \) is a Cauchy sequence.

**Definition 4.6.14** The sequence, \( \{a_n\} \), is monotone increasing if for all \( n \), \( a_n \leq a_{n+1} \). The sequence is monotone decreasing if for all \( n \), \( a_n \geq a_{n+1} \).

If someone says a sequence is monotone, it usually means monotone increasing.

There exist different descriptions of the completeness axiom. If you like you can simply define any of the criteria in the following theorem to be what you mean by completeness and skip the proof. All versions of completeness involve the notion of filling in holes and they are really just different ways of expressing this idea.

In practice, it is often more convenient to use the first of the three equivalent versions of completeness in the following theorem which states that every Cauchy sequence converges. In fact, this version of completeness, although it is equivalent to the completeness axiom for the real line, also makes sense in many situations where Definition 1.11.1 on Page 26 does not make sense. For example, the concept of completeness is often needed in settings where there is no order. This happens as soon as one does multivariable calculus. From now on completeness will mean any of the three conditions in the following theorem.

**Theorem 4.6.15** The following conditions are equivalent to completeness.

1. Every Cauchy sequence converges
2. Every monotone increasing sequence which is bounded above converges.
3. Every monotone decreasing sequence which is bounded below converges.

**Proof:** Suppose every Cauchy sequence converges and let \( S \) be a non empty set which is bounded above. In what follows, \( s_n \in S \) and \( b_n \) will be an upper bound of \( S \). If, in the process about to be described, \( s_n = b_n \), this will have shown the existence of a least upper bound to \( S \). Therefore, assume \( s_n < b_n \) for all \( n \). Let \( b_1 \) be an upper bound of \( S \) and let \( s_1 \) be an element of \( S \). Suppose \( s_1, \ldots, s_n \) and \( b_1, \ldots, b_n \) have been chosen such that \( s_k \leq s_{k+1} \) and \( b_k \geq b_{k+1} \). Consider \( \frac{s_n + b_n}{2} \), the point on \( \mathbb{R} \) which is mid way between \( s_n \) and \( b_n \). If this point is an upper bound, let
\[
b_{n+1} = \frac{s_n + b_n}{2}
\]
and $s_{n+1} = s_n$. If the point is not an upper bound, let

$$s_{n+1} \in \left( \frac{s_n + b_n}{2}, b_n \right)$$

and let $b_{n+1} = b_n$. It follows this specifies an increasing sequence $\{s_n\}$ and a decreasing sequence $\{b_n\}$ such that

$$0 \leq b_n - s_n \leq 2^{-n} (b_1 - s_1).$$

Now if $n > m$,

$$0 \leq b_n - b_m = |b_n - b_m|$$

$$= \sum_{k=m}^{n-1} b_k - b_{k+1} \leq \sum_{k=m}^{n-1} b_k - s_k \leq \sum_{k=m}^{n-1} 2^{-k} (b_1 - s_1)$$

$$= \frac{2^{-m} - 2^{-n}}{2-1} (b_1 - s_1) \leq 2^{-m+1} (b_1 - s_1)$$

and $\lim_{m \to \infty} 2^{-m} = 0$ by Theorem 4.6.9. Therefore, $\{b_n\}$ is a Cauchy sequence. Similarly, $\{s_n\}$ is a Cauchy sequence. Let $l \equiv \lim_{n \to \infty} s_n$ and let $l_1 \equiv \lim_{n \to \infty} b_n$. If $n$ is large enough,

$$|l - s_n| < \varepsilon/3, |l_1 - b_n| < \varepsilon/3, \text{ and } |b_n - s_n| < \varepsilon/3.$$  

Then

$$|l - l_1| \leq |l - s_n| + |s_n - b_n| + |b_n - l_1|$$

$$< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon.$$  

Since $\varepsilon > 0$ is arbitrary, $l = l_1$. Why? Then $l$ must be the least upper bound of $S$. It is an upper bound because if there were $s > l$ where $s \in S$, then by the definition of limit, $b_n < s$ for some $n$, violating the assumption that each $b_n$ is an upper bound for $S$. On the other hand, if $l_0 < l$, then for all $n$ large enough, $s_n > l_0$, which implies $l_0$ is not an upper bound. This shows (1) implies completeness.

First note that (2) and (3) are equivalent. Why? Suppose (2) and consequently (3). Then the same construction yields the two monotone sequences, one increasing and the other decreasing. The sequence $\{b_n\}$ is bounded below by $s_m$ for all $m$ and the sequence $\{s_n\}$ is bounded above by $b_m$ for all $m$. Why? Therefore, the two sequences converge. The rest of the argument is the same as the above. Thus (2) and (3) imply completeness.

Now suppose completeness and let $\{a_n\}$ be an increasing sequence which is bounded above. Let $a$ be the least upper bound of the set of points in the sequence. If $\varepsilon > 0$ is given, there exists $n_{\varepsilon}$ such that $a - \varepsilon < a_{n_{\varepsilon}}$. Since $\{a_n\}$ is a monotone sequence, it follows that whenever $n > n_{\varepsilon}$, $a - \varepsilon < a_n \leq a$. This proves $\lim_{n \to \infty} a_n = a$ and proves convergence. Since (3) is equivalent to (2), this is also established. If follows (3) and (2) are equivalent to completeness. It remains to show that completeness implies every Cauchy sequence converges.

Suppose completeness and let $\{a_n\}$ be a Cauchy sequence. Let

$$\inf \{a_k : k \geq n\} \equiv A_n, \sup \{a_k : k \geq n\} \equiv B_n$$

Then $A_n$ is an increasing sequence while $B_n$ is a decreasing sequence and $B_n \geq A_n$. Furthermore,

$$\lim_{n \to \infty} B_n - A_n = 0.$$  

The details of these assertions are easy and are left to the reader. Also, $\{A_n\}$ is bounded below by any lower bound for the original Cauchy sequence while $\{B_n\}$ is bounded above
by any upper bound for the original Cauchy sequence. By the equivalence of completeness with \(3\) and \(2\), it follows there exists \(a\) such that \(a = \lim_{n \to \infty} A_n = \lim_{n \to \infty} B_n\). Since \(B_n \geq a_n \geq A_n\), the squeezing theorem implies \(\lim_{n \to \infty} a_n = a\) and this proves the equivalence of these characterizations of completeness.

The following theorem is a very important result.

**Theorem 4.6.16** Let \(\{a_n\}\) be a monotone increasing sequence which is bounded above. Then \(\lim_{n \to \infty} a_n = \sup \{a_n : n \geq 1\}\)

**Proof:** Let \(a = \sup \{a_n : n \geq 1\}\) and let \(\varepsilon > 0\) be given. Then from Proposition 1.11.3 on Page 27 there exists \(m\) such that \(a - \varepsilon < a_m \leq a\). Since the sequence is increasing, it follows that for all \(n \geq m\), \(a - \varepsilon < a_n \leq a\). Thus \(a = \lim_{n \to \infty} a_n\).

### 4.6.2 Decimals

You are familiar with decimals. In the United States these are written in the form \(a_1a_2a_3\cdots\) where the \(a_i\) are integers between 0 and 9.\(^3\) Thus .23417432 is a number written as a decimal. You also recall the meaning of such notation in the case of a terminating decimal. For example, .234 is defined as \(\frac{2}{10} + \frac{3}{10^2} + \frac{4}{10^3}\). Now what is meant by a nonterminating decimal?

**Definition 4.6.17** Let \(a_1a_2\cdots\) be a decimal. Define

\[
.a_1a_2\cdots \equiv \lim_{n \to \infty} \sum_{k=1}^{n} \frac{a_k}{10^k}.
\]

**Proposition 4.6.18** The above definition makes sense.

**Proof:** Note the sequence \(\left\{ \sum_{k=1}^{n} \frac{a_k}{10^k} \right\}_{n=1}^{\infty}\) is an increasing sequence. Therefore, if there exists an upper bound, it follows from Theorem 4.6.16 that this sequence converges and so the definition is well defined.

\[
\sum_{k=1}^{n} \frac{a_k}{10^k} \leq \sum_{k=1}^{n} \frac{9}{10^k} = 9 \sum_{k=1}^{n} \frac{1}{10^k}.
\]

Now

\[
\frac{9}{10} \left( \sum_{k=1}^{n} \frac{1}{10^k} \right) = \sum_{k=1}^{n} \frac{1}{10^k} - \frac{1}{10} \sum_{k=1}^{n} \frac{1}{10^k} = \sum_{k=1}^{n} \frac{1}{10^k} - \sum_{k=2}^{n+1} \frac{1}{10^k} = \frac{1}{10} - \frac{1}{10^{n+1}}
\]

and so

\[
\sum_{k=1}^{n} \frac{1}{10^k} \leq \frac{10}{9} \left( \frac{1}{10} - \frac{1}{10^{n+1}} \right) \leq \frac{10}{9} \left( \frac{1}{10} \right) = \frac{1}{9}.
\]

Therefore, since this holds for all \(n\), it follows the above sequence is bounded above. It follows the limit exists.

\(^3\)In France and Russia they use a comma instead of a period. This looks very strange but that is just the way they do it.
4.6.3 Infinite Series

The limit of the sequence of partial sums is very important so it is given a special symbol. As explained above, a repeating decimal is really a limit of a sequence of partial sums. I want to explain this a little more generally.

**Definition 4.6.19** Let \( \{a_n\}_{n=1}^{\infty} \) be a sequence. Let

\[
S_n \equiv \sum_{k=1}^{n} a_k.
\]

Then

\[
\sum_{k=1}^{\infty} a_n \equiv \lim_{n\to\infty} \sum_{k=1}^{n} a_k = \lim_{n\to\infty} S_n.
\]

This is sometimes referred to as an infinite sum. It is also called an infinite series. The infinite sum is said to exist or converge exactly when the limit of the sequence of partial sums exists.

**Example 4.6.20** Find the sum

\[
\sum_{k=1}^{100} \left( \frac{1}{2} \right)^{k-1}
\]

Then find

\[
\sum_{k=1}^{\infty} \left( \frac{1}{2} \right)^{k-1}
\]

According to the above procedure, you form

\[
\frac{\left( \frac{1}{2} \right) - \left( \frac{1}{2} \right)^{101}}{1 - \left( \frac{1}{2} \right)} = \left( 1 - \left( \frac{1}{2} \right)^{100} \right)
\]

Similarly,

\[
\sum_{k=1}^{n} \left( \frac{1}{2} \right)^{k-1} = \left( 1 - \frac{1}{2^n} \right)
\]

The limit of this sequence equals 1 by Theorem 4.6.9. More generally, here is a theorem about infinite geometric series.

**Theorem 4.6.21** Let \( |r| < 1 \). Then

\[
\sum_{n=1}^{\infty} ar^{n-1} = \frac{a}{1 - r}
\]

**Proof:** From Procedure 4.4.4,

\[
\sum_{k=1}^{n} ar^{k-1} = \frac{a - ar^n}{1 - r}
\]

and so the conclusion follows from Theorem 4.6.9. This proves the theorem.
4.7 Exercises

1. Find \( \lim_{n \to \infty} \frac{n}{3n+2} \).
2. Find \( \lim_{n \to \infty} \frac{3n^3 + 7n + 1000}{n^3 + 1} \).
3. Find \( \lim_{n \to \infty} \frac{2^n + 7(1^n)}{2^n + 2(5^n)} \).
4. Find \( \lim_{n \to \infty} \sqrt{n^2 + 6n} - n \). \textbf{Hint:} Multiply and divide by \( \sqrt{n^2 + 6n} + n \).
5. Find \( \lim_{n \to \infty} \sum_{k=1}^{n} \frac{1}{10^n} \).

6. Suppose \( x = 0.343434\ldots \) where the bar over the last 34 signifies that this repeats forever. In elementary school you were probably given the following procedure for finding the number \( x \) as a quotient of integers. First multiply by 100 to get \( 100x = 34.343434\ldots \) and then subtract to get \( 99x = 34 \). From this you conclude that \( x = 34/99 \). Fully justify this procedure. \textbf{Hint:} \( 343434 = \lim_{n \to \infty} 34 \sum_{k=1}^{n} \left( \frac{1}{10^k} \right) \).

7. If \( \lim_{n \to \infty} a_n = a \), does it follow that \( \lim_{n \to \infty} |a_n| = |a| \)? Prove or else give a counter example.

8. Suppose \( \lim_{n \to \infty} x_n = x \). Show that then \( \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} x_k = x \). Give an example where \( \lim_{n \to \infty} x_n \) does not exist but \( \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} x_k \) does.

9. Suppose \( x_n \to x \) and \( x_n \leq c \). Show that \( x \leq c \). Also show that if \( x_n \to x \) and \( x_n \geq c \), then \( x \geq c \). \textbf{Hint:} If this is not true, argue that for all \( n \) large enough \( x_n > c \).

10. What is the wrong with the following? Let \( a_{n+1} = 2a_n, a_1 = 1 \). Then letting \( a = \lim_{n \to \infty} a_n \), it follows from the limit theorems that \( a = 2a \)

so \( a = -1 \), thus contradicting the conclusion of the above problem.

11. Is the decimal expansion of a number unique? Consider \( .2 \) and \( .1999\bar{9} \) where the bar on top of the last 9 indicates this repeats forever.

12. Recall the axiom of completeness states that a set which is bounded above has a least upper bound and a set which is bounded below has a greatest lower bound. Show that a monotone decreasing sequence which is bounded below converges to its greatest lower bound. \textbf{Hint:} Let \( a \) denote the greatest lower bound and recall that because of this, it follows that for all \( \varepsilon > 0 \) there exist points of \( \{a_n\} \) in \( [a, a + \varepsilon] \).

13. Let \( a_n = (1 + \frac{1}{n})^n \). Show \( \{a_n\} \) is an increasing sequence and show also that it is bounded above by 3. By Theorem 4.6.16 this sequence converges. The number to which it converges is called \( e \). \textbf{Hint:} Use the binomial theorem. Argue that the expansion for \( a_{n+1} \) has each term larger than the corresponding term for \( a_n \) and that in addition, all the terms are positive and there are more of them for \( a_{n+1} \).

14. By Theorem 1.1.5 \( \sqrt{n} \) exists. Show \( \lim_{n \to \infty} \sqrt[n]{n} = 1 \).

\textbf{Hint:} First explain why \( \sqrt[n]{n} > 1 \). Then let \( e_n = \sqrt[n]{n} - 1 > 0 \) so that \( (1 + e_n)^n = n \).

Now use the binomial theorem. Argue \( n = (1 + e_n)^n \geq 1 + n e_n + n(n-1)e_n^2/2 \).

You finish the argument. Show \( e_n \to 0 \).
15. Suppose \( a_n \) is defined for all \( n \) less than some integer. Give the appropriate definition for
\[
\lim_{n \to -\infty} a_n
\]
Find
\[
\lim_{n \to -\infty} \frac{n}{2^n}
\]

16. Recall the following table illustrating the possible outcomes of rolling a pair of dice.

<table>
<thead>
<tr>
<th>1</th>
<th>1, 1</th>
<th>1, 2</th>
<th>1, 3</th>
<th>1, 4</th>
<th>1, 5</th>
<th>1, 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2, 1</td>
<td>2, 2</td>
<td>2, 3</td>
<td>2, 4</td>
<td>2, 5</td>
<td>2, 6</td>
</tr>
<tr>
<td>3</td>
<td>3, 1</td>
<td>3, 2</td>
<td>3, 3</td>
<td>3, 4</td>
<td>3, 5</td>
<td>3, 6</td>
</tr>
<tr>
<td>4</td>
<td>4, 1</td>
<td>4, 2</td>
<td>4, 3</td>
<td>4, 4</td>
<td>4, 5</td>
<td>4, 6</td>
</tr>
<tr>
<td>5</td>
<td>5, 1</td>
<td>5, 2</td>
<td>5, 3</td>
<td>5, 4</td>
<td>5, 5</td>
<td>5, 6</td>
</tr>
<tr>
<td>6</td>
<td>6, 1</td>
<td>6, 2</td>
<td>6, 3</td>
<td>6, 4</td>
<td>6, 5</td>
<td>6, 6</td>
</tr>
</tbody>
</table>

Find the probability that you roll a 7 before you roll either a 3 or an 11. **Hint:** It can happen in infinitely many distinct ways. You don’t roll either a 3 or an 11 for \( k \) rolls and then on the \( k \)th roll you get a 7. Here \( k = 0, 1, 2, 3, \ldots \) so you need to take a limit of the partial sums associated with the different values of \( k \) and then take a limit. So what is the probability of getting a 7 on try \( k + 1 \) and not getting either a 3 or an 11 before this? Argue it is \( \left( \frac{13}{18} \right)^k \frac{1}{6} \).

17. Show \( \sum_{n=2}^{\infty} \frac{1}{n(n-1)} \) exists. **Hint:** Use partial fractions to estimate the partial sums.

18. Show \( \sum_{n=1}^{\infty} \frac{1}{n^2} \) exists. **Hint:**
\[
\sum_{k=1}^{n} \frac{1}{k^2} = 1 + \sum_{k=2}^{n} \frac{1}{k^2} \leq 1 + \sum_{k=2}^{n} \frac{1}{k(k-1)}
\]
Now use Problem 17 to show the partial sums are bounded above. Next explain why these partial sums are increasing and use Theorem 4.6.16.

19. Explain why if \( p \geq 2 \), the sum
\[
\sum_{n=1}^{\infty} \frac{1}{n^p}
\]
must converge (exist).

20. Does \( \sum_{k=1}^{\infty} \frac{1}{k} \) converge? **Hint:**
\[
1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} + \cdots
\]
\[
\geq 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} + \cdots
\]
Show a large enough partial sum exceeds \( 1 + \frac{k}{2} \) for any positive integer \( k \).
Basic Geometry And Trigonometry

This section is a review some basic geometry which is especially useful in the study of calculus. The purpose here is not to give a complete treatment of plane geometry, just a suitable introduction. To do this right, you should consult the books of Euclid written about 300 B.C.

5.1 Similar Triangles And Pythagorean Theorem

Definition 5.1.1 Two triangles are similar if they have the same angles. For example, in the following picture, the two triangles are similar because the angles are the same.

\[ \frac{a}{b} = \frac{a^*}{b^*} \]

The fundamental axiom for similar triangles is the following.

Axiom 5.1.2 If two triangles are similar then the ratios of corresponding parts are the same.

Definition 5.1.3 Two lines in the plane are said to be parallel if no matter how far they are extended, they never intersect.

Definition 5.1.4 If two lines \( l_1 \) and \( l_2 \) are parallel and if they are intersected by a line, \( l_3 \), the alternate interior angles are shown in the following picture labeled as \( \alpha \).
As suggested by the above picture, the following axiom will be used.

**Axiom 5.1.5** If $l_1$ and $l_2$ are parallel lines intersected by $l_3$, then alternate interior angles are equal.

**Definition 5.1.6** An angle is a right angle if when either side is extended, the new angle formed by the extension equals the original angle.

**Axiom 5.1.7** Suppose $l_1$ and $l_2$ both intersect a third line, $l_3$ in a right angle. Then $l_1$ and $l_2$ are parallel.

**Definition 5.1.8** A right triangle is one in which one of the angles is a right angle.

**Axiom 5.1.9** Given a straight line and a point, there exists a straight line which contains the point and intersects the given line in two right angles. This line is called perpendicular to the given line.

**Theorem 5.1.10** Let $\alpha$, $\beta$, and $\gamma$ be the angles of a right triangle with $\gamma$ the right angle. Then if the angles, $\alpha$ and $\beta$ are placed next to each other, the resulting angle is a right angle.

**Proof:** Consider the following picture.

In the picture the top horizontal line is obtained from Axiom 5.1.9. It is a line perpendicular to the line determined by the line segment joining $B$ and $C$ which passes through the point, $B$. Thus from Axiom 5.1.7 this line is parallel to the line joining $A$ and $B$ and by Axiom 5.1.5 the angle between the line joining $A$ and $B$ and this new line is $\alpha$ as shown in the picture. Therefore, the angle formed by placing $\alpha$ and $\beta$ together is a right angle as claimed.

**Definition 5.1.11** When an angle $\alpha$ is placed next to an angle $\beta$ as shown above, then the resulting angle is denoted by $\alpha + \beta$. A right angle is said to have $90^\circ$ or to be a $90^\circ$ angle.

With this definition, Theorem 5.1.10 says the sum of the two non $90^\circ$ angles in a right triangle is $90^\circ$.

In a right triangle the long side is called the hypotenuse. The similar triangles axiom can be used to prove the Pythagorean theorem.

**Theorem 5.1.12** (Pythagoras) In a right triangle the square of the length of the hypotenuse equals the sum of the squares of the lengths of the other two sides.
**Proof:** Consider the following picture in which the large triangle is a right triangle and $D$ is the point where the line through $C$ perpendicular to the line from $A$ to $B$ intersects the line from $A$ to $B$. Then $c$ is defined to be the length of the line from $A$ to $B$, $a$ is the length of the line from $B$ to $C$, and $b$ is the length of the line from $A$ to $C$. Denote by $DB$ the length of the line from $D$ to $B$.

Then from Theorem 5.1.10, $\delta + \gamma = 90^\circ$ and $\beta + \gamma = 90^\circ$. Therefore, $\delta = \beta$. Also from this same theorem, $\alpha + \delta = 90^\circ$ and so $\alpha = \gamma$. Therefore, the three triangles shown in the picture are all similar. By Axiom 5.1.2,

$$\frac{c}{a} = \frac{a}{DB}, \text{ and } \frac{c}{b} = \frac{b}{c - DB}.$$ 

Therefore, $cDB = a^2$ and

$$\quad c(c - DB) = b^2$$

so

$$c^2 = cDB + b^2 = a^2 + b^2.$$ 

This proves the Pythagorean theorem. [1]

This theorem implies there should exist some such number which deserves to be called $\sqrt{a^2 + b^2}$ as mentioned earlier in the discussion on completeness of $\mathbb{R}$.

### 5.2 Distance Formula And Trigonometric Functions

As just explained, points in the plane may be identified by giving a pair of numbers. Suppose there are two points in the plane and it is desired to find the distance between them. There are actually many ways used to measure this distance but the best way, and the only way used in this book is determined by the Pythagorean theorem. Consider the following picture.

---

[1] This theorem is due to Pythagoras who lived about 572-497 B.C. This was during the Babylonian captivity of the Jews. Thus Pythagoras was probably a contemporary of the prophet Daniel, sometime before Ezra and Nehemiah. Alexander the great would not come along for more than 100 years. There was, however, an even earlier Greek mathematician named Thales, 624-547 B.C. who also did fundamental work in geometry. Greek geometry was organized and published by Euclid about 300 B.C.
In this picture, the distance between the points denoted by \((x_0, y_0)\) and \((x_1, y_1)\) should be the square root of the sum of the squares of the lengths of the two sides. The length of the side on the bottom is \(|x_0 - x_1|\) while the length of the side on the right is \(|y_0 - y_1|\). Therefore, by the Pythagorean theorem the distance between the two indicated points is \(\sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}\). Note you could write \(\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}\) or even \(\sqrt{(x_0 - x_1)^2 + (y_1 - y_0)^2}\) and it would make no difference in the resulting number. The distance between the two points is written as \(|(x_0, y_0) - (x_1, y_1)|\) or sometimes when \(P_0\) is the point determined by \((x_0, y_0)\) and \(P_1\) is the point determined by \((x_1, y_1)\), as \(d(P_0, P_1)\) or \(|P_0P|\).

The trigonometric functions \(\cos\) and \(\sin\) are defined next. Consider the following picture in which the small circle has radius 1, the large circle has radius \(R\), and the right side of each of the two triangles is perpendicular to the bottom side which lies on the \(x\) axis.

By Theorem 5.1.10 on Page 112 the two triangles have the same angles and so they are similar. Now define by \((\cos \theta, \sin \theta)\) the coordinates of the top vertex of the smaller triangle. Therefore, it follows the coordinates of the top vertex of the larger triangle are as shown. This shows the following definition is well defined.

**Definition 5.2.1** For \(\theta\) an angle, define \(\cos \theta\) and \(\sin \theta\) as follows. Place the vertex of the angle (The vertex is the point.) at the point whose coordinates are \((0, 0)\) in such a way that one side of the angle lies on the positive \(x\) axis and the other side extends upward. Extend this other side until it intersects a circle of radius \(R\). Then the point of
5.2. DISTANCE FORMULA AND TRIGONOMETRIC FUNCTIONS

intersection, is given as \((R \cos \theta, R \sin \theta)\). In particular, this specifies \(\cos \theta\) and \(\sin \theta\) by simply letting \(R = 1\).

**Proposition 5.2.2** For any angle, \(\theta\), \(\cos^2 \theta + \sin^2 \theta = 1\).

**Proof:** This follows immediately from the above definition and the distance formula. Since \((\cos \theta, \sin \theta)\) is a point on the circle which has radius 1, the distance of this point to \((0, 0)\) equals 1. Thus the above identity holds.

The other trigonometric functions are defined as follows.

\[
\tan \theta \equiv \frac{\sin \theta}{\cos \theta}, \quad \cot \theta \equiv \frac{\cos \theta}{\sin \theta}, \quad \sec \theta \equiv \frac{1}{\cos \theta}, \quad \csc \theta \equiv \frac{1}{\sin \theta}.
\] (5.1)

It is also important to understand these functions in terms of a right triangle. Consider the following picture of a right triangle.

\[
\begin{array}{c}
A \\
\downarrow \quad \downarrow \\
B \\
\downarrow \quad \downarrow \\
C
\end{array}
\]

You should verify \(\sin A \equiv a/c, \cos A \equiv b/c, \tan A \equiv a/b, \sec A \equiv c/b,\) and \(\csc A \equiv c/a\).

Having defined the cos and sin there is a very important generalization of the Pythagorean theorem known as the law of cosines. Consider the following picture of a triangle in which \(a, b\) and \(c\) are the lengths of the sides and \(A, B,\) and \(C\) denote the angles indicated.

The law of cosines is the following.

**Theorem 5.2.3** Let \(ABC\) be a triangle as shown above. Then

\[
c^2 = a^2 + b^2 - 2ab \cos C
\]

**Proof:** Situate the triangle so the vertex of the angle, \(C\), is on the point whose coordinates are \((0, 0)\) and so the side opposite the vertex, \(B\) is on the positive \(x\) axis as shown in the above picture. Then from the definition of the \(\cos C\), the coordinates of the vertex, \(B\) are \((a \cos C, a \sin C)\) while it is clear the coordinates of \(A\) are \((b, 0)\). Therefore, from the distance formula, and Proposition 5.2.2

\[
c^2 = (a \cos C - b)^2 + a^2 \sin^2 C
= a^2 \cos^2 C - 2ab \cos C + b^2 + a^2 \sin^2 C
= a^2 + b^2 - 2ab \cos C
\]
Corollary 5.2.4 Let \(ABC\) be any triangle as shown above. Then the length of any side is no longer than the sum of the lengths of the other two sides.

**Proof:** This follows immediately from the law of cosines. From Proposition 5.2.2 \( |\cos \theta| \leq 1 \) and so \( c^2 = a^2 + b^2 - 2ab \cos C \leq a^2 + b^2 + 2ab = (a + b)^2 \). This proves the corollary.

Corollary 5.2.5 Suppose \(T\) and \(T'\) are two triangles such that one angle is the same in the two triangles and in each triangle, the sides forming that angle are equal. Then the corresponding sides are proportional.

**Proof:** Let \(T = ABC\) with the two equal sides being \(AC\) and \(AB\). Let \(T'\) be labeled in the same way but with primes on the letters. Thus the angle at \(A\) is equal to the angle at \(A'\). The following picture is descriptive of the situation.

Denote by \(a, a', b, b', c\) and \(c'\) the sides indicated in the picture. Then by the law of cosines,

\[
a^2 = b^2 + c^2 - 2bc \cos A
\]

\[
= 2b^2 - 2b^2 \cos A
\]

and so \(a/b = \sqrt{2(1 - \cos A)}\). Similar reasoning shows \(a'/b' = \sqrt{2(1 - \cos A)}\) and so \(a/b = a'/b'\).

Similarly, \(a/c = a'/c'\). By assumption \(c/b = 1 = c'/b'\).

Such triangles in which two sides are equal are called isosceles.

### 5.3 The Circular Arc Subtended By An Angle

How can angles be measured? This will be done by considering arcs on a circle. To see how this will be done, let \(\theta\) denote an angle and place the vertex of this angle at the center of the circle. Next, extend its two sides till they intersect the circle. Note the angle could be opening in any of infinitely many different directions. Thus this procedure could yield any of infinitely many different circular arcs. Each of these arcs is said to subtend the angle. In fact each of these arcs has the same length. When this has been shown, it will be easy to measure angles. Angles will be measured in terms of lengths of arcs subtended by the angle. Of course it is also necessary to define what is meant by the length of a circular arc in order to do any of this. First I will describe an intuitive way of thinking about this and then give a rigorous definition and proof. If the intuitive way of thinking about this satisfies you, no harm will be done by skipping the more technical discussion which follows.
Take an angle and place its vertex (the point) at the center of a circle of radius \( r \). Then, extending the sides of the angle if necessary till they intersect the circle, this determines an arc on the circle. If \( r \) were changed to \( R \), this really amounts to a change of units of length. Think, for example, of keeping the numbers the same but changing centimeters to meters in order to produce an enlarged version of the same picture. Thus the picture looks exactly the same, only larger. It is reasonable to suppose, based on this reasoning that the way to measure the angle is to take the length of the arc subtended in whatever units being used and divide this length by the radius measured in the same units, thus obtaining a number which is independent of the units of length used, just as the angle itself is independent of units of length. After all, it is the same angle regardless of how far its sides are extended. This is in fact how to define the radian measure of an angle and the definition is well defined. Thus, in particular, the ratio between the circumference (length) of a circle and its radius is a constant which is independent of the radius of the circle.\(^2\) Since the time of Euler in the 1700’s, this constant has been denoted by \( 2\pi \). In summary, if \( \theta \) is the radian measure of an angle, the length of the arc subtended by the angle on a circle of radius \( r \) is \( r\theta \).

This is a little sloppy right now because no precise definition of the length of an arc of a circle has been given. For now, imagine taking a string, placing one end of it on one end of the circular arc and then wrapping the string till you reach the other end of the arc. Stretching this string out and measuring it would then give you the length of the arc. Such intuitive discussions involving string may or may not be enough to convey understanding. If you need to see more discussion, read on. Otherwise, skip to the next section.

To give a precise description of what is meant by the length of an arc, consider the following picture.

In this picture, there are two circles, a big one having radius, \( R \) and a little one having radius \( r \). The angle, \( \theta \) is situated in two different ways subtending the arcs \( A_1 \) and \( A_2 \) as shown.

Letting \( A \) be an arc of a circle, like those shown in the above picture, \( A \) is a subset of \( A, \{p_0, \cdots, p_n\} \) is a partition of \( A \) if \( p_0 \) is one endpoint, \( p_n \) is the other end point, and the points are encountered in the indicated order as one moves in the counter clockwise direction along the arc. To illustrate, see the following picture.

\(^2\)In 2 Chronicles 4:2 the “molten sea” used for “washing” by the priests and found in Solomon’s temple is described. It sat on 12 oxen, was round, 5 cubits high, 10 across and 30 around. Thus the Bible, taken literally, gives the value of \( \pi \) as 3. This is not too far off. Later, methods will be given which allow one to calculate \( \pi \) more precisely. A better value is 3.1415926535 and presently this number is known to thousands of decimal places. It was proved by Lindeman in the 1880’s that \( \pi \) is transcendental which is the worst sort of irrational number.
Also, denote by \( \mathcal{P}(A) \) the set of all such partitions. For \( P = \{p_0, \cdots, p_n\} \), denote by \( |p_i - p_{i-1}| \) the distance between \( p_i \) and \( p_{i-1} \). Then for \( P \in \mathcal{P}(A) \), define \( |P| \equiv \sum_{i=1}^{n} |p_i - p_{i-1}| \). Thus \( |P| \) consists of the sum of the lengths of the little lines joining successive points of \( P \) and appears to be an approximation to the length of the circular arc, \( A \). By Corollary \[5.2.4\] the length of any of the straight line segments joining successive points in a partition is smaller than the sum of the two sides of a right triangle having the given straight line segment as its hypotenuse. For example, see the following picture.

![Diagram](image)

The sum of the lengths of the straight line segments in the part of the picture found in the right rectangle above is less than \( A + B \) and the sum of the lengths of the straight line segments in the part of the picture found in the left rectangle above is less than \( C + D \) and this would be so for any partition. Therefore, for any \( P \in \mathcal{P}(A) \), \( |P| \leq M \) where \( M \) is the perimeter of a rectangle containing the arc, \( A \). To be a little sloppy, simply pick \( M \) to be the perimeter of a rectangle containing the whole circle of which \( A \) is a part. The only purpose for doing this is to obtain the existence of an upper bound. Therefore, \( \{|P| : P \in \mathcal{P}(A)\} \) is a set of numbers which is bounded above by \( M \) and by completeness of \( \mathbb{R} \) it is possible to define the length of \( A \), \( l(A) \), by \( l(A) \equiv \sup \{|P| : P \in \mathcal{P}(A)\} \).

A fundamental observation following from Corollary \[5.2.4\] is that if \( P, Q \in \mathcal{P}(A) \) and \( P \subseteq Q \), then \( |P| \leq |Q| \). To see this, add in one point at a time to \( P \). This effect of adding in one point is illustrated in the following picture.
Also, letting \( \{p_0, \cdots, p_n\} \) be a partition of \( A \), specify angles, \( \theta_i \) as follows. The angle \( \theta_i \) is formed by the two lines, one from the center of the circle to \( p_i \) and the other line from the center of the circle to \( p_{i-1} \). Furthermore, a specification of these angles yields the partition of \( A \) in the following way. Place the vertex of \( \theta_1 \) on the center of the circle, letting one side lie on the line from the center of the circle to \( p_0 \) and the other side extended resulting in a point further along the arc in the counter clockwise direction. When the angles, \( \theta_1, \cdots, \theta_{i-1} \) have produced points, \( p_0, \cdots, p_{i-1} \) on the arc, place the vertex of \( \theta_i \) on the center of the circle and let one side of \( \theta_i \) coincide with the side of the angle \( \theta_{i-1} \) which is most counter clockwise, the other side of \( \theta_i \) when extended, resulting in a point further along the arc, \( A \) in the counterclockwise direction as shown below.

Now let \( \varepsilon > 0 \) be given and pick \( P_1 \in P(A_1) \) such that \( |P_1| + \varepsilon > l(A_1) \). Then determining the angles as just described, use these angles to produce a corresponding partition of \( A_2 \), \( P_2 \). If \( |P_2| + \varepsilon > l(A_2) \), then stop. Otherwise, pick \( Q \in P(A_2) \) such that \( |Q| + \varepsilon > l(A_2) \) and let \( P'_2 = P_2 \cup Q \). Then use the angles determined by \( P'_2 \) to obtain \( P'_1 \in P(A_1) \). Then \( |P'_1| + \varepsilon > l(A_1), |P'_2| + \varepsilon > l(A_2) \), and both \( P'_1 \) and \( P'_2 \) determine the same sequence of angles. Using Corollary 5.2.5

\[
\frac{|P'_1|}{|P'_2|} = \frac{R}{r}
\]

and so

\[
l(A_2) < |P'_2| + \varepsilon = \frac{R}{r} |P'_1| + \varepsilon \leq \frac{R}{r} l(A_1) + \varepsilon.
\]

Since \( \varepsilon \) is arbitrary, this shows \( Rl(A_2) \leq rl(A_1) \). But now reverse the argument and write

\[
l(A_1) < |P'_1| + \varepsilon = \frac{R}{r} |P'_2| + \varepsilon \leq \frac{R}{r} l(A_2) + \varepsilon
\]

which implies, since \( \varepsilon \) is arbitrary that \( Rl(A_2) \geq rl(A_1) \) and this has proved the following theorem.

**Theorem 5.3.1** Let \( \theta \) be an angle which subtends two arcs, \( A_R \) on a circle of radius \( R \) and \( A_r \) on a circle of radius \( r \). Then denoting by \( l(A) \) the length of a circular arc as described above, \( Rl(A_r) = rl(A_R) \).

Before proceeding further, note the proof of the above theorem involved showing \( l(A_1) < \frac{R}{r} l(A_2) + \varepsilon \) where \( \varepsilon > 0 \) was arbitrary and from this, the conclusion that
l(A_1) \leq \frac{R}{2} l(A_2). This is a very typical way of showing one number is no larger than another. To show \( a \leq b \) first show that for every \( \varepsilon > 0 \) it follows that \( a < b + \varepsilon \). This implies \( a - b < \varepsilon \) for all positive \( \varepsilon \) and so it must be the case that \( a - b \leq 0 \) since otherwise, you could take \( \varepsilon = \frac{a-b}{2} \) and conclude \( 0 < a - b < \frac{a-b}{2} \), a contradiction.

With this preparation, here is the definition of the measure of an angle.

**Definition 5.3.2** Let \( \theta \) be an angle. The measure of \( \theta \) is defined to be the length of the circular arc subtended by \( \theta \) on a circle of radius \( r \) divided by \( r \). This is also called the radian measure of the angle. In this way, one degree consists of an angle which subtends an arc which goes \( 1/360 \) of the way around the circle. The measure of the angle consists of the number of degrees which correspond to the given angle. Thus \( 45^\circ \), read 45 degrees corresponds to an angle which subtends an arc which goes \( 1/8 \) the way around a circle. The correct way to describe this angle is in terms of the radian measure of \( \pi/4 \). Nevertheless, people continue to use this way of measuring angles so you need to be aware of it\(^3\). When you look on a map, lines of latitude are measured in terms of degrees from the equator. This way of measuring angles is not the right one for calculus but is often very convenient. The degree is divided into 60 minutes and these are divided into 60 seconds. (Minutes are denoted by placing a prime on the number of minutes. For example \( 45^\prime \). ) Seconds are denoted by two primes. Note the earth has a diameter of about 4000 miles and so its circumference is about \( 8000 \pi = 25132.7412 \) miles. Thus a degree corresponds to about \( \frac{25132}{360} = 69.8 \) miles and a minute is about \( \frac{69.8}{60} = 1.163 \) miles. The arc on the earth corresponding to a minute is sometimes called a knot.

You should note the radian measure of \( \theta \) is independent of dimension. This is because the units of length cancel when the division takes place.

**Proposition 5.3.3** The above definition is well defined and also, if \( A \) is an arc subtended by the angle \( \theta \) on a circle of radius \( r \) then the length of \( A \), denoted by \( l(A) \) is given by \( l(A) = r \theta \).

**Proof:** That the definition is well defined follows from Theorem 5.3.1. The formula also follows from Theorem 5.3.1 and letting \( R = 1 \).

Now is a good time to present a useful inequality which may or may not be self evident. Here is a picture which illustrates the conclusion of this corollary.

The following corollary states that the length of the subtended arc shown in the picture is longer than the vertical side of the triangle and smaller than the sum of the vertical side with the segment having length \( 1 - \cos \theta \). To me, this seems abundantly clear but in case it is hard to believe, the following corollary gives a proof.

\(^3\)This way of measuring angles originated with the ancient Babylonians. They liked to do things in terms of 60.
5.4. The Trigonometric Functions

Now the Trigonometric functions will be defined as functions of an arbitrary real variable. Up till now these have been defined as functions of pointy things called angles. The following theorem will make possible the definition.

**Theorem 5.4.1** Let \( b \in \mathbb{R} \). Then there exists a unique integer \( p \) and real number \( r \) such that \( 0 \leq r < 2\pi \) and \( b = p2\pi + r \).

**Proof:** First suppose \( b \geq 0 \). Then from Theorem 1.9.11 on Page 23 there exists a unique integer, \( p \) such that \( b = 2\pi p + r \) where \( 0 \leq r < 2\pi \). Now suppose \( b < 0 \). Then there exists a unique integer, \( p \) such that \( -b = 2\pi p + r_1 \) where \( r_1 \in [0, 2\pi) \). If \( r_1 = 0 \), then \( b = (-p)2\pi \). Otherwise, \( b = (-p)2\pi + (-r_1) = (-p - 1)2\pi + \frac{2\pi - r_1}{r} \) and \( r \equiv 2\pi - r_1 \in (0, 2\pi) \).

The following definition is for \( \sin b \) and \( \cos b \) for \( b \in \mathbb{R} \).

**Definition 5.4.2** Let \( b \in \mathbb{R} \). Then \( \sin b \equiv \sin r \) and \( \cos b \equiv \cos r \) where \( b = 2\pi p + r \) for \( p \) an integer, and \( r \in [0, 2\pi) \).

Several observations are now obvious from this.

**Observation 5.4.3** Let \( k \in \mathbb{Z} \), then the following formulas hold.

\[
\sin b = -\sin (-b), \quad \cos b = \cos (-b), \quad (5.3)
\]

\[
\sin (b + 2k\pi) = \sin b, \quad \cos (b + 2k\pi) = \cos b \quad (5.4)
\]

\[
\cos^2 b + \sin^2 b = 1 \quad (5.5)
\]

The other trigonometric functions are defined in the usual way as in [5.1] provided they make sense.

From the observation that the \( x \) and \( y \) axes intersect at right angles the four arcs on the unit circle subtended by these axes are all of equal length. Therefore, the measure of a right angle must be \( 2\pi/4 = \pi/2 \). The measure of the angle which is determined by the arc from \((1, 0)\) to \((-1, 0)\) is seen to equal \( \pi \) by the same reasoning. From the definition of the trig functions, \( \cos(\pi/2) = 0 \) and \( \sin(\pi/2) = 1 \). You can easily find other values for \( \cos \) and \( \sin \) at all the other multiples of \( \pi/2 \).

The next topic is the important formulas for the trig. functions of sums and differences of numbers. For \( b \in \mathbb{R} \), denote by \( r_b \) the element of \([0, 2\pi)\) having the property that \( b = 2\pi p + r_b \) for \( p \) an integer.
Lemma 5.4.4 Let \( x, y \in \mathbb{R} \). Then \( r_{x+y} = r_x + r_y + 2k\pi \) for some \( k \in \mathbb{Z} \).

**Proof:** By definition,

\[
x + y = 2\pi p + r_{x+y}, \quad x = 2\pi p_1 + r_x, \quad y = 2\pi p_2 + r_y.
\]

From this the result follows because

\[
0 = ((x + y) - x) - y = 2\pi((p - p_1) - p_2) + r_{x+y} - (r_x + r_y).
\]

Let \( z \in \mathbb{R} \) and let \( p(z) \) denote the point on the unit circle determined by the length \( r_z \) whose coordinates are \( \cos z \) and \( \sin z \). Thus, starting at \((1,0)\) and moving counter clockwise a distance of \( r_z \) on the unit circle yields \( p(z) \). Note also that \( p(z) = p(r_z) \).

Lemma 5.4.5 Let \( x, y \in \mathbb{R} \). Then the length of the arc between \( p(x+y) \) and \( p(x) \) is equal to the length of the arc between \( p(y) \) and \((1,0)\).

**Proof:** The length of the arc between \( p(x+y) \) and \( p(x) \) is \( |r_{x+y} - r_x| \). There are two cases to consider here.

First assume \( r_{x+y} \geq r_x \). Then \( |r_{x+y} - r_x| = r_{x+y} - r_x = r_y + 2k\pi \). Since both \( r_{x+y} \) and \( r_x \) are in \([0, 2\pi]\), their difference is also in \([0, 2\pi]\) and so \( k = 0 \). Therefore, the arc joining \( p(x) \) and \( p(x+y) \) is of the same length as the arc joining \( p(y) \) and \((1,0)\). In the other case, \( r_{x+y} < r_x \) and in this case \( |r_{x+y} - r_x| = |r_x - r_{x+y}| = -r_y - 2k\pi \). Since \( r_x \) and \( r_{x+y} \) are both in \([0, 2\pi]\) their difference is also in \([0, 2\pi]\) and so in this case \( k = -1 \). Therefore, in this case, \( |r_{x+y} - r_x| = 2\pi - r_y \). Now since the circumference of the unit circle is \( 2\pi \), the length of the arc joining \( p(2\pi - r_y) \) to \((1,0)\) is the same as the length of the arc joining \( p(r_y) = p(y) \) to \((1,0)\). This proves the lemma.

The following theorem is the fundamental identity from which all the major trig. identities involving sums and differences of angles are derived.

**Theorem 5.4.6** Let \( x, y \in \mathbb{R} \). Then

\[
\cos(x+y)\cos y + \sin(x+y)\sin y = \cos x. \tag{5.6}
\]

**Proof:** Recall that for a real number, \( z \), there is a unique point, \( p(z) \) on the unit circle and the coordinates of this point are \( \cos z \) and \( \sin z \). Now from the above lemma, the length of the arc between \( p(x+y) \) and \( p(x) \) has the same length as the arc between \( p(y) \) and \( p(0) \). As an illustration see the following picture.

![Diagram](image)

It follows from the definition of the radian measure of an angle that the two angles determined by these arcs are equal and so, by Corollary 5.2.5, the distance between the
points \( p(x+y) \) and \( p(x) \) must be the same as the distance from \( p(y) \) to \( p(0) \). Writing this in terms of the definition of the trig functions and the distance formula,

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

\[
\cos^2(x+y) + \cos^2 x - 2 \cos(x+y) \cos x + \sin^2(x+y) + \sin^2 x - 2 \sin(x+y) \sin x \\
= \cos^2 y - 2 \cos y + 1 + \sin^2 y
\]

From Observation 5.4.3 this implies 5.6. This proves the theorem.

Letting \( y = \pi/2 \), this shows that

\[
\sin(x + \pi/2) = \cos x.
\] (5.7)

Now let \( u = x + y \) and \( v = y \). Then 5.6 implies

\[
\cos u \cos v + \sin u \sin v = \cos (u - v)
\] (5.8)

Also, from this and 5.3,

\[
\cos(u + v) = \cos(u - (-v)) \\
= \cos u \cos(-v) + \sin u \sin(-v) \\
= \cos u \cos v - \sin u \sin v
\] (5.9)

Thus, letting \( v = \pi/2 \),

\[
\cos\left(u + \frac{\pi}{2}\right) = -\sin u.
\] (5.10)

It follows

\[
\sin(x + y) = -\cos\left(x + \frac{\pi}{2} + y\right) \\
= -\left[\cos\left(x + \frac{\pi}{2}\right) \cos y - \sin\left(x + \frac{\pi}{2}\right) \sin y\right] \\
= \sin x \cos y + \sin y \cos x
\] (5.11)

Then using Observation 5.4.3 again, this implies

\[
\sin(x - y) = \sin x \cos y - \cos x \sin y.
\] (5.12)

In addition to this, Observation 5.4.3 implies

\[
\cos 2x = \cos^2 x - \sin^2 x \\
= 2 \cos^2 x - 1 \\
= 1 - 2 \sin^2 x
\] (5.13) (5.14) (5.15)

Therefore, making use of the above identities, and Observation 5.4.3

\[
\cos(3x) = \cos 2x \cos x - \sin 2x \sin x \\
= (2 \cos^2 x - 1) \cos x - 2 \cos x \sin^2 x \\
= 4 \cos^3 x - 3 \cos x
\] (5.16)

With these fundamental identities, it is easy to obtain the cosine and sine of many special angles, called reference angles. First, \( \cos\left(\frac{\pi}{4}\right) \).

\[
0 = \cos\left(\frac{\pi}{2}\right) = \cos\left(\frac{\pi}{4} + \frac{\pi}{4}\right) = 2 \cos^2\left(\frac{\pi}{4}\right) - 1
\]
and so \( \cos \left( \frac{\pi}{4} \right) = \sqrt{2}/2 \). (Why does it’s not equal to \( -\sqrt{2}/2 \)? \textbf{Hint:} Draw a picture.) Thus \( \sin \left( \frac{\pi}{4} \right) = \sqrt{2}/2 \) also. (Why?) Here is another one. From 5.16,

\[
0 = \cos \left( \frac{\pi}{2} \right) = \cos 3 \left( \frac{\pi}{6} \right) = 4 \cos^3 \left( \frac{\pi}{6} \right) - 3 \cos \left( \frac{\pi}{6} \right).
\]

Therefore, \( \cos \left( \frac{\pi}{6} \right) = \sqrt{3}/2 \) and consequently, \( \sin \left( \frac{\pi}{6} \right) = 1/2 \). Here is a short table including these and a few others. You should make sure you can obtain all these entries. In the table, \( \theta \) refers to the radian measure of the angle. From now on, angles are considered as real numbers, not as pointy things.

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>0</th>
<th>( \frac{\pi}{6} )</th>
<th>( \frac{\pi}{4} )</th>
<th>( \frac{\pi}{3} )</th>
<th>( \frac{\pi}{2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \cos \theta )</td>
<td>1</td>
<td>( \sqrt{3}/2 )</td>
<td>( \sqrt{2}/2 )</td>
<td>( 1/2 )</td>
<td>0</td>
</tr>
<tr>
<td>( \sin \theta )</td>
<td>0</td>
<td>( 1/2 )</td>
<td>( \sqrt{2}/2 )</td>
<td>( \sqrt{3}/2 )</td>
<td>1</td>
</tr>
</tbody>
</table>

Here is a picture which describes the above table.

You only need to memorize the cosines and sines of angles in the first quadrant, the one for which the values are given in the above picture in which both \( x \) and \( y \) are nonnegative. You can figure out the others from this. For example, \( \sin \left( \frac{2\pi}{3} \right) = \sqrt{3}/2 \) and \( \cos \left( \frac{2\pi}{3} \right) = -1/2 \). You can see this from the picture right away. Similarly, \( \cos \left( \frac{5\pi}{6} \right) = -\sqrt{3}/2 \).

Another thing to notice is that if \( k \) is any integer, \( \theta + 2k\pi \) corresponds to the same point on the unit circle as \( \theta \) and so the sines and cosines of the two angles are the same. In particular, the sines and cosines of, for example, \( \frac{2\pi}{3} \) and \( -\frac{4\pi}{6} \) are the same because \( 5\pi - (-\frac{4\pi}{6}) = 8\pi = 4 \times 2\pi \) and so the two correspond to the same point on the unit circle. Starting at \( \frac{\pi}{6} \), you can go clockwise four revolutions and wind up back at the same point of the circle which corresponds to \( -\frac{4\pi}{6} \). By convention the negative direction is clockwise and the positive direction is counterclockwise.

### 5.5 Exercises

1. Find \( \cos \theta \) and \( \sin \theta \) for \( \theta \in \{ \frac{2\pi}{3}, \frac{3\pi}{4}, \pi, \frac{5\pi}{6}, \pi, \frac{5\pi}{4}, \frac{4\pi}{3}, \frac{3\pi}{2}, \frac{2\pi}{3}, \frac{7\pi}{6}, \frac{11\pi}{6}, 2\pi \} \).

2. Prove \( \cos^2 \theta = \frac{1 + \cos 2\theta}{2} \) and \( \sin^2 \theta = \frac{1 - \cos 2\theta}{2} \).

3. \( \pi/12 = \pi/3 - \pi/4 \). Therefore, from Problem 2, \( \cos (\pi/12) = \sqrt{1 + \frac{\sqrt{3}/2}{2}} \). On the other hand,

\[
\cos (\pi/12) = \cos (\pi/3 - \pi/4) = \cos \pi/3 \cos \pi/4 + \sin \pi/3 \sin \pi/4
\]
and so \( \cos(\pi/12) = \sqrt{2}/4 + \sqrt{6}/4 \). Is there a problem here? Please explain.

4. Prove \( 1 + \tan^2 \theta = \sec^2 \theta \) and \( 1 + \cot^2 \theta = \csc^2 \theta \).

5. Prove that \( \sin x \cos y = \frac{1}{2} (\sin (x + y) + \sin (x - y)) \).

6. Prove that \( \sin x \sin y = \frac{1}{2} (\cos (x - y) - \cos (x + y)) \).

7. Prove that \( \cos x \cos y = \frac{1}{2} (\cos (x + y) + \cos (x - y)) \).

8. Using Problem 5, find an identity for \( \sin x - \sin y \).

9. Suppose \( \sin x = a \) where \( 0 < a < 1 \). Find all possible values for
   (a) \( \tan x \)
   (b) \( \cot x \)
   (c) \( \sec x \)
   (d) \( \csc x \)
   (e) \( \cos x \)

10. Solve the equations and give all solutions.
    (a) \( \sin (3x) = \frac{1}{2} \)
    (b) \( \cos (5x) = \frac{\sqrt{3}}{2} \)
    (c) \( \tan (x) = \sqrt{3} \)
    (d) \( \sec (x) = 2 \)
    (e) \( \sin (x + 7) = \frac{\sqrt{2}}{2} \)
    (f) \( \cos^2 (x) = \frac{1}{2} \)
    (g) \( \sin^4 (x) = 4 \)

11. Sketch a graph of \( y = \sin x \).

12. Sketch a graph of \( y = \cos x \).

13. Sketch a graph of \( y = \sin 2x \).

14. Sketch a graph of \( y = \tan x \).

15. Find a formula for \( \sin x \cos y \) in terms of sines and cosines of \( x + y \) and \( x - y \).

16. Using Problem 2 graph \( y = \cos^2 x \).

17. If \( f(x) = A \cos \alpha x + B \sin \alpha x \), show there exists \( \phi \) such that
    \[ f(x) = \sqrt{A^2 + B^2} \sin (\alpha x + \phi) . \]

    Show there also exists \( \psi \) such that \( f(x) = \sqrt{A^2 + B^2} \cos (\alpha x + \psi) \). This is a very
    important result, enough that some of these quantities are given names. \( \sqrt{A^2 + B^2} \)
    is called the amplitude and \( \phi \) or \( \psi \) are called phase shifts.

18. Using Problem 17 graph \( y = \sin x + \sqrt{3} \cos x \).

19. Give all solutions to \( \sin x + \sqrt{3} \cos x = \sqrt{3} \). **Hint:** Use Problem 18.
20. As noted above 45° is the same angle as π/4 radians. Explain why 90° is the same angle as π/2 radians. Next find a simple formula which will change the degree measure of an angle to radian measure and radian measure into degree measure.

21. If ABC is a triangle where the capitol letters denote vertices of the triangle and the angle at the vertex. Let a be the length of the side opposite A and b is the length of the side opposite B and c is the length of the side opposite the vertex, C. The law of sines says \( \sin(A)/a = \sin(B)/b = \sin(C)/c \). Prove the law of sines from the definition of the trigonometric functions.

22. In the picture, \( a = 5 \), \( b = 3 \), and \( \theta = \frac{2}{3} \pi \). Find \( c \).

23. In the picture, \( \theta = \frac{1}{4} \pi \), \( \alpha = \frac{2}{3} \pi \) and \( c = 3 \). Find \( a \).

24. An isosceles triangle is one which has two equal sides. For example the following picture is of an isosceles triangle the two equal sides having length \( a \). Show the “base angles” \( \theta \) and \( \alpha \) are equal. **Hint:** You might want to use the law of sines.

25. Find a formula for \( \tan(\theta + \beta) \) in terms of \( \tan(\theta) \) and \( \tan(\beta) \).

26. Find a formula for \( \tan(2\theta) \) in terms of \( \tan(\theta) \).

27. Find a formula for \( \tan\left(\frac{\theta}{2}\right) \) in terms of \( \tan(\theta) \).

28. Show \( \tan(4\theta) = \frac{4\tan(\theta) - 4\tan^3(\theta)}{1 - 6\tan^2(\theta) + \tan^4(\theta)} \). Now find \( x \) such that if \( \tan(\theta) = x \), and \( \tan(\beta) = \frac{1}{4} \), then \( 4\beta + \theta = \frac{3}{4} \). This is the basis for a wonderful formula which has been used to compute \( \pi \) for hundreds of years.

29. The function, \( \sin \) has domain equal to \( \mathbb{R} \) and range \( [-1, 1] \). However, this function is not one to one because \( \sin(x + 2\pi) = \sin x \). Show that if the domain of the function is restricted to be \( [-\frac{\pi}{2}, \frac{\pi}{2}] \), then \( \sin \) still maps onto \( [-1, 1] \) but is now also one to one on this restricted domain. Therefore, there is an inverse function, called arcsin which is defined by \( \arcsin(x) \equiv \) the angle whose \( \sin \) is \( x \) which is in the interval, \( [-\frac{\pi}{2}, \frac{\pi}{2}] \). Thus \( \arcsin\left(\frac{1}{2}\right) \) is the angle whose \( \sin \) is \( \frac{1}{2} \) which is in
5.5. EXERCISES

\([-\pi/6, \pi/6]\). This angle is \(\pi/6\). Suppose you wanted to find \(\tan(\arcsin(x))\). How would you do it? Consider the following picture which corresponds to the case where \(x > 0\).

\[
\begin{align*}
\theta &= \arcsin(x) \\
\sin \theta &= x \\
\cos \theta &= \sqrt{1 - x^2}
\end{align*}
\]

Then letting \(\theta = \arcsin(x)\), the thing which is wanted is \(\tan \theta\). Now from the picture, you see this is \(\frac{x}{\sqrt{1 - x^2}}\). If \(x\) were negative, you would have the little triangle pointing down rather than up as in the picture. The result would be the same for \(\tan \theta\). Find the following:

(a) \(\cot(\arcsin(x))\)
(b) \(\sec(\arcsin(x))\)
(c) \(\csc(\arcsin(x))\)
(d) \(\cos(\arcsin(x))\)

30. Using Problem 29 and the formulas for the trig functions of a sum of angles, find the following.

(a) \(\cot(\arcsin(2x))\)
(b) \(\sec(\arcsin(x + y))\)
(c) \(\csc(\arcsin(x^2))\)
(d) \(\cos(2\arcsin(x))\)
(e) \(\tan(\arcsin(x) + \arcsin(y))\)
(f) \(\csc(\arcsin(x) - \arcsin(y))\)

31. The function, \(\cos\), is onto \([-1, 1]\) but fails to be one to one. Show that if the domain of \(\cos\) is restricted to be \([0, \pi]\), then \(\cos\) is one to one on this restricted domain and still is onto \([-1, 1]\). Define \(\arccos(x)\) \(\equiv\) the angle whose cosine is \(x\) which is in \([0, \pi]\). Find the following.

(a) \(\tan(\arccos(x))\)
(b) \(\cot(\arccos(x))\)
(c) \(\sin(\arccos(x))\)
(d) \(\csc(\arccos(x))\)
(e) \(\sec(\arccos(x))\)

32. Using Problem 31 and the formulas for the trig functions of a sum of angles, find the following.

(a) \(\cot(\arccos(2x))\)
(b) \(\sec(\arccos(x + y))\)
(c) \(\csc(\arccos(x^2))\)
(d) \(\cos(\arcsin(x) + \arccos(y))\)
33. The function, arctan is defined as arctan \( (x) \equiv \text{the angle whose tangent is } x \) which is in \( \left( -\frac{\pi}{2}, \frac{\pi}{2} \right) \). Show this is well defined and is the inverse function for tan if the domain of tan is restricted to be \( \left( -\frac{\pi}{2}, \frac{\pi}{2} \right) \). Find

(a) \( \cos (\arctan (x)) \)
(b) \( \cot (\arctan (x)) \)
(c) \( \sin (\arctan (x)) \)
(d) \( \csc (\arctan (x)) \)
(e) \( \sec (\arctan (x)) \)

34. Using Problem 33 and the formulas for the trig functions of a sum of angles, find the following.

(a) \( \cot (\arctan (2x)) \)
(b) \( \sec (\arctan (x + y)) \)
(c) \( \csc (\arccos (x^2)) \)
(d) \( \cos (2 \arctan (x) + \arcsin (y)) \)
(e) \( \tan (\arctan (x) + 2 \arccos (y)) \)
(f) \( \csc (2 \arctan (x) - \arccos (y)) \)

5.6 Some Basic Area Formulas

5.6.1 Areas Of Triangles And Parallelograms

This section is a review of how to find areas of some simple figures. The discussion will be somewhat informal since it is assumed the reader has seen this sort of thing already. First of all, consider a right triangle as indicated in the following picture.

![Right Triangle](image)

The area of this triangle shown above must equal \( \frac{ab}{2} \) because it is half of a rectangle having sides \( a \) and \( b \). Now consider a general triangle in which a line perpendicular to the line from \( A \) to \( C \) has been drawn through \( B \).

![General Triangle](image)

The area of this triangle would be the sum of the two right triangles formed. Thus this area would be \( \frac{1}{2} (BD)(AD + CD) = \frac{1}{2} (BD)b \). In words, the area of the triangle...
5.6. SOME BASIC AREA FORMULAS

equals one half the base times the height. This also holds if the height and base are chosen with respect to any other side of the triangle.

A parallelogram is a four sided figure which is formed when two identical triangles are joined along a corresponding side with the corresponding angles not adjacent. For example, see the picture in which the two triangles are $ABC$ and $CDA$.

Note the height of triangle $ABC$ taken with respect to side $AB$ is the same as the height of the parallelogram taken with respect to this same side. Therefore, the area of this parallelogram equals twice the area of one of these triangles which equals $2AB\cdot \text{height of parallelogram} \cdot \frac{1}{2} = AB$ (height of parallelogram). Similarly the area equals height times base where the base is any side of the parallelogram and the height is taken with respect to that side, as just described in the case where $AB$ is the side.

5.6.2 The Area Of A Circular Sector

Consider an arc, $A$, of a circle of radius $r$ which subtends an angle, $\theta$. The circular sector determined by $A$ is obtained by joining the ends of the arc, $A$, to the center of the circle. The sector, $S(\theta)$ denotes the points which lie between the arc, $A$ and the two lines just mentioned. The angle between the two lines is called the central angle of the sector. The problem is to define the area of this shape. First a fundamental inequality must be obtained.

Lemma 5.6.1 Let $1 > \varepsilon > 0$ be given. Then whenever the positive number, $\alpha$, is small enough,

$$1 \leq \frac{\alpha}{\sin \alpha} \leq 1 + \varepsilon \quad (5.17)$$

and

$$1 + \varepsilon \geq \frac{\alpha}{\tan \alpha} \geq 1 - \varepsilon \quad (5.18)$$

Proof: This follows from Corollary [5.3.4] on Page [121]. In this corollary, $l(A) = \alpha$ and so

$$1 - \cos \alpha + \sin \alpha \geq \alpha \geq \sin \alpha.$$

Therefore, dividing by $\sin \alpha$,

$$\frac{1 - \cos \alpha}{\sin \alpha} + 1 \geq \frac{\alpha}{\sin \alpha} \geq 1. \quad (5.19)$$

Now using the properties of the trig functions,

$$\frac{1 - \cos \alpha}{\sin \alpha} = \frac{1 - \cos^2 \alpha}{\sin \alpha (1 + \cos \alpha)} = \frac{\sin^2 \alpha}{\sin \alpha (1 + \cos \alpha)} = \frac{\sin \alpha}{1 + \cos \alpha}.$$

From the definition of the sin and cos, whenever $\alpha$ is small enough,

$$\frac{\sin \alpha}{1 + \cos \alpha} < \varepsilon$$
and so \(5.19\) implies that for such \(\alpha, 5.17\) holds. To obtain \(5.18\) let \(\alpha\) be small enough that \(5.17\) holds and multiply by \(\cos \alpha\). Then for such \(\alpha\),

\[
\cos \alpha \leq \frac{\alpha}{\tan \alpha} \leq (1 + \varepsilon) \cos \alpha
\]

Taking \(\alpha\) smaller if necessary and noting that for all \(\alpha\) small enough, \(\cos \alpha\) is very close to 1, yields \(5.18\). This proves the lemma.

This lemma is very important in another context.

**Theorem 5.6.2** Let \(S(\theta)\) denote the sector of a circle of radius \(r\) having central angle, \(\theta\). Then the area of \(S(\theta)\) equals \(\frac{r^2}{2} \theta\).

**Proof:** Let the angle which \(A\) subtends be denoted by \(\theta\) and divide this sector into \(n\) equal sectors each of which has a central angle equal to \(\theta/n\). The following is a picture of one of these.

![Diagram of a sector](image)

In the picture, there is a circular sector, \(S(\theta/n)\) and inside this circular sector is a triangle while outside the circular sector is another triangle. Thus any reasonable definition of area would require

\[
\frac{r^2}{2} \sin (\theta/n) \leq \text{area of } S(\theta/n) \leq \frac{r^2}{2} \tan (\theta/n).
\]

It follows the area of the whole sector having central angle \(\theta\) must satisfy the following inequality.

\[
\frac{nr^2}{2} \sin (\theta/n) \leq \text{area of } S(\theta) \leq \frac{nr^2}{2} \tan (\theta/n).
\]

Therefore, for all \(n\), the area of \(S(\theta)\) is trapped between the two numbers,

\[
\frac{r^2}{2} \theta \sin (\theta/n) \quad \frac{r^2}{2} \theta \tan (\theta/n).
\]

Now let \(\varepsilon > 0\) be given, a small positive number less than 1, and let \(n\) be large enough that

\[
1 \geq \frac{\sin (\theta/n)}{(\theta/n)} \geq \frac{1}{1 + \varepsilon}
\]

and

\[
\frac{1}{1 + \varepsilon} \leq \frac{\tan (\theta/n)}{(\theta/n)} \leq \frac{1}{1 - \varepsilon}.
\]

Therefore,

\[
\frac{r^2}{2} \theta \left(\frac{1}{1 + \varepsilon}\right) \leq \text{Area of } S(\theta) \leq \left(\frac{1}{1 - \varepsilon}\right) \frac{r^2}{2} \theta.
\]

Since \(\varepsilon\) is an arbitrary small positive number, it follows the area of the sector equals \(\frac{r^2}{2} \theta\) as claimed. (Why?)
5.7 Exercises

1. Give another argument which verifies the Pythagorean theorem by supplying the details for the following argument\(^4\). Take the given right triangle and situate copies of it as shown below. The big four sided figure which results is a rectangle because all the angles are equal. Now from the picture, the area of the big square equals \(c^2\), the area of each triangle equals \(ab/2\), since it is half of a rectangle of area \(ab\), and the area of the inside square equals \((b - a)^2\). Here \(a, b,\) and \(c\) are the lengths of the respective sides. Therefore,

\[
c^2 = 4 \left( \frac{ab}{2} \right) + (b - a)^2
= 2ab + b^2 + a^2 - 2ab
= a^2 + b^2.
\]

2. Another very simple and convincing proof of the Pythagorean theorem\(^5\) is based on writing the area of the following trapezoid two ways. Sum the areas of three triangles in the following picture or write the area of the trapezoid as \((a + b) a + \frac{1}{2} (a + b) (b - a)\) which is the sum of a triangle and a rectangle as shown. Do it both ways and see the pythagorean theorem appear.

3. Make up your own proof of the pythagorean theorem based on the following picture.

\(^4\)This argument is old and was known to the Indian mathematician Bhaskar who lived 1114-1185 A.D.
\(^5\)This argument involving the area of a trapezoid is due to James Garfield who was one of the presidents of the United States.
4. A right circular cone has radius $r$ and height $h$. This is like an ice cream cone. Find the area of the side of this cone in terms of $h$ and $r$. **Hint:** Think of painting the side of the cone and while the paint is still wet, rolling it on the floor yielding a circular sector.

5. An equilateral triangle is one in which all sides are of equal length. Find the area of an equilateral triangle whose sides have length $l$.

6. Draw two parallel lines one having length $a$ and the other having length $b$ suppose also these lines are at a distance of $h$ from each other. Now join the ends of these lines to obtain a four sided figure. What is the area of this four sided figure?

7. Explain why the area of a circle of radius $r$ is $\pi r^2$.

8. Explain why through any point in the plane there exists a line parallel to a given line in the plane.

9. Explain why the sum of the radian measures of the angles in any triangle equals $\pi$. **Hint:** Consider the following picture and use the result of Problem 8.

10. The following picture is of an “inscribed angle”, denoted by $\theta$ in a circle of radius $a$. Drawing a line from the center as shown in the picture, it follows from Problem 24 on Page 126 the two base angles are equal. These are denoted as $\theta$ in the picture.

Now the radian measure of $\alpha$ is $l/a$. Using the result of Problem 9 show the radian measure of $\theta$ equals $l/2a$.

11. The inscribed angle in Problem 10 has the special property that one side is a diameter of the circle. A general inscribed angle is just like the one shown in this
problem but without the requirement that either of the sides of the angle are a diameter. Show that for a completely arbitrary inscribed angle a similar result holds to the one in Problem [10].
Exponential Functions And Logarithms

6.1 The Exponential Function

What is \(2^x\) for \(x = m/n\) with \(m, n\) integers? You may remember from High School this is defined as \(\sqrt[n]{2^m}\). From Theorem 1.11.5 the symbol at least makes sense because that theorem says \(m^{th}\) roots exist. However, this is only a theoretical issue. It has no usefulness in terms of computing the \(2^{m/n}\) in general. For example, could you use it to find \(2^{100003/1000000}\)? You would first need to take \(2^{100003}\) and then take the indicated root. However, you have a problem right away. This number is too big to compute very well. However, the computer knows \(2^{100003} = 2.1436\). You can be sure it is not using the definition to do this computation. Something else is going on. It gets even worse of course if \(x\) is an irrational number like \(\sqrt{2}\), one which cannot be written as the quotient of two integers. In this case, you don’t even have a theoretical definition of what \(2^x\) is.

This entire section is devoted to proving the existence of the exponential function, \(\exp\). The main thing to know is its properties and that this function exists. It is a function which satisfies the following.

\[
\exp \text{ is defined on all of } \mathbb{R} \quad (6.1)
\]
\[
\exp (x + y) = \exp (x) \exp (y) \quad (6.2)
\]
\[
\exp : \mathbb{R} \rightarrow (0, \infty) \text{ is one to one and onto} \quad (6.3)
\]
\[
\exp (0) = 1 \quad (6.4)
\]

It has other properties which with the above make it unique. However, we only need the existence of the function to deal with the question of the exponential functions \(x \rightarrow b^x\) for \(x \in \mathbb{R}\). If you are a person who does not require explanations and is happy to believe on faith that such a function as the above exist, then you can simply skip the entire section and go directly to the next.

For those who think this way, consider the Chimera, an animal which has a goat’s head on its back, a lion’s body, a lion’s head in the front and a tail which has a snake’s head on the end. It breathes fire and is very dangerous. You might also consider the function \(\mathcal{N}\) which satisfies all of 6.1-6.4 except \(6.2\) is changed to

\[
\mathcal{N} (x + y) = \frac{\mathcal{N} (y)}{\mathcal{N} (x)} \quad (6.5)
\]

All functions \(\mathcal{N}\) satisfying these properties are very remarkable functions indeed. See Problem 23. The point is, you can prove anything you want about something which doesn’t exist but a theory based on such Chimeras is not very useful.

The exponential function \(\exp\) is defined as follows.
Definition 6.1.1 For \( x \in \mathbb{R} \), \( \exp(x) \equiv \lim_{n \to \infty} (1 + \frac{x}{n})^n \).

Of course I have to show the limit exists and then establish properties of this function. Note that for \( x > 0 \), \( P \left( 1 + \frac{x}{n} \right)^n \) is the future value of a payment after one year if the interest rate is \( x \) per year and there are \( n \) payment periods in the year.

Lemma 6.1.2 Suppose \( \{a_n\} \) is a sequence and each \( a_n > 0 \). Suppose also that

\[
\lim_{k \to \infty} \frac{a_{k+1}}{a_k} = r < 1
\]

Then there exists a constant \( C \) such that for all \( n \),

\[
\sum_{k=1}^{n} a_k < C
\]

and so by Theorem 4.6.16, \( \lim_{n \to \infty} \sum_{k=1}^{n} a_k \) exists.

Proof: Let \( R \) be such that \( r < R < 1 \). From the definition of the limit in Definition 4.6.1 there exists \( N \) such that if \( n \geq N \), then

\[
a_{n+1}/a_n < R.
\]

Therefore, in particular,

\[
a_{N+1} < Ra_N, a_{N+2} < Ra_{N+1} < R^2a_N, \ldots a_{N+p} < R^p a_N.
\]

It follows that for \( n > N \),

\[
\sum_{k=1}^{n} a_k \leq \sum_{k=1}^{N} a_k + \sum_{k=N+1}^{n} a_k
\]

Now using (6.7)

\[
\leq \sum_{k=1}^{N} a_k + \sum_{k=1}^{n-N} a_{N+k} \leq \sum_{k=1}^{N} a_k + \sum_{k=1}^{n} a_N R^k
\]

and by the formula for the sum of a geometric series,

\[
\leq \sum_{k=1}^{N} a_k + \frac{a_N R - a_N R^{n+1}}{1 - R} \leq \sum_{k=1}^{N} a_k + \frac{a_N R}{1 - R} \equiv C
\]

The inequality also holds if \( n \leq N \). Now the sequence whose \( n^{th} \) term is the sum

\[
\sum_{k=1}^{n} a_k
\]

is an increasing sequence because each of the \( a_k \geq 0 \). Since it is bounded above, it follows from Theorem 4.6.16 the limit in (6.6) exists. This proves the lemma.

Now here is the main theorem on properties of \( \exp(x) \).
Theorem 6.1.3 For each $x \in \mathbb{R}$, the sequence 
\[
(1 + \frac{x}{n})^n
\]
converges and so $\exp(x)$ given in Definition 6.1.1 is well defined. If $x \geq 0$, the sequence is increasing and bounded above. Furthermore,
\[
\exp(x) > 0
\]
for all $x \in \mathbb{R}$ and
\[
\exp(-x) = 1/\exp(x)
\]
Also for all $x, y \in \mathbb{R}$,
\[
\exp(x + y) = \exp(x) \exp(y), \exp(0) = 1.
\]
The function $x \to \exp(x)$ is strictly increasing.

Proof: First consider the case where $x \geq 0$. Consider the binomial expansions of 
\[
(1 + \frac{x}{n})^n
\]
and 
\[
(1 + \frac{x}{n+1})^{n+1}
\]. The first of these is
\[
\sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) \left( \frac{x}{n} \right)^k
\]
\[
= \sum_{k=0}^{n} \frac{n(n-1)\cdots(n-k+1)(n-k)!}{k!n\cdot n\cdot \cdots \cdot n(n-k)!} x^k
\]
\[
= \sum_{k=0}^{n} \frac{n(n-1)\cdots(n-k+1)}{k!n\cdot n\cdot \cdots \cdot n} x^k
\]
The binomial expansion of the second of these is
\[
\sum_{k=0}^{n+1} \left( \begin{array}{c} n+1 \\ k \end{array} \right) \left( \frac{x}{n+1} \right)^k
\]
\[
= \sum_{k=0}^{n+1} \frac{(n+1)n(n-1)\cdots(n-k+2)(n+1-k)!}{k!(n+1)\cdot(n+1)\cdots(n+1)(n+1-k)!} x^k
\]
\[
= \sum_{k=0}^{n+1} \frac{(n+1)n(n-1)\cdots(n-k+2)}{k!(n+1)\cdot(n+1)\cdots(n+1)} x^k
\]
Now here is what you immediately notice. There are more positive terms added in the second than in the first, $n+1$ as opposed to $n$. Also if you consider the $k^{th}$ term of each for $k \leq n$, you have
\[
\frac{1}{k!} \left( \begin{array}{c} n \\ k \end{array} \right) \left( \frac{n-1}{n} \right) \left( \frac{n-2}{n} \right) \cdots \left( \frac{n-k+1}{n} \right) x^k
\]
in the first and the slightly larger
\[
\frac{1}{k!} \left( \begin{array}{c} n+1 \\ k \end{array} \right) \left( \frac{n}{n+1} \right) \left( \frac{n-1}{n+1} \right) \cdots \left( \frac{n-k+2}{n+1} \right) x^k
\]
for the second. Thus the sequence is increasing.
The binomial expansion of $(1 + \frac{x}{n})^n$ equals
\[
\sum_{k=0}^{n} \frac{n(n-1)\cdots(n-k+1)}{k!n^k} x^k
\]
and
\[
\lim_{k \to \infty} \frac{n(n-1) \cdots (n-k)}{k!n^k} x^k = \lim_{k \to \infty} \frac{1}{k+1} \frac{1}{n} x = 0 < 1
\]
and so by Lemma 6.1.2, there exists an upper bound for \((1 + \frac{x}{n})^n\). Therefore, by Theorem 4.6.16 \(\lim_{n \to \infty} (1 + \frac{x}{n})^n\) exists and \(\exp(x)\) is well defined for all \(x \geq 0\).

Now I want to verify this holds for all real \(x\).

\[
(1 - \frac{x}{n})^n (1 + \frac{x}{n})^n = \left(1 - \frac{x^2}{n^2}\right)^n \quad (6.8)
\]

because
\[
(1 - \frac{x}{n}) \left(1 + \frac{x}{n}\right) = 1 - \frac{x^2}{n^2}.
\]

Now
\[
\left| \left(1 - \frac{x^2}{n^2}\right)^n - 1 \right| \leq \left| \sum_{k=1}^{n} \binom{n}{k} (-1)^k \left(\frac{x^2}{n^2}\right)^k \right| \leq \sum_{k=1}^{n} \frac{n(n-1) \cdots (n-k+1)}{k!n^k} \left(\frac{x^2}{n}\right)^k \quad (6.9)
\]

For \(n > x^2\),
\[
\left(\frac{x^2}{n}\right)^k \leq \left(\frac{x^2}{n}\right)
\]

and so for such \(n\), (6.9) is less than
\[
\left(\frac{x^2}{n}\right) \sum_{k=1}^{n} \frac{n(n-1) \cdots (n-k+1)}{k!n^k} \leq \left(\frac{x^2}{n}\right) \sum_{k=1}^{n} \frac{1}{k!}
\]

Now
\[
\lim_{k \to \infty} \frac{1}{(k+1)!} = \lim_{k \to \infty} \frac{1}{k+1} = 0 < 1
\]

and so by Lemma 6.1.2 it follows there exists a constant \(C\) independent of \(n\) such that for all \(n\) sufficiently large,
\[
\left| \left(1 - \frac{x^2}{n^2}\right)^n - 1 \right| \leq \left(\frac{x^2}{n}\right) C
\]

and so for \(n\) sufficiently large,
\[
\left| \left(1 - \frac{x^2}{n^2}\right)^n - 1 \right| < \varepsilon
\]

and from the definition of the limit and (6.8) this says
\[
\lim_{n \to \infty} \left(1 - \frac{x}{n}\right)^n \left(1 + \frac{x}{n}\right)^n = \lim_{n \to \infty} \left(1 - \frac{x^2}{n^2}\right)^n = 1
\]

Now it is clear \(\exp(x) \geq 1\) if \(x \geq 0\) because for such \(x\), it was shown above that the sequence \((1 + \frac{x}{n})^n\) is increasing and each term of the sequence is at least as large as 1. (Just replace \(x\) by 0.) It follows from Theorem 4.6.6 that \(\lim_{n \to \infty} (1 - \frac{x}{n})^n\) exists and in fact
\[
\lim_{n \to \infty} \left(1 - \frac{x}{n}\right)^n = \lim_{n \to \infty} \frac{(1 - \frac{x}{n})^n (1 + \frac{x}{n})^n}{(1 + \frac{x}{n})^n} = \frac{1}{\exp(x)}.
\]
This shows $\exp(-x) = 1/\exp(x)$. This also shows that for all $x \in \mathbb{R}$, $\exp(x) \neq 0$ because $\exp(-x) \exp(x) = 1$.

It remains to show $\exp(x + y) = \exp(x) \exp(y)$ and $\exp(x) > 0$ for all $x$. The second assertion follows right away from the first because $\exp(x) = \exp(x^2 + x^2) = \exp(x^2) \exp(x^2) > 0$ because it was shown above that $\exp(x) \neq 0$. Thus it only remains to verify the first of these assertions, $\exp(x + y) = \exp(x) \exp(y)$.

From the definition,

$$\exp(x + y) = \lim_{n \to \infty} \left( 1 + \frac{x+y}{n} \right)^n$$

$$= \lim_{n \to \infty} \left( 1 + \frac{x}{n} \right)^n \left( 1 + \frac{y}{n + x} \right)^n \quad (6.10)$$

That last factor is interesting.

$$\left( 1 + \frac{y}{n + x} \right)^n = \left( 1 + \frac{y}{n} \right)^n \left( 1 + \frac{y}{n + x} - \frac{y}{x} \right)^n$$

$$= \left( 1 + \frac{y}{n} \right)^n \left( 1 + \frac{-yx}{(n + x)(n + y)} \right)^n$$

I claim

$$\lim_{n \to \infty} \left( 1 + \frac{-yx}{(n + x)(n + y)} \right)^n = 1. \quad (6.12)$$

Assuming this, the desired result is obtained from Theorem 4.6.6. Here is why.

$$\exp(x + y) = \lim_{n \to \infty} \left( 1 + \frac{x}{n} \right)^n \left( 1 + \frac{y}{n + x} \right)^n$$

$$= \lim_{n \to \infty} \left( 1 + \frac{x}{n} \right)^n \left( 1 + \frac{y}{n} \right)^n \left( 1 + \frac{-yx}{(n + x)(n + y)} \right)^n$$

$$= \exp(x) \exp(y) \cdot 1 = \exp(x) \exp(y).$$

Thus the theorem is proved if (6.12) is shown.

$$\left| \left( 1 + \frac{-yx}{(n + x)(n + y)} \right)^n - 1 \right| = \sum_{k=1}^{n} \binom{n}{k} \left( \frac{-yx}{(n + x)(n + y)} \right)^k$$

Let $n > 2$ be large enough that $n + x$ and $n + y$ are both larger than $n/2$. Then the above is no larger than

$$\sum_{k=1}^{n} \binom{n}{k} |xy|^k \left( \frac{1}{n + x} \right)^k \left( \frac{1}{n + y} \right)^k$$

$$\leq \sum_{k=1}^{n} \binom{n}{k} |xy|^k \left( \frac{2}{n} \right)^k \left( \frac{2}{n} \right)^k$$
\[
\leq \sum_{k=1}^{n} \frac{n(n-1) \cdots (n-k+1)}{k!n^k} |xy|^k 2^k \left( \frac{2}{n} \right)^k
\]
\[
\leq \left( \frac{2}{n} \right) \sum_{k=1}^{n} \frac{1}{k!} |xy|^k 2^k
\]

Now the conditions of Lemma 6.1.2 are satisfied for the sum and so there is a constant \( C \) independent of \( n \) sufficiently large such that for such \( n \)
\[
\left| \left( 1 + \frac{-yx}{(n+x)(n+y)} \right)^n - 1 \right| \leq C \left( \frac{2}{n} \right)
\]
and this establishes 6.12 It is obvious from the definition that \( \exp(0) = 1 \).

Finally, why is the function strictly increasing? Let \( x < y \). Then
\[
\exp(y) = \exp(y-x) \exp(x) > \exp(x)
\]
because \( \exp(\alpha) > 1 \) if \( \alpha > 0 \). This proves the theorem.

Next I want to show that \( \exp : \mathbb{R} \to (0, \infty) \) and is onto. I just showed it is strictly increasing. It remains to verify it is onto.

**Lemma 6.1.4** Let \( b > 0 \) be given. If \( \exp(a) - b \neq 0 \) then there exists \( \delta > 0 \) such that for all \( x \in (a-\delta, a+\delta) \), \( \exp(x) - b \neq 0 \) and the sign of \( \exp(x) - b \) is constant on this interval.

**Proof:** First consider \( \exp(h) \) for \( |h| < 1 \).
\[
\left| \left( 1 + \frac{h}{n} \right)^n - 1 \right| = \sum_{k=1}^{n} \binom{n}{k} \frac{1}{n^k} |h|^k
\]
\[
< \sum_{k=1}^{n} \binom{n}{k} \frac{1}{n^k} |h|^k \leq |h| \sum_{k=1}^{n} \binom{n}{k} \frac{1}{n^k}
\]
\[
\leq |h| \sum_{k=1}^{n} \frac{n(n-1) \cdots (n-k+1)}{k!n^k} \leq |h| \sum_{k=1}^{n} \frac{1}{k!} < C|h|,
\]
where \( C \) does not depend on \( n \), the last inequality following from an application of Lemma 6.1.2. Therefore, taking a limit of both sides as \( n \to \infty \),
\[
|\exp(h) - 1| \leq C|h| \tag{6.13}
\]
whenever \( |h| < 1 \).

Now suppose \( \exp(a) - b > 0 \) and let \( 0 < \delta < 1 \). Then using 6.13 for \( x \in (a-\delta, a+\delta) \) and the triangle inequality,
\[
|\exp(x) - b| \geq |\exp(a) - b| - |\exp(x) - \exp(a)|
\]
\[
= |\exp(a) - b| - |\exp(x-a) \exp(a) - \exp(a)|
\]
\[
= |\exp(a) - b| - (\exp(a)|\exp(x-a) - 1|
\]
\[
\geq |\exp(a) - b| - C \exp(a) \delta
\]
Suppose now that in addition to \( \delta < 1 \), it is also the case that
\[
\delta < \frac{|\exp(a) - b|}{2C \exp(a)}
\]
Thus for \( x \in (a - \delta, a + \delta) \),
\[ |\exp(x) - b| \geq \frac{|\exp(a) - b|}{2} > 0 \]
This proves the lemma.

**Corollary 6.1.5** \( \exp \) maps \( \mathbb{R} \) onto \((0, \infty)\).

**Proof:** Consider \( \exp(1) \). It is at least as large as
\[ \left(1 + \frac{1}{1}\right)^1 = 2 \]
(6.14)
This is because it was shown in Theorem 6.1.3 the sequence \((1 + \frac{x}{n})^n\) is increasing as \( n \) increases. It follows from that theorem that
\[ \exp(n) = \exp(1 + 1 + \cdots + 1) \geq 2^n \]
and so \( \exp(x) \) achieves arbitrarily large values. From Theorem 6.1.3
\[ \exp(-n) = \frac{1}{2^n} \]
so it also achieves arbitrarily small positive values. It only remains to verify that it achieves all the values in between.

Let \( b > 0 \). Then from what was just shown, there exists \( x \) such that \( \exp(x) < b \) and \( y \) such that \( \exp(y) < b \). Let \( S = \{x \in \mathbb{R} : \exp(x) - b < 0\} \). Then \( S \neq \emptyset \) and it is bounded above. It follows it has a least upper bound \( x \). If \( \exp(x) - b > 0 \), then by Lemma 6.1.4 there exists an interval \((x - \delta, x + \delta)\) such that for all \( y \) in this interval, \( \exp(y) - b > 0 \). But then \( x \) is not the least upper bound of \( S \). If \( \exp(x) - b < 0 \), then by Lemma 6.1.4 again, there is an interval \((x - \delta, x + \delta)\) such that for \( y \) in this interval, \( \exp(y) - b < 0 \) and so in this case \( x \) is not even an upper bound of \( S \). Therefore, \( \exp(x) - b = 0 \). This proves the corollary.

Corollary 6.1.5 and Theorem 6.1.3 imply there exists a function \( \exp \) defined on all of \( \mathbb{R} \) which maps one to one and onto \((0, \infty)\) and has the properties
\[ \exp(0) = 1 \]
\[ \exp(x + y) = \exp(x) \exp(y) \]
More could be said about this function but to do so would involve calculus. Let \( \ln \) denote its inverse.

**Definition 6.1.6** \( \ln : (0, \infty) \rightarrow \mathbb{R} \) is defined as the inverse function of \( \exp \). Thus
\[ \exp(\ln(x)) = x, x \in (0, \infty) \]
\[ \ln(\exp(x)) = x, x \in \mathbb{R} \]

**Theorem 6.1.7** \( \ln(1) = 0 \) and \( \ln(xy) = \ln(x) + \ln(y) \). Also \( \ln \) is increasing and \( \ln(b) > 0 \) if \( b > 1 \) and \( \ln(b) < 0 \) if \( b < 1 \).

**Proof:** The first claim follows right away from the observation that
\[ \ln(1) = \ln(\exp(0)) = 0 \]
Consider the second claim. Do exp to both sides.
\[
\exp(\ln(xy)) = xy
\]
and also
\[
\exp(\ln(x) + \ln(y)) = \exp(\ln(x))\exp(\ln(y)) = xy
\]
Since exp is one to one, it follows
\[
\ln(xy) = \ln(x) + \ln(y)
\]
If \(x < y\), then
\[
x = \exp(\ln(x)) < \exp(\ln(y)) = y
\]
and so, since exp is increasing, it follows
\[
\ln(x) < \ln(y).
\]
If \(b > 1\), then since \(\ln\) is increasing,
\[
\ln(b) > \ln(1) = 0
\]
and if \(b < 1\), then
\[
\ln(b) < \ln(1) = 0.
\]
This proves the theorem.

Now with this theorem, it becomes possible to define what it means to raise a positive number to any real exponent. First of all, here is a lemma.

**Lemma 6.1.8** Let \(r\) be a rational number and let \(b > 0\). Then
\[
\ln(b^r) = r \ln(b).
\]
Thus
\[
b^r = \exp(r \ln(b)).
\]

**Proof:** Let \(r = m/n\) where \(n \geq 1\). Then from Theorem 6.1.7
\[
\ln(b) = \ln\left(\sqrt[n]{b}\right)^n = n \ln\left(\sqrt[n]{b}\right)
\]
and so
\[
\ln\left(\sqrt[n]{b}\right) = \frac{1}{n} \ln(b)
\]
Therefore,
\[
\ln(b^r) = \ln\left(b^{m/n}\right) = \ln\left(\left(\sqrt[n]{b}\right)^m\right)
\]
\[
= m \ln\left(\sqrt[n]{b}\right) = \frac{m}{n} \ln(b) = r \ln(b)
\]
This proves the lemma.

With this lemma, here is the definition of \(b^r\) for \(r \in \mathbb{R}\).

**Definition 6.1.9** For \(b > 0\) and \(r \in \mathbb{R}\), \(b^r\) is that positive number which satisfies
\[
\ln(b^r) = r \ln(b).
\]
In other words, doing \(\exp\) to both sides,
\[
b^r = \exp(r \ln(b))\]
With this definition, all the rules of exponents hold and in addition to this, the definition does not contradict the known definition of a positive number raised to a rational exponent. This is the following theorem.

**Theorem 6.1.10** For every \( b > 0 \), one can define \( b^x \) for any \( x \in \mathbb{R} \). It satisfies the usual rules of exponents

\[
b^{x+y} = b^x b^y, \quad b^0 = 1, \quad (b^x)^y = b^{xy}
\]

and it agrees with the usual definition in the case where \( x \) is rational. That is,

\[
b^{m/n} = \sqrt[n]{b^m}
\]

If \( b > 1 \), the function \( x \to b^x \) is strictly increasing and if \( b < 1 \), the function \( x \to b^x \) is strictly decreasing. In either case, this function maps \( \mathbb{R} \) onto \((0, \infty)\).

**Proof:** First consider the laws of exponents. From properties of \( \exp \) in Theorem 6.1.3

\[
b^{x+y} \equiv \exp((x + y) \ln(b)) \equiv \exp(x \ln(b) + y \ln(b)) = \exp(x \ln(b)) \exp(y \ln(b)) \equiv b^x b^y
\]

Also \( b^0 \equiv \exp(0 \ln(b)) = \exp(0) = 1 \).

\[
(b^x)^y \equiv \exp(y \ln(b^x)) \equiv \exp(y \ln(\exp(x \ln(b)))) \equiv \exp(y x \ln(b)) \equiv b^{xy}
\]

By Lemma 6.1.8

\[
b^{m/n} \equiv \exp\left(\frac{m}{n} \ln(b)\right) = \exp\left(\ln\left(\sqrt[n]{b^m}\right)\right) = \sqrt[n]{b^m}
\]

Showing the traditional definition of \( b^{m/n} \) gives the same as the new definition in the case of rational exponents. Now suppose \( b > 1 \). Then for \( x < y \), using the property of \( \exp \) which says it is an increasing function,

\[
b^x \equiv \exp(x \ln(b)) < \exp(y \ln(b)) \equiv b^y
\]

The reason for the inequality is that by Theorem 6.1.7 \( \ln(b) > 0 \) if \( b > 1 \). If \( b < 1 \), then \( \ln(b) < 0 \) and the inequality above will be turned around. To see \( b^x \) maps onto \((0, \infty)\), let \( y > 0 \) and solve the equation \( y = b^x \) by taking \( \ln \) of both sides. Thus

\[
\ln(y) = \ln\left(\exp(x \ln(b))\right) = x \ln(b).
\]

Then

\[
x = \ln(y) / \ln(b).
\]

This proves the theorem.

The functions \( \exp \) and \( \ln \) have been tabulated and they are also available on your calculator or computer. It is quite easy to compute \( \exp(x) \) if \( x \) is small. For example, consider the problem of finding \( \exp(1) \). Let's try the definition. From the definition, this should be close to

\[
\left(1 + \frac{1}{n}\right)^n
\]

for \( n \) large. Let's use the binomial theorem to approximate this by writing the first several terms. Then writing these gives

\[
1 + n \left(\frac{1}{n}\right) + \frac{n(n-1)}{2!} + \frac{n(n-1)(n-2)}{3!} + \frac{n(n-1)(n-2)(n-3)}{4!}
\]
Note that you would always have these terms for any value of $n$ and that as $n$ gets larger, those expressions which multiply $1/2!$, $1/3!$, etc. converge to 1. Thus a lower bound for $\exp(1)$ is

$$\frac{n}{5!} + \frac{n}{5!} + \frac{n}{5!} + \frac{n}{5!} + \frac{n}{6!}$$

and if desired you could include more terms than this. Thus $\exp(1)$ is bounded below by

$$\sum_{k=0}^{n} \frac{1}{k!}$$

for each $n$. Letting $n = 20$ for example, a lower bound for $\exp(1)$ is

$$\sum_{k=0}^{20} \frac{1}{k!} = 2.7183$$

What if I had only used 12 terms in the sum?

$$\sum_{k=0}^{12} \frac{1}{k!} = 2.7183$$

Thus it seems clear that $\exp(1)$ is approximately equal to 2.7183. The number $\exp(1)$ is called Euler’s number.

By similar reasoning to the above, an upper bound for $\exp(x)$, $x \geq 0$ is

$$\lim_{m \to \infty} \sum_{k=0}^{m} \frac{x^k}{k!}$$

because each term in the series is larger than the corresponding term in the binomial expansion of the sequence $(1 + \frac{x}{n})^n$ while a lower bound is

$$\sum_{k=0}^{m} \frac{x^k}{k!}$$

Thus

$$\exp(x) = \lim_{m \to \infty} \sum_{k=0}^{m} \frac{x^k}{k!}$$

and so, at least for $x \geq 0$ one can approximate $\exp(x)$ by taking enough terms of the above sum. This is also the case if $x < 0$ and in fact yields a more efficient way of computing $\exp(x)$ but I will not pursue this issue any further. The above shows it is routine to compute $\exp(x)$. Here is a graph of $y = \exp(x)$.

---

Euler was a Swiss mathematician who lived from 1707-1783. He was the most prolific mathematician who ever lived. His works fill some 80 volumes. He possessed the ability to do enormous computations in his head and had a photographic memory which enabled him to quote from memory Virgil’s Aeneid. These abilities were of great use because he went blind. Euler also was interested in physics and religion.
6.2 Applications

6.2.1 Interest Compounded Continuously

From the formula for compound interest compounded \( n \) times a year, what is the future value after \( t \) years? There are \( n \) payment periods in each year and so there are \( nt \) payment periods in \( t \) years. Hence the future value if the interest rate is \( r \) is given by

\[
P \left( 1 + \frac{r}{n} \right)^{nt} = P \left( 1 + \frac{rt}{nt} \right)^{nt}
\]

and so, taking the limit of the sequence as \( n \to \infty \) yields from the definition of \( \exp \),

\[
P \exp (rt)
\]

This motivates the following procedure which tells how to compound interest continuously.

Procedure 6.2.1 If the interest rate is \( r \) and the interest is compounded continuously, to find the future value after \( t \) years you compute

\[
P \exp (rt)
\]

Example 6.2.2 The interest rate is 10% and the payment is \$1000. What is the future value after 5 years if interest is compounded daily and if interest is compounded continuously.

To compound daily, you would have 365 payment periods in each year so the future value is

\[
1000 \left( 1 + \frac{1}{360} \right)^{5 \times 360} = \$1648.60
\]

Now compounding it continuously you get

\[
1000 \exp (5 \times .1) = \$1648.70
\]

You see, compounding continuously is better than compounding daily. If you wait 5 years you get an extra 10 cents. Well, every little bit helps.
6.2.2 Exponential Growth And Decay

Mathematically, this is identical to the process of exponential growth. Suppose you have a bacteria culture and you feed it all it needs and there is no restriction on its growth due to crowding for example. Then in this case, the rate of growth is proportional to the amount of bacteria present. This is because the more you have, the more bacteria there are to divide and make new bacteria. Consider equally spaced intervals of time such that $n$ of them equal one unit of time where $n$ is large. The unit might be years, days, etc. Also let $A_k$ denote the amount of whatever is growing at the end of the $k^{th}$ increment of time. In exponential growth

$$A_{k+1} - A_k \approx r A_k \left(1/n\right), \quad A_0 = P$$

where $r$ is a proportionality constant called the growth rate and the above difference equation should give a good description of the amount provided $n$ is large enough. Now this is easy to solve.

$$A_{k+1} \approx \left(1 + \frac{r}{n}\right) A_k, \quad A_0 = P$$

and you look for $A_k = \alpha^k$ and find $\alpha$. This is easily seen to be $\left(1 + \frac{r}{n}\right)$ and so

$$A_k \approx P \left(1 + \frac{r}{n}\right)^k$$

Now let $A(t)$ denote the amount at time $t$. What is it? There are $tn$ time intervals so $k$ goes up to $tn$ and you get

$$A(t) \approx P \left(1 + \frac{r}{n}\right)^{nt} \approx P \left(1 + \frac{rt}{nt}\right)^{nt}$$

the approximation for $A(t)$ getting better as $n \to \infty$. This converges as $n \to \infty$ to

$$A(t) = P \exp (rt)$$

which is the formula for exponential growth.

When the rate of change is negative, the process is called exponential decay. This is the process which governs radioactive substances. It is the same formula which results only this time it is of the form

$$A(t) = P \exp (-rt)$$

where $r > 0$.

**Exercise 6.2.3** Carbon 14 has a half life of 5730 years. This means that if you start with a given amount of it and wait 5730 years, there will be half as much left. Carbon 14 is assumed to be constantly created by cosmic rays hitting the atmosphere so that the proportion of carbon in a living organism is the same now as it was a long time ago. This is of course an assumption and there is evidence it is not true but this does not concern us here. When the living thing dies, it quits replenishing the carbon 14 and so that which it has decays according to the above half life. By measuring the amount in the remains of the dead thing and comparing with what it had when it was alive, one can determine an estimate for how long it has been dead. Suppose then you measure the amount of carbon 14 in some dead wood and find there is $1/3$ the amount there would have been when it was alive. How long ago did the tree from which the wood came die?
Let $A(t)$ be the amount of carbon 14 in the sample and let $A_0$ be the amount when it died. Then
\[ A(t) = A_0 \exp(-rt) \]
By assumption
\[ .33A_0 = A_0 \exp(-rt) \]
and cancelling the $A_0$ one can solve for $t$ as follows.
\[ \ln(.33) = -rt \]
If I knew what $r$ was, I could then solve for $t$. The half life is 5730 and so
\[ .5 = \exp(-r5730) \]
and so
\[ \ln(.5) = -r(5730) \]
from properties of $\ln$ described above, $-\ln(1/2) = \ln((1/2)^{-1}) = \ln(2)$. Therefore,
\[ r = \frac{\ln(2)}{5730} = 1.2097 \times 10^{-4} \]
To get this number, I just used the computer. As mentioned above $\ln$ has been tabulated. Therefore, in the problem of interest,
\[ \ln(.33) = -\left(1.2097 \times 10^{-4}\right)t \]
and so
\[ t = \frac{\ln(.33)}{-1.2097 \times 10^{-4}} = 9164.8 \text{ years}. \]

So how did they find the half life of carbon 14? Did Noah have a sample of recently dead wood in the ark and make some measurements which he recorded in the Book of the Law of Noah which were then compared to measurements made in the twentieth century using chronology determined by Bishop Usher to determine that exactly 5730 years had passed? Actually, this is not the way it was done. The half life was also not established by the decree of scientists.

**Example 6.2.4** Find the half life of a radioactive substance if after 5 years there is .999395 of the original amount present.

This says
\[ .999395 = e^{-r5} \]
and so
\[ r = \frac{\ln(.999395)}{-5} = 1.21036617 \times 10^{-4} \]
Now to find the half life $T$, you need to solve the equation
\[ \frac{1}{2} = e^{-\left(1.21036617 \times 10^{-4}\right)T} \]
Thus
\[ \ln(.5) = -\left(1.21036617 \times 10^{-4}\right)T \]
and so
\[ T = \frac{\ln(.5)}{-\left(1.21036617 \times 10^{-4}\right)} = 5726 \]
Using the known properties of exponential decay, you can compute the half life without waiting for over 5000 years.
6.2.3 Logarithms

By Theorem 6.1.10 the function \( x \to b^x \) is one to one and onto \((0, \infty)\) whenever \( b \neq 1, b > 0 \). Therefore, it has an inverse function. This inverse function is denoted as \( \log_b \).

Thus \( \log_b \) maps \((0, \infty)\) to \( \mathbb{R} \) and is one to one and onto, satisfying

\[
\log_b(b^x) = x, \ x \in \mathbb{R}, \ b^{\log_b (x)} = x, \ x > 0
\]  

(6.15)

**Definition 6.2.5** Let \( b > 0, b \neq 1 \). Then \( \log_b : (0, \infty) \to \mathbb{R} \) is defined by

**Proposition 6.2.6** The following relationship holds between \( \log_b \) and \( \ln \).

\[
\log_b (x) = \frac{\ln (x)}{\ln (b)}
\]  

(6.16)

**Proof:** This comes from the definition. For \( b \neq 1, b > 0 \),

\[
b^{\log_b (x)} \equiv x
\]

and also

\[
b^{\ln(x)/\ln(b)} \equiv \exp \left( \frac{\ln (x)}{\ln (b)} \ln (b) \right) = \exp (\ln (x)) = x
\]

and since the function \( x \to b^x \) is one to one, (6.16) follows. This proves the proposition.

**Corollary 6.2.7** The rules of logarithms hold for \( \log_b \). That is \( \log_b(xy) = \log_b (x) + \log_b (y) \) and \( \log_b (1) = 0 \). If \( b > 1 \), \( \log_b \) is increasing and if \( b < 1 \) then \( \log_b \) is decreasing. Also \( \ln (x) = \log_e (x) \) where \( e = \exp (1) \).

**Proof:** The first part follows from Proposition 6.2.6 and Theorem 6.1.7. The second part follows from

\[
\log_e (x) \equiv \frac{\ln (x)}{\ln (e)} = \frac{\ln (x)}{\ln (\exp (1))} = \frac{\ln (x)}{1} = \ln (x).
\]

This proves the corollary. Here are graphs of some of these log functions.

In the following exercises, feel free to use a calculator which can compute logs.
6.3 Exercises

1. Show that for each positive real number $n$,

\[
\sum_{k=1}^{n} \frac{1}{k!} \leq 2, \quad \sum_{k=0}^{n} \frac{1}{k!} \leq 3
\]

2. Using logarithms and their properties, solve the equation

\[2^{x+3} = 3^x\]

3. Find $x$ such that $\log_x (8) = 3$.

4. Find $x$ such that $\log_x \left( \frac{1}{16} \right) = 4$.

5. If $1 < a < b$ and $x > 1$, how are $\log_a (x)$ and $\log_b (x)$ related? Which is larger? Explain why.

6. Find without using a calculator $\log_3 (27)$, $\log_2 (64)$, $\log_{10} (1000)$, $\log_{1/2} (8)$.

7. Find the domain of the function of $x$ given by

\[\log_3 \left( \frac{x + 1}{(x - 1)(x + 2)} \right)\]

**Hint:** You need $x$ to be such that the expression inside the parenthesis is positive and makes sense. Thus you can’t have for example $x = 1$.

8. Find the domain of the function $f(x) = \sqrt{\ln \left( \frac{x + 1}{x + 2} \right)}$.

9. Show the change of base formula

\[\log_b (x) = \frac{\log_a (x)}{\log_a (b)}\]

10. Find all solutions to $\log_2 (x + 4) = \log_4 (x + 16)$.

11. At 4% compounded monthly, how long does it take an amount of money to double?

12. At 6% compounded quarterly, how long will it take an initial amount to double?

13. The population of bacteria grows exponentially. It is observed that every hour this population doubles. How long will it take to have eight times as many bacteria as at the beginning?

14. A pesticide has a half life of 27 years. How long will it take to have only 1/4 the initial amount?

15. Suppose 5% interest is compounded continuously and you make a payment of $100 at the end of every year, starting with an initial $1000. How much will you have at the end of 10 years?

16. Measurements are taken of an exponentially decaying substance and it is found that after 5 years there is .9 of the amount which was present at the start. What is the half life of this substance?
17. Let \( \{ h_n \}_{n=1}^{\infty} \) denote a sequence which converges to 0, each \( h_n \neq 0 \). Show
\[
\lim_{n \to \infty} \frac{\exp(h_n) - 1}{h_n} = 1.
\]
**Hint:** You should begin by estimating
\[
\frac{1}{h} \left( \left( 1 + \frac{h}{n} \right)^n - 1 \right)
\]
It turns out this property along with the other properties of \( \exp \) are enough to get uniqueness of \( \exp \).

18. In the situation of the previous problem, show
\[
\lim_{n \to \infty} \frac{\exp(x + h_n) - \exp(x)}{h_n} = \exp(x).
\]
When you have done this, you have shown the derivative of \( \exp(x) \) equals \( \exp(x) \). It is this fact that makes this function so important in the study of differential equations.

19. Find the following limit. For \( a > 0 \) and \( \lim_{n \to \infty} h_n = 0 \), each \( h_n \neq 0 \),
\[
\lim_{n \to \infty} \frac{a^{x+h_n} - a^x}{h_n} = ?
\]

20. Let \( p \) be an integer. Also let \( \lim_{n \to \infty} h_n = 0 \), each \( h_n \neq 0 \). Find
\[
\lim_{n \to \infty} \frac{(x + h_n)^p - x^p}{h_n}.
\]
**Hint:** There are two ways to do this problem. One way which is fun is to use the binomial theorem. The other way is to use the definition of what it means given above to raise a number to the \( p^{th} \) power.

21. Do the above problem in the case where \( x > 0 \) and \( p \) is an arbitrary real number.

22. Consider a binomial probability distribution. Thus \( X \) is a random variable which can take integer values between 0 and \( n \) and
\[
P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}
\]
Suppose \( E(X) \) which equals \( np \) equals \( \lambda \) and \( n \) is very large. Show that a good approximation to \( P(X = k) \) is
\[
\frac{1}{k!} \lambda^k e^{-\lambda} = \frac{1}{k!} \lambda^k \exp(-\lambda).
\]
Next verify that
\[
\lim_{n \to \infty} \sum_{k=0}^{n} \frac{1}{k!} \lambda^k \exp(-\lambda) = 1.
\]
A random variable \( X \) which has values equal to 0,1,\( \cdots \) is called a Poisson random variable if \( P(X = k) = \frac{1}{k!} \lambda^k e^{-\lambda} \).

23. Show there is no such thing as a function \( N \) which satisfies 6.1 - 6.4 except for 6.2 being changed to 6.5.
24. Show from the definition of the exponential function that for every \( n \) a positive integer,
\[
\lim_{k \to \infty} \frac{k^n}{\exp(k)} = 0.
\]

25. Show from the definition of the exponential function that if \( \{a_k\} \) is an increasing sequence of positive terms which is also unbounded, then for \( \beta > 0 \)
\[
\lim_{k \to \infty} \frac{a_k}{\exp(\beta a_k)} = 0.
\]
Next show this implies that for all \( \alpha > 0 \)
\[
\lim_{k \to \infty} \frac{a_k^\alpha}{\exp(a_k)} = 0
\]
Thus \( \exp \) grows faster than the function \( f(x) = x^\alpha \) for every \( \alpha > 0 \). \textbf{Hint:} You might consider \( n > \alpha \) where \( n \) is a positive integer. Then use the first part to argue that
\[
\lim_{k \to \infty} \frac{a_k}{\exp \left( \frac{1}{n} a_k \right)} = 0
\]
and then use limit theorems to argue
\[
0 = \lim_{k \to \infty} \left( \frac{a_k}{\exp \left( \frac{1}{n} a_k \right)} \right)^n = \lim_{k \to \infty} \frac{a_k^n}{\exp(a_k)}
\]

26. Show that for all \( \alpha > 0 \)
\[
\lim_{k \to \infty} \frac{\ln(k)}{k^\alpha} = 0.
\]
\textbf{Hint:} Use \( k = \exp(\ln(k)) \). Thus \( \ln \) grows slower than any positive power.

27. Show that if \( p > 1, \sum_{k=1}^\infty \frac{1}{k^p} \) converges. \textbf{Hint:}
\[
1 + \frac{1}{2^p} + \frac{1}{3^p} + \frac{1}{4^p} + \frac{1}{5^p} + \frac{1}{6^p} + \frac{1}{7^p} + \frac{1}{8^p} + \frac{1}{9^p} + \ldots
\leq 2(2^p) \leq 4(4^p)
\leq 1 + \frac{1}{2^p} + \frac{1}{3^p} + \frac{1}{4^p} + \frac{1}{5^p} + \frac{1}{6^p} + \frac{1}{7^p} + \frac{1}{8^p} + \frac{1}{9^p} + \ldots
\]
Show every partial sum is bounded above by
\[
\sum_{k=1}^\infty 2^k (2^{-kp}) = \sum_{k=1}^\infty (2^{1-p})^k
\]
These are called the \( p \) series in calculus. Some people also refer to a series like this as \( \zeta(p) \). There are profound questions related to this function.

28. Show that if \( a_n, b_n \) are positive for all \( n \) and
\[
\lim_{n \to \infty} \frac{a_n}{b_n} = c,
\]
where \( c \) is a positive number, then the two infinite series
\[
\sum_{n=1}^\infty a_n, \sum_{n=1}^\infty b_n
\]
converge or diverge together. **Hint:** From the definition of the limit, there exists $N$ large enough that if $n \geq N$,
\[
\frac{c}{2} < \frac{a_n}{b_n} < 2c
\]

Therefore,
\[
\frac{c}{2} b_n < a_n < 2c b_n
\]

Now make an auspicious use of Theorem 4.6.16 applied to the sequences of partial sums for $\sum a_n$ and $\sum b_n$. This is called the limit comparison test and is very useful because by Problem 27, you have examples of many series which converge and some which diverge.

29. Determine whether the following series converge and explain why.

(a) $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n+1}}$

(b) $\sum_{n=1}^{\infty} \frac{1}{\sqrt{5n+1}}$

(c) $\sum_{n=1}^{\infty} \frac{1}{(\sum_{k=1}^{n} k)^2}$

(d) $\sum_{n=1}^{\infty} \frac{2^{-n}}{\sqrt{n}}$

(e) $\sum_{n=1}^{\infty} \frac{1}{(1+\frac{1}{n})^{n^2}}$

30. Use partial fractions and the definition of an infinite series to show
\[
\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 1
\]

Now if you wanted to compute $\zeta(2) \equiv \sum_{k=1}^{\infty} \frac{1}{k^2}$, you could write
\[
\sum_{n=1}^{\infty} \frac{1}{n^2} = \left( \sum_{n=1}^{\infty} \frac{1}{n^2} - \sum_{n=1}^{\infty} \frac{1}{n(n+1)} \right) + \sum_{n=1}^{\infty} \frac{1}{n(n+1)}
\]

Explain using the definition why
\[
\sum_{n=1}^{\infty} \frac{1}{n^2} - \sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \sum_{n=1}^{\infty} \left( \frac{1}{n^2} - \frac{1}{n(n+1)} \right) = \sum_{n=1}^{\infty} \frac{1}{n^2(n+1)}
\]

Thus you could compute this sum and add to 1 to get $\zeta(2)$. Is there an advantage to doing this?
Parabolas, Ellipses, and Hyperbolas

These are also called the conic sections because they result when a cone is sliced by a plane. Their importance is determined by the fact that planets, comets, asteroids, etc. move along a conic section. There are also very interesting properties connected to the way they focus light and sound. These properties are all obtained using the methods of calculus which is not contained in this book. Therefore, all that is attempted here is a brief description of what these curves are and how to identify some of the interesting parameters which have geometrical and physical significance.

7.0.1 The Parabola

A parabola is a collection of points, $P$ in the plane such that the distance from $P$ to a fixed line, called the directrix, is the same as the distance from $P$ to a given point, $P_0$, called the focus, as shown in the following picture.

\[ y = c \]

\[ P_0 = (a, b) \]

From this geometric description, it is possible to obtain a formula for a parabola. For simplicity, assume the top (vertex) of the parabola is $(0, 0)$ and the line is $y = c$. It follows the focus must be $(0, -c)$. Then if $(x, y)$ is a point on the parabola, the geometric description requires

\[ \left( x^2 + (y + c)^2 \right)^{1/2} = |y - c| . \]

Squaring both sides yields

\[ x^2 + y^2 + 2yc + c^2 = y^2 - 2yc + c^2 \]

and so

\[ x^2 = -4yc \] (7.1)

If $c > 0$ then the largest $y$ can be is 0. This follows from solving the above equation for $y$. Thus $y = -x^2/4c$ which is in this case is always less than or equal to 0. Therefore, the largest $y$ can be and still satisfy the equation for the parabola is 0 when $x = 0$. Thus the vertex is $(0, 0)$. Similarly, if $c < 0$, the smallest $y$ can be is 0 and the vertex is still $(0, 0)$. 

153
This yields the following equation for a parabola.

**Equation Of A Parabola:** The equation for a parabola which has directrix $y = c$, focus $(0, -c)$ and vertex $(0, 0)$ is $x^2 = -4cy$.

This simple description can be used to consider more general situations in which, for example the vertex is not at $(0, 0)$.

**Example 7.0.1** Find the focus and directrix of the parabola $2x^2 - 3x + 1 = 5y$.

First complete the square on the left. Thus $2 \left( x^2 - \frac{3}{4}x + \frac{9}{16} \right) - \frac{9}{8} + 1 = 5y$.

Dividing both sides by 2,

$$(x - \frac{3}{4})^2 = \frac{5}{2} \left( y + \frac{1}{40} \right).$$

Now if this were just $x^2 = -4cy$, you would know the directrix is $y = c$ and the focus is $(0, -c)$. It doesn’t look like this, however. Therefore, change the variables, letting $u = x - \frac{3}{4}$ and $v = y + \frac{1}{40}$. Then the equation in terms of $u$ and $v$ is of the form $u^2 = -4 \left( -\frac{5}{8} \right) v$.

Changing variables letting $u = x - \frac{3}{4}$ and $v = y + \frac{1}{40}$, this gives

$$-4 \left( -\frac{1}{12} \right) u = v^2.$$ (7.3)

Changing variables letting $u = x - \frac{3}{4}$ and $v = y + \frac{1}{40}$, this gives

$$-4 \left( -\frac{1}{12} \right) u = v^2.$$

Therefore, in the $uv$ plane, the directrix is $u = \frac{-1}{12}$ and the focus is $\left( \frac{1}{12}, 0 \right)$ while the vertex is $(0, 0)$. In terms of the original variables the directrix is $x = \frac{3}{5} - \frac{1}{12} = \frac{7}{12}$; the focus is $\left( \frac{1}{12} + \frac{2}{7}, 0 - \frac{1}{4} \right) = \left( \frac{7}{4}, -\frac{1}{4} \right)$ while the vertex is $\left( \frac{3}{5}, -\frac{1}{4} \right)$.

This illustrates the following procedure.
Procedure 7.0.3 To find the focus, directrix and vertex of a parabola of the form

\[ x = ay^2 + by + c, \]

you complete the square and otherwise massage things to get it in the form

\[ -4c(x - p) = (y - q)^2 \]

and then the vertex is at \((p, q)\) and the focus is at \((p - c, q)\) while the directrix is \(x = p + c\).

To find the focus directrix and vertex of a parabola of the form \(y = ax^2 + bx + c\), you complete the square and otherwise massage things to obtain an equation of the form

\[ (x - p)^2 = -4c(y - q). \]

then the vertex is \((p, q)\) the focus is \((p, q - c)\) and the directrix is \(y = q + c\).

7.0.2 The Ellipse

With an ellipse, there are two points, \(P_1\) and \(P_2\) which are fixed and the ellipse consists of the set of points, \(P\) such that \(d(P, P_1) + d(P, P_2) = d\), where \(d\) is a fixed positive number. These two points are called the foci of the ellipse. Each is called a focus point by itself. Here is a picture in case the two points are of the form \((\alpha - h, \beta)\) and \((\alpha + h, \beta)\).

Note in the picture \(2a\) is the length of this ellipse and \(2b\) is its height. The major axis is thus \(2a\) and the minor axis is \(2b\). You can think of a string stretched tight and the pencil being at the point on the graph of the ellipse. Thus you can see from the picture that \(d = 2a\). Just imagine the point is at \((a + \alpha, \beta)\), the right vertex of the ellipse.

For simplicity assume the foci are at \((-c, 0)\) and \((c, 0)\) where \(c \geq 0\). Then from the description and letting \(2a\) be the length of the ellipse, the equation of the ellipse is given by

\[ \sqrt{(x + c)^2 + y^2} + \sqrt{(x - c)^2 + y^2} = 2a \]

Therefore,

\[ \sqrt{(x + c)^2 + y^2} = 2a - \sqrt{(x - c)^2 + y^2} \]

and so squaring both sides,

\[ x^2 + 2xc + c^2 + y^2 = 4a^2 + x^2 - 2xc + c^2 + y^2 - 4a\sqrt{(x - c)^2 + y^2} \]

Subtracting like terms from both sides and then dividing by 4

\[ xc = a^2 - a\sqrt{(x - c)^2 + y^2} \]

It follows

\[ xc - a^2 = -a\sqrt{(x - c)^2 + y^2}. \]
Now square both sides of this to obtain
\[ x^2c^2 - 2xca^2 + a^4 = a^2 (x^2 - 2cx + c^2 + y^2). \]
Next simplify this some more to obtain
\[ a^2 (a^2 - c^2) = (a^2 - c^2) x^2 + a^2 y^2. \]
Therefore,
\[ 1 = \frac{x^2}{a^2} + \frac{y^2}{b^2} \]
where \( b^2 = a^2 - c^2 \). This verifies the following equation of an ellipse.

**Equation Of An Ellipse:** Suppose the ellipse is centered at \((0, 0)\) has major axis parallel to the \(x\) axis and minor axis parallel to the \(y\) axis and that the major axis is \(2a\) while the minor axis is \(2b\). Then the equation of the ellipse is
\[ 1 = \frac{x^2}{a^2} + \frac{y^2}{b^2} \]
and the focus points are at
\[ \left(-\sqrt{a^2 - b^2}, 0\right) \]
and
\[ \left(\sqrt{a^2 - b^2}, 0\right). \]
Switching the variables, if the ellipse is centered at \((0, 0)\) and the major axis of length \(2a\) is parallel to the \(y\) axis while the minor axis of length \(2b\) is parallel to the \(x\) axis. Then the equation of the ellipse is
\[ 1 = \frac{x^2}{b^2} + \frac{y^2}{a^2} \]
and the focus points are at
\[ \left(0, -\sqrt{a^2 - b^2}\right) \]
and
\[ \left(0, \sqrt{a^2 - b^2}\right). \]
As in the case of the parabola, you can use this to find the foci of ellipses which are not centered at \((0, 0)\).

**Example 7.0.4** Find the focus points for the ellipse \(\frac{(x-1)^2}{9} + \frac{(y-2)^2}{4} = 1.\)
Let \(u = x - 1\) and \(v = y - 2\). Then the focus points of the identical ellipse in the \(uv\) plane would be \((\sqrt{5}, 0)\) and \((-\sqrt{5}, 0)\).

**Example 7.0.5** Show the following is the equation of an ellipse and find its major and minor axes along with the focus points.
\[ 2x^2 + 4x + y^2 + 2y + 2 = 0. \]
As usual you have to complete the square in this. Thus
\[ 2 \left(x^2 + 2x + 1\right) + y^2 + 2y + 1 = 1. \]
Therefore,
\[
\frac{(x+1)^2}{(1/\sqrt{2})^2} + (y+1)^2 = 1
\]

The major axis has length 2 and is parallel to the \(y\) axis and the minor axis has length \(\sqrt{2}\) and is parallel to the \(x\) axis. Let \(x+1 = u\) and \(y+1 = v\). Then in terms of \(u, v\) this equation is of the form
\[
\frac{u^2}{(1/\sqrt{2})^2} + v^2 = 1.
\]

The focus points in the \(uv\) plane are
\[
\left(0, \pm \sqrt{1 - (1/2)}\right) = \left(0, \pm \frac{1}{\sqrt{2}}\right).
\]

Therefore, the focus points of the ellipse in the \(xy\) plane are
\[
\left(-1, -1 \pm \frac{1}{\sqrt{2}}\right).
\]

This illustrates the following procedure.

**Procedure 7.0.6** For \(a, b > 0\) and an ellipse in the form
\[
Ax^2 + By^2 + Cx + Dy + E = 0
\]
you first complete the square and then massage to obtain something of the form
\[
\frac{(x - p)^2}{a^2} + \frac{(y - q)^2}{b^2} = 1
\]
where \(a, b\) are positive numbers. If \(a > b\) then the major axis has length \(2a\) and is parallel to the \(x\) axis. In this case the focus points are at
\[
(p - \sqrt{a^2 - b^2}, q)
\]
and
\[
(p + \sqrt{a^2 - b^2}, q)
\]
and the minor axis has length \(2b\) and is parallel to the \(y\) axis. If \(a < b\) then the major axis is parallel to the \(y\) axis and has length \(2b\). In this case the focus points are
\[
(p, q + \sqrt{b^2 - a^2})
\]
and
\[
(p, q - \sqrt{b^2 - a^2})
\]
and the minor axis is parallel to the \(x\) axis and has length \(2a\). If \(a = b\), the ellipse is a circle and it has center at \((p, q)\) with radius equal to \(a = b\).
7.0.3 The Hyperbola

With a hyperbola, there are two points, \( P_1 \) and \( P_2 \) which are fixed and the hyperbola consists of the set of points, \( P \) such that \( d(P, P_1) - d(P, P_2) = d \), where \( d \) is a fixed positive number. These two points are called the foci of the hyperbola. Each is called a focus point by itself. The following picture is descriptive of the above situation.

Now one can obtain an equation which will describe a hyperbola. For simplicity, consider the case where the focus points are \((-c, 0)\) and \((c, 0)\) and let \( d \) denote the difference between the two distances. Then from the above description of the hyperbola,

\[
\sqrt{(x + c)^2 + y^2} - \sqrt{(x - c)^2 + y^2} = d.
\]

Then

\[
x^2 + 2cx + y^2 = d^2 + 2d\sqrt{(x - c)^2 + y^2} + x^2 - 2cx + y^2.
\]

Subtracting like terms,

\[
4cx - d^2 = 2d\sqrt{(x - c)^2 + y^2}.
\]

Therefore, squaring both sides again,

\[
16c^2 x^2 + d^4 - 8xcd^2 = 4d^2 (x^2 - 2cx + c^2 + y^2)
\]

Another simplification gives

\[
d^2 (d^2 - 4c^2) = 4 (d^2 - 4c^2) x^2 + 4d^2 y^2.
\]

Therefore,

\[
1 = \frac{x^2}{(d^2/4)} - \frac{y^2}{(4c^2 - d^2)/4}
\]

Therefore, the vertices of this hyperbola, one of which is labeled \((\alpha, \beta)\) in the above picture are in this case at \((\pm c, 0)\) and \((-\frac{c}{2}, 0)\) so \( d \) is also the distance between the two vertices which is less than the distance between the two focus points, \(2c\). Therefore, \(4c^2 - d^2 > 0\). Let \( a^2 = d^2/4 \) and let \( b^2 = (4c^2 - d^2)/4 \) so the equation of the hyperbola is of the form

\[
1 = \frac{x^2}{a^2} - \frac{y^2}{b^2}.
\]

Observe also the relation between \(a, b\) and \(c\). \(a^2 + b^2 = c^2\).

Another interesting thing about hyperbolas is the asymptotes. Solve the equation above for \(x\).

\[
x = \pm \left(a^2 + \frac{a^2 y^2}{b^2}\right)^{1/2}
\]

Now for large \(|y|\), this is essentially equal to \(\pm \frac{ay}{b}\). The asymptotes of the hyperbola are \(x = \frac{a}{b} y\) and \(x = -\frac{a}{b} y\). These are the dotted lines in the above picture.
**Equation For A Hyperbola:** Suppose you have the relation,

\[ \frac{x^2}{a^2} - \frac{y^2}{b^2} = 1. \]

Then this is a hyperbola having vertices at \((a,0)\) and \((-a,0)\) and foci at

\( (−\sqrt{a^2 + b^2}, 0) \) and \((\sqrt{a^2 + b^2}, 0)\). Its asymptotes are given by

\[ x = \pm \frac{a}{b} y. \]

Switching the variables, suppose now you have the relation

\[ \frac{y^2}{b^2} - \frac{x^2}{a^2} = 1. \]

Then it is a hyperbola having vertices at \((0,b)\) and \((0,-b)\) and foci at

\( (0,−\sqrt{a^2 + b^2}) \) and \((0,\sqrt{a^2 + b^2})\). Its asymptotes are

\[ y = \pm \frac{bx}{a}. \]

The two kinds of hyperbolas are illustrated in the following picture.

\[ \frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \quad \text{and} \quad \frac{y^2}{b^2} - \frac{x^2}{a^2} = 1 \]

In the first kind you can’t have \(x = 0\) and in the second, you can’t have \(y = 0\). Now the hyperbolas can be slid around as indicated in the following example. This has the effect of replacing the \(x\) and \(y\) axes by lines of the form \(x = k\) and \(y = c\).

**Example 7.0.7** Find the focus points and vertices for the hyperbola

\[ \frac{(x - 1)^2}{4} - \frac{(y - 2)^2}{9} = 1. \]

Change the variables. Let \(u = x - 1\) and \(v = y - 2\). Then the focus points of the identical hyperbola in the \(uv\) plane would be \((\sqrt{13},0)\) and \((-\sqrt{13},0)\). Now \(x = u + 1, y = v + 2\) so the focus points of the original hyperbola in the \(xy\) plane are \((\sqrt{13} + 1, 2)\) and \((-\sqrt{13} + 1, 2)\). The vertices are obtained by taking \(y = 2, (3,2)\) and \((-1,2)\).

**Example 7.0.8** Show that the relation, \(2y^2 + 4y - 4x^2 + 8x = 3\) is a hyperbola and find its foci.
As usual, you complete the square. Thus

\[2 \left(y^2 + 2y + 1\right) - 4 \left(x^2 - 2x + 1\right) = 3 + 2 - 4 = 1.\]

Therefore,

\[
\frac{(y + 1)^2}{(1/\sqrt{2})^2} - \frac{(x - 1)^2}{(1/2)^2} = 1.
\]

Therefore, the vertices are \(\left(1, -1 + \frac{1}{\sqrt{2}}\right)\) and \(\left(1, -1 - \frac{1}{\sqrt{2}}\right)\). The foci are at \(\left(1, -1 - \frac{\sqrt{3}}{2}\right)\)

and \(\left(1, -1 + \frac{\sqrt{3}}{2}\right)\).

This illustrates the following procedure for a hyperbola.

Procedure 7.0.9 Suppose \(Ax^2 + By^2 + Cx + Dy + E = 0\) and \(A\) and \(B\) have different signs. Complete the square and massage to obtain an expression of one of the following forms.

\[
\frac{(x-p)^2}{a^2} - \frac{(y-q)^2}{b^2} = 1 \text{ a.}
\]

\[
\frac{(y-q)^2}{b^2} - \frac{(x-p)^2}{a^2} = 1 \text{ b.}
\]

where \(a, b\) are positive numbers. In the case of a.) the focus points are at \(\left(p + \sqrt{a^2 + b^2}, q\right)\)

and \(\left(p - \sqrt{a^2 + b^2}, q\right)\),

the vertices are at \((p + a, q)\), \((p - a, q)\), and the asymptotes are of the form

\[
\frac{(x-p)}{a} \pm \frac{(y-q)}{b} = 1
\]

In the case of b.) the focus points are at \(\left(p, q + \sqrt{a^2 + b^2}\right)\)

and \(\left(p, q - \sqrt{a^2 + b^2}\right)\),

the vertices are at \((p, q + b)\), \((p, q - b)\), and the asymptotes are of the form

\[
\frac{(y-q)}{b} \pm \frac{(x-p)}{a} = 1.
\]
7.1 Exercises

1. Consider \( y = 2x^2 + 3x + 7 \). Find the focus and the directrix of this parabola.

2. Sketch a graph of the ellipse whose equation is \( \frac{(x-1)^2}{4} + \frac{(y-2)^2}{9} = 1 \).

3. Sketch a graph of the ellipse whose equation is \( \frac{(x-1)^2}{9} + \frac{(y-2)^2}{4} = 1 \).

4. Sketch a graph of the hyperbola, \( \frac{x^2}{4} - \frac{y^2}{9} = 1 \).

5. Sketch a graph of the hyperbola, \( \frac{y^2}{4} - \frac{x^2}{9} = 1 \).

6. Find the focus points for the hyperbola \( \frac{(x-1)^2}{4} - \frac{(y-2)^2}{9} = 1 \).

7. Find a formula for the focus points for a hyperbola of the form \( \frac{(x-p)^2}{a^2} - \frac{(y-q)^2}{b^2} = 1 \) where \( a \geq b \).

8. Find a formula for the focus points for a hyperbola of the form \( \frac{(y-p)^2}{a^2} - \frac{(x-q)^2}{b^2} = 1 \) where \( b \geq a \).

9. Find a formula for the focus points for an ellipse of the form \( \frac{(x-p)^2}{a^2} + \frac{(y-q)^2}{b^2} = 1 \) where \( a \geq b \).

10. Find a formula for the focus points for an ellipse of the form \( \frac{(y-p)^2}{a^2} + \frac{(x-q)^2}{b^2} = 1 \) where \( b \geq a \).

11. Consider the hyperbola, \( \frac{y^2}{4} - \frac{x^2}{9} = 1 \). Show that \( y = \pm \sqrt{\frac{b^2x^2}{a^2} + \frac{b^2x^2}{a^2}} \). The straight lines \( y = \frac{bx}{a} \) and \( y = -\frac{bx}{a} \) are called the asymptotes of the hyperbola. Show that for large \( x \), \( \sqrt{\frac{b^2x^2}{a^2} + \frac{b^2x^2}{a^2}} - \frac{bx}{a} \) is very small.

12. What is the diameter of the ellipse, \( \frac{(x-1)^2}{9} + \frac{(y-2)^2}{4} = 1 \)? Remember the diameter of a set, \( S \) is defined as \( \sup \{|x - y| : x, y \in S\} \).

13. Consider \( 9x^2 - 36x + 32 - 4y^2 - 8y = 36 \). Identify this as either an ellipse, a hyperbola or a parabola. Then find its focus point(s) and its directrix if it is a parabola. \textbf{Hint:} First complete the square.

14. Consider \( 4x^2 - 8x + 68 + 16y^2 - 64y = 64 \). Identify this as either an ellipse, a hyperbola or a parabola. Then find its focus point(s) and its directrix if it is a parabola. \textbf{Hint:} First complete the square.

15. Consider \(-8x + 8 + 16y^2 - 64y = 64 \). Identify this as either an ellipse, a hyperbola or a parabola. Then find its focus point(s) and its directrix if it is a parabola. \textbf{Hint:} First complete the square.

16. Consider \( 5x + 3y^2 + 2y = 7 \). Identify this as either an ellipse, a hyperbola or a parabola. Then find its focus point(s) and its directrix if it is a parabola. \textbf{Hint:} First complete the square.

17. Consider the two points, \( P_1 = (0, 0) \) and \( P_2 = (1, 1) \). Find the equation of the ellipse defined by \( d(P, P_1) + d(P, P_2) = 4 \) which has these as focus points.

18. Find the equation of the parabola which has focus \((0, 0)\) and directrix \( x + y = 1 \). This is pretty hard. To do it you need to figure out how to find the distance between a point and the given line.
19. As explained earlier, \((\cos t, \sin t)\) for \(t \in \mathbb{R}\) is a point on the circle of radius 1. Find a formula for the coordinates of a point on the ellipse, \(\frac{(x-2)^2}{4} + \frac{(y+1)^2}{8} = 1\). **Hint:** This says \(\left(\frac{x-2}{2}, \frac{y+1}{\sqrt{8}}\right)\) is a point on the unit circle.
The Complex Numbers

Just as a real number should be considered as a point on the line, a complex number is considered a point in the plane which can be identified in the usual way using the Cartesian coordinates of the point. Thus \((a, b)\) identifies a point whose \(x\) coordinate is \(a\) and whose \(y\) coordinate is \(b\). In dealing with complex numbers, such a point is written as \(a + ib\). For example, in the following picture, I have graphed the point \(3 + 2i\). You see it corresponds to the point in the plane whose coordinates are \((3, 2)\).

\[3 + 2i\]

Multiplication and addition are defined in the most obvious way subject to the convention that \(i^2 = -1\). Thus,

\[(a + ib) + (c + id) = (a + c) + i(b + d)\]

and

\[(a + ib)(c + id) = ac + iad + ibc + i^2bd = (ac - bd) + i(bc + ad)\].

Every non zero complex number, \(a + ib\), with \(a^2 + b^2 \neq 0\), has a unique multiplicative inverse.

\[
\frac{1}{a + ib} = \frac{a - ib}{a^2 + b^2} = \frac{a}{a^2 + b^2} - \frac{b}{a^2 + b^2}i.
\]

You should prove the following theorem.

**Theorem 8.0.1** The complex numbers with multiplication and addition defined as above form a field satisfying all the field axioms listed on Page 9.

The field of complex numbers is denoted as \(\mathbb{C}\). An important construction regarding complex numbers is the complex conjugate denoted by a horizontal line above the number. It is defined as follows.

\[\overline{a + ib} \equiv a - ib\].

What it does is reflect a given complex number across the \(x\) axis. Algebraically, the following formula is easy to obtain.

\[(a + ib)(a + ib) = a^2 + b^2\].
Definition 8.0.2 Define the absolute value of a complex number as follows.

\[ |a + ib| \equiv \sqrt{a^2 + b^2}. \]

Thus, denoting by \( z \) the complex number, \( z = a + ib \),

\[ |z| = (z\overline{z})^{1/2}. \]

The triangle inequality holds for the absolute value for complex numbers just as it does for the ordinary absolute value.

**Proposition 8.0.3** Let \( z, w \) be complex numbers. Then the triangle inequality holds.

\[ |z + w| \leq |z| + |w|, \quad ||z| - |w|| \leq |z - w|. \]

**Proof:** From the definition,

\[ |z + w|^2 = (z + w)(\overline{z} + \overline{w}) \]

\[ = z\overline{z} + w\overline{w} + z\overline{w} + \overline{z}w = |z|^2 + 2\text{Re}(z\overline{w}) + |w|^2 \]

\[ \leq |z|^2 + 2|z||w| + |w|^2 = (|z| + |w|)^2 \]

and taking the square root, this gives the first inequality above. To get the second,

\[ z = z - w + w, \quad w = w - z + z \]

and so by the first form of the inequality

\[ |z| \leq |z - w| + |w|, \quad |w| \leq |z - w| + |z| \]

and so both \( |z| - |w| \) and \( |w| - |z| \) are no larger than \( |z - w| \) and this proves the second version because \( ||z| - |w|| \) is one of \( |z| - |w| \) or \( |w| - |z| \). This proves the proposition.

With this definition, it is important to note the following. Be sure to verify this. It is not too hard but you need to do it.

**Remark 8.0.4**: Let \( z = a + ib \) and \( w = c + id \). Then \( |z - w| = \sqrt{(a - c)^2 + (b - d)^2}. \)

Thus the distance between the point in the plane determined by the ordered pair, \( (a, b) \) and the ordered pair \( (c, d) \) equals \( |z - w| \) where \( z \) and \( w \) are as just described.

For example, consider the distance between \( (2, 5) \) and \( (1, 8) \). From the distance formula this distance equals \( \sqrt{(2 - 1)^2 + (5 - 8)^2} = \sqrt{10} \). On the other hand, letting \( z = 2 + i5 \) and \( w = 1 + i8 \), \( z - w = 1 - i3 \) and so \( (z - w)(\overline{z - w}) = (1 - i3)(1 + i3) = 10 \) so \( |z - w| = \sqrt{10} \), the same thing obtained with the distance formula.

Complex numbers, are often written in the so called polar form which is described next. Suppose \( x + iy \) is a complex number. Then

\[ x + iy = \sqrt{x^2 + y^2}\left(\frac{x}{\sqrt{x^2 + y^2}} + i\frac{y}{\sqrt{x^2 + y^2}}\right). \]

Now note that

\[ \left(\frac{x}{\sqrt{x^2 + y^2}}\right)^2 + \left(\frac{y}{\sqrt{x^2 + y^2}}\right)^2 = 1 \]
and so
\[
\left( \frac{x}{\sqrt{x^2 + y^2}}, \frac{y}{\sqrt{x^2 + y^2}} \right)
\]
is a point on the unit circle. Therefore, there exists a unique angle, \( \theta \in [0, 2\pi) \) such that
\[
\cos \theta = \frac{x}{\sqrt{x^2 + y^2}}, \quad \sin \theta = \frac{y}{\sqrt{x^2 + y^2}}.
\]
The polar form of the complex number is then
\[
r (\cos \theta + i \sin \theta)
\]
where \( \theta \) is this angle just described and \( r = \sqrt{x^2 + y^2} \).

A fundamental identity is the formula of De Moivre which follows.

**Theorem 8.0.5** Let \( r > 0 \) be given. Then if \( n \) is a positive integer,
\[
[r (\cos t + i \sin t)]^n = r^n (\cos nt + i \sin nt).
\]

**Proof:** It is clear the formula holds if \( n = 1 \). Suppose it is true for \( n \).
\[
[r (\cos t + i \sin t)]^{n+1} = [r (\cos t + i \sin t)]^n [r (\cos t + i \sin t)]
\]
which by induction equals
\[
= r^{n+1} (\cos nt + i \sin nt) (\cos t + i \sin t)
\]
\[
= r^{n+1} ((\cos nt \cos t - \sin nt \sin t) + i (\sin nt \cos t + \cos nt \sin t))
\]
\[
= r^{n+1} (\cos (n + 1)t + i \sin (n + 1)t)
\]
by the formulas for the cosine and sine of the sum of two angles.

**Corollary 8.0.6** Let \( z \) be a non zero complex number. Then there are always exactly \( k \) \( k^{th} \) roots of \( z \) in \( \mathbb{C} \).

**Proof:** Let \( z = x + iy \) and let \( z = |z| (\cos t + i \sin t) \) be the polar form of the complex number. By De Moivre’s theorem, a complex number,
\[
r (\cos \alpha + i \sin \alpha),
\]
is a \( k^{th} \) root of \( z \) if and only if
\[
r^k (\cos k\alpha + i \sin k\alpha) = |z| (\cos t + i \sin t).
\]
This requires \( r^k = |z| \) and so \( r = |z|^{1/k} \) and also both \( \cos (k\alpha) = \cos t \) and \( \sin (k\alpha) = \sin t \). This can only happen if
\[
k\alpha = t + 2l\pi
\]
for \( l \) an integer. Thus
\[
\alpha = \frac{t + 2l\pi}{k}, l \in \mathbb{Z}
\]
and so the \( k^{th} \) roots of \( z \) are of the form
\[
|z|^{1/k} \left( \cos \left( \frac{t + 2l\pi}{k} \right) + i \sin \left( \frac{t + 2l\pi}{k} \right) \right), l \in \mathbb{Z}.
\]
Since the cosine and sine are periodic of period \( 2\pi \), there are exactly \( k \) distinct numbers which result from this formula.
Example 8.0.7 Find the three cube roots of \(i\).

First note that \(i = 1 \left( \cos \left( \frac{\pi}{2} \right) + i \sin \left( \frac{\pi}{2} \right) \right)\). Using the formula in the proof of the above corollary, the cube roots of \(i\) are

\[
1 \left( \cos \left( \frac{(\pi/2) + 2l\pi}{3} \right) + i \sin \left( \frac{(\pi/2) + 2l\pi}{3} \right) \right)
\]

where \(l = 0, 1, 2\). Therefore, the roots are

\[
\cos \left( \frac{\pi}{6} \right) + i \sin \left( \frac{\pi}{6} \right), \cos \left( \frac{5\pi}{6} \right) + i \sin \left( \frac{5\pi}{6} \right), \text{ and } \cos \left( \frac{3\pi}{2} \right) + i \sin \left( \frac{3\pi}{2} \right).
\]

Thus the cube roots of \(i\) are \(\sqrt{3}/2 + i \left( \frac{1}{2} \right)\), \(-\sqrt{3}/2 + i \left( \frac{1}{2} \right)\), and \(-i\).

The ability to find \(k\)th roots can also be used to factor some polynomials.

Example 8.0.8 Factor the polynomial \(x^3 - 27\).

First find the cube roots of 27. By the above procedure using De Moivre’s theorem, these cube roots are \(3 \left( \frac{1}{2} + i \frac{\sqrt{3}}{2} \right)\), and \(3 \left( \frac{-1}{2} - i \frac{\sqrt{3}}{2} \right)\). Therefore, \(x^3 + 27 =

\[
(x - 3) \left( x - 3 \left( \frac{-1}{2} + i \frac{\sqrt{3}}{2} \right) \right) \left( x - 3 \left( \frac{-1}{2} - i \frac{\sqrt{3}}{2} \right) \right).
\]

Note also \(x - 3 \left( \frac{1}{2} + i \frac{\sqrt{3}}{2} \right) \left( x - 3 \left( \frac{1}{2} - i \frac{\sqrt{3}}{2} \right) \right) = x^2 + 3x + 9\) and so

\[
x^3 - 27 = (x - 3) (x^2 + 3x + 9)
\]

where the quadratic polynomial, \(x^2 + 3x + 9\) cannot be factored without using complex numbers.

Note that even though the polynomial \(x^3 - 27\) has all real coefficients, it has some complex zeros, \(\frac{1}{2} + i \frac{\sqrt{3}}{2}\) and \(\frac{-1}{2} - i \frac{\sqrt{3}}{2}\). These zeros are complex conjugates of each other. It is always this way. You should show this is the case. To see how to do this, see Problems 13 and 14 below.

Another fact for your information is the fundamental theorem of algebra. This theorem says that any polynomial of degree at least 1 having any complex coefficients always has a root in \(\mathbb{C}\). This is sometimes referred to by saying \(\mathbb{C}\) is algebraically complete. This is proved in an appendix. Gauss is usually credited with giving a proof of this theorem in 1797 but many others worked on it and the first completely correct proof was due to Argand in 1806. For more on this theorem, you can google fundamental theorem of algebra and look at the interesting Wikipedia article on it.

Theorems 3.4.2 and 3.4.4 remain true in the context of polynomials having coefficients in \(\mathbb{C}\) because as mentioned above \(\mathbb{C}\) is a field.

8.1 Exercises

1. Let \(z = 5 + i9\). Find \(z^{-1}\).

2. Let \(z = 2 + i7\) and let \(w = 3 - i8\). Find \(zw, z + w, z^2,\) and \(w/z\).
3. Give the complete solution to \( x^4 + 16 = 0 \).

4. Graph the complex cube roots of 8 in the complex plane. Do the same for the four fourth roots of 16.

5. If \( z \) is a complex number, show there exists \( \omega \) a complex number with \( |\omega| = 1 \) and \( \omega z = |z| \).

6. De Moivre’s theorem says \([r (\cos t + i \sin t)]^n = r^n (\cos nt + i \sin nt)\) for \( n \) a positive integer. Does this formula continue to hold for all integers, \( n \), even negative integers? Explain.

7. You already know formulas for \( \cos (x + y) \) and \( \sin (x + y) \) and these were used to prove De Moivre’s theorem. Now using De Moivre’s theorem, derive a formula for \( \sin (5x) \) and one for \( \cos (5x) \). \( \textbf{Hint:} \) Use Problem 18 on Page 25 and if you like, you might use Pascal’s triangle to construct the binomial coefficients.

8. If \( z \) and \( w \) are two complex numbers and the polar form of \( z \) involves the angle \( \theta \) while the polar form of \( w \) involves the angle \( \phi \), show that in the polar form for \( zw \) the angle involved is \( \theta + \phi \). Also, show that in the polar form of a complex number, \( z, r = |z| \).

9. Factor \( x^3 + 8 \) as a product of linear factors.

10. Write \( x^3 + 27 \) in the form \((x + 3) (x^2 + ax + b)\) where \( x^2 + ax + b \) cannot be factored any more using only real numbers.

11. Completely factor \( x^4 + 16 \) as a product of linear factors.

12. Factor \( x^4 + 16 \) as the product of two quadratic polynomials each of which cannot be factored further without using complex numbers.

13. If \( z, w \) are complex numbers prove \( \overline{\overline{z}} = \overline{z} \) and then show by induction that \( \overline{z_1 \cdots z_m} = \overline{z_1} \cdots \overline{z_m} \). Also verify that \( \sum_{k=1}^{m} \overline{z_k} = \overline{\sum_{k=1}^{m} z_k} \). In words this says the conjugate of a product equals the product of the conjugates and the conjugate of a sum equals the sum of the conjugates.

14. Suppose \( p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \) where all the \( a_k \) are real numbers. Suppose also that \( p(z) = 0 \) for some \( z \in \mathbb{C} \). Show it follows that \( p(\overline{z}) = 0 \) also.

15. Show that \( 1 + i, 2 + i \) are the only two zeros to
\[
p(x) = x^2 - (3 + 2i) x + (1 + 3i)
\]
so the zeros do not necessarily come in conjugate pairs if the coefficients are not real.

16. I claim that \( 1 = -1 \). Here is why.
\[
-1 = i^2 = \sqrt{-1} \sqrt{-1} = \sqrt{(-1)^2} = \sqrt{1} = 1.
\]
This is clearly a remarkable result but is there something wrong with it? If so, what is wrong?

17. De Moivre’s theorem is really a grand thing. I plan to use it now for rational exponents, not just integers.
\[
1 = 1^{(1/4)} = (\cos 2\pi + i \sin 2\pi)^{1/4} = \cos (\pi/2) + i \sin (\pi/2) = i.
\]
Therefore, squaring both sides it follows $1 = -1$ as in the previous problem. What does this tell you about De Moivre’s theorem? Is there a profound difference between raising numbers to integer powers and raising numbers to non integer powers?

18. Review Problem 6 at this point. Now here is another question: If $n$ is an integer, is it always true that $(\cos \theta - i \sin \theta)^n = \cos (n\theta) - i \sin (n\theta)$? Explain.

19. Suppose you have any polynomial in $\cos \theta$ and $\sin \theta$. By this I mean an expression of the form $\sum_{\alpha=0}^{m} \sum_{\beta=0}^{n} a_{\alpha\beta} \cos^\alpha \theta \sin^\beta \theta$ where $a_{\alpha\beta} \in \mathbb{C}$. Can this always be written in the form $\sum_{\gamma=-n-m}^{n+m} b_{\gamma} \cos \gamma \theta + \sum_{\tau=-n-m}^{n+m} c_{\tau} \sin \tau \theta$? Explain.

20. Suppose $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ is a polynomial and it has $n$ zeros, 
$$z_1, z_2, \ldots, z_n$$
listed according to multiplicity. ($z$ is a root of multiplicity $m$ if the polynomial $f(x) = (x - z)^m$ divides $p(x)$ but $(x - z)^{m+1}$ does not.) Show that 
$$p(x) = a_n (x - z_1) (x - z_2) \cdots (x - z_n).$$

21. Prove the fundamental theorem of algebra for quadratic polynomials having coefficients in $\mathbb{C}$. 
The notation, \( \mathbb{C}^n \) refers to the collection of ordered lists of \( n \) complex numbers. Since every real number is also a complex number, this simply generalizes the usual notion of \( \mathbb{R}^n \), the collection of all ordered lists of \( n \) real numbers. In order to avoid worrying about whether it is real or complex numbers which are being referred to, the symbol \( \mathbb{F} \) will be used. If it is not clear, always pick \( \mathbb{C} \).

**Definition 9.0.1** Define \( \mathbb{F}^n \equiv \{(x_1, \ldots , x_n) : x_j \in \mathbb{F} \text{ for } j = 1, \ldots , n\} \).

\[ (x_1, \ldots , x_n) = (y_1, \ldots , y_n) \]

if and only if for all \( j = 1, \ldots , n, \) \( x_j = y_j \). When \( (x_1, \ldots , x_n) \in \mathbb{F}^n \), it is conventional to denote \( (x_1, \ldots , x_n) \) by the single bold face letter, \( \mathbf{x} \). The numbers, \( x_j \) are called the coordinates. The set \( \{(0, \ldots , 0, t, 0, \ldots , 0) : t \in \mathbb{F}\} \) for \( t \) in the \( i \)th slot is called the \( i \)th coordinate axis. The point \( \mathbf{0} \equiv (0, \ldots , 0) \) is called the origin. Elements in \( \mathbb{F}^n \) are called vectors.

Thus \( (1, 2, 4i) \in \mathbb{F}^3 \) and \( (2, 1, 4i) \in \mathbb{F}^3 \) but \( (1, 2, 4i) \neq (2, 1, 4i) \) because, even though the same numbers are involved, they don’t match up. In particular, the first entries are not equal.

The geometric significance of \( \mathbb{R}^n \) for \( n \leq 3 \) has been encountered already in calculus or in pre-calculus. Here is a short review. First consider the case when \( n = 1 \). Then from the definition, \( \mathbb{R}^1 = \mathbb{R} \). Recall that \( \mathbb{R} \) is identified with the points of a line. Look at the number line again. Observe that this amounts to identifying a point on this line with a real number. In other words a real number determines where you are on this line. Now suppose \( n = 2 \) and consider two lines which intersect each other at right angles as shown in the following picture.

Notice how you can identify a point shown in the plane with the ordered pair, \( (2, 6) \). You go to the right a distance of 2 and then up a distance of 6. Similarly, you can identify another point in the plane with the ordered pair \( (-8, 3) \). Go to the left a distance of 8.
and then up a distance of 3. The reason you go to the left is that there is a − sign on the
eight. From this reasoning, every ordered pair determines a unique point in the plane.
Conversely, taking a point in the plane, you could draw two lines through the point,
one vertical and the other horizontal and determine unique points, \( x_1 \) on the horizontal
line in the above picture and \( x_2 \) on the vertical line in the above picture, such that
the point of interest is identified with the ordered pair, \((x_1, x_2)\). In short, points in
the plane can be identified with ordered pairs similar to the way that points on the real
line are identified with real numbers. Now suppose \( n = 3 \). As just explained, the first
two coordinates determine a point in a plane. Letting the third component determine
how far up or down you go, depending on whether this number is positive or negative,
this determines a point in space. Thus, \((1, 4, -5)\) would mean to determine the point
in the plane that goes with \((1, 4)\) and then to go below this plane a distance of 5 to
obtain a unique point in space. You see that the ordered triples correspond to points
in space just as the ordered pairs correspond to points in a plane and single real numbers
correspond to points on a line.

You can’t stop here and say that you are only interested in \( n \leq 3 \). What if you were
interested in the motion of two objects? You would need three coordinates to describe
where the first object is and you would need another three coordinates to describe
where the other object is located. Therefore, you would need to be considering \( \mathbb{R}^6 \). If
the two objects moved around, you would need a time coordinate as well. As another
example, consider a hot object which is cooling and suppose you want the temperature
of this object. How many coordinates would be needed? You would need one for the
temperature, three for the position of the point in the object and one more for the
time. Thus you would need to be considering \( \mathbb{R}^5 \). Many other examples can be given.
Sometimes \( n \) is very large. This is often the case in applications to business when they
are trying to maximize profit subject to constraints. It also occurs in numerical analysis
when people try to solve hard problems on a computer.

There are other ways to identify points in space with three numbers but the one
presented is the most basic. In this case, the coordinates are known as Cartesian
coordinates after Descartes\(^1\) who invented this idea in the first half of the seventeenth
century. I will often not bother to draw a distinction between the point in space and
its Cartesian coordinates.

The geometric significance of \( \mathbb{C}^n \) for \( n > 1 \) is not available because each copy of \( \mathbb{C} \)
corresponds to the plane or \( \mathbb{R}^2 \).

### 9.1 Algebra in \( \mathbb{F}^n \)

There are two algebraic operations done with elements of \( \mathbb{F}^n \). One is addition and the
other is multiplication by numbers, called scalars. In the case of \( \mathbb{C}^n \) the scalars are
complex numbers while in the case of \( \mathbb{R}^n \) the only allowed scalars are real numbers.
Thus, the scalars always come from \( \mathbb{F} \) in either case.

**Definition 9.1.1** If \( \mathbf{x} \in \mathbb{F}^n \) and \( a \in \mathbb{F} \), also called a scalar, then \( a\mathbf{x} \in \mathbb{F}^n \) is defined by
\[
    a\mathbf{x} = a(x_1, \ldots, x_n) \equiv (ax_1, \ldots, ax_n).
\]

This is known as scalar multiplication. If \( \mathbf{x}, \mathbf{y} \in \mathbb{F}^n \) then \( \mathbf{x} + \mathbf{y} \in \mathbb{F}^n \) and is defined by
\[
    \mathbf{x} + \mathbf{y} = (x_1, \ldots, x_n) + (y_1, \ldots, y_n) \\
    \equiv (x_1 + y_1, \ldots, x_n + y_n)
\]

\(^1\) René Descartes 1596-1650 is often credited with inventing analytic geometry although it seems
the ideas were actually known much earlier. He was interested in many different subjects, physiology,
chemistry, and physics being some of them. He also wrote a large book in which he tried to explain
9.2. GEOMETRIC MEANING OF VECTORS

$\mathbb{F}^n$ is often called $n$ dimensional space. With this definition, the algebraic properties satisfy the conclusions of the following theorem.

**Theorem 9.1.2** For $v, w \in \mathbb{F}^n$ and $\alpha, \beta$ scalars, (real numbers), the following hold.

$$v + w = w + v,$$  \hspace{1cm} (9.3)

the commutative law of addition,

$$(v + w) + z = v + (w + z),$$ \hspace{1cm} (9.4)

the associative law for addition,

$$v + 0 = v,$$ \hspace{1cm} (9.5)

the existence of an additive identity,

$$v + (-v) = 0,$$ \hspace{1cm} (9.6)

the existence of an additive inverse, Also

$$\alpha (v + w) = \alpha v + \alpha w,$$ \hspace{1cm} (9.7)

$$(\alpha + \beta) v = \alpha v + \beta v,$$ \hspace{1cm} (9.8)

$$\alpha (\beta v) = \alpha \beta (v),$$ \hspace{1cm} (9.9)

$$1v = v.$$ \hspace{1cm} (9.10)

In the above $0 = (0, \cdots, 0)$.

You should verify these properties all hold. For example, consider $9.7$

$$\alpha (v + w) = \alpha (v_1 + w_1, \cdots, v_n + w_n)$$
$$= (\alpha v_1 + w_1, \cdots, \alpha v_n + w_n)$$
$$= \alpha v + \alpha w.$$

As usual subtraction is defined as $x - y \equiv x + (-y)$.

### 9.2 Geometric Meaning Of Vectors

The geometric meaning is especially significant in the case of $\mathbb{R}^n$ for $n = 2, 3$. Here is a short discussion of this topic.

**Definition 9.2.1** Let $x = (x_1, \cdots, x_n)$ be the coordinates of a point in $\mathbb{R}^n$. Imagine an arrow with its tail at $0 = (0, \cdots, 0)$ and its point at $x$ as shown in the following picture in the case of $\mathbb{R}^3$.

![Position Vector](image)

$(x_1, x_2, x_3) = x$

Then this arrow is called the **position vector** of the point, $x$. Given two points, $P, Q$ whose coordinates are $(p_1, \cdots, p_n)$ and $(q_1, \cdots, q_n)$ respectively, one can also determine the position vector from $P$ to $Q$ defined as follows.

$$\overrightarrow{PQ} \equiv (q_1 - p_1, \cdots, q_n - p_n)$$
Thus every point determines a vector and conversely, every such vector (arrow) which has its tail at 0 determines a point of \( \mathbb{R}^n \), namely the point of \( \mathbb{R}^n \) which coincides with the point of the vector. Also two different points determine a position vector going from one to the other as just explained.

Imagine taking the above position vector and moving it around, always keeping it pointing in the same direction as shown in the following picture.

After moving it around, it is regarded as the same vector because it points in the same direction and has the same length. Thus each of the arrows in the above picture is regarded as the same vector. The components of this vector are the numbers, \( x_1, \ldots, x_n \). You should think of these numbers as directions for obtaining an arrow. Starting at some point, \( (a_1, a_2, \ldots, a_n) \) in \( \mathbb{R}^n \), you move to the point \( (a_1 + x_1, \ldots, a_n) \) and then move from there to the point \( (a_1 + x_1, a_2 + x_2, a_3, \ldots, a_n) \) and then to \( (a_1 + x_1, a_2 + x_2, a_3 + x_3 + 1, \ldots, a_n) \) and continue this way until you obtain the point \( (a_1 + x_1, a_2 + x_2, \ldots, a_n + x_n) \). The arrow having its tail at \( (a_1, a_2, \ldots, a_n) \) and its point at \( (a_1 + x_1, a_2 + x_2, \ldots, a_n + x_n) \) looks just like the arrow which has its tail at 0 and its point at \( (x_1, \ldots, x_n) \) so it is regarded as the same vector.

### 9.3 Geometric Meaning Of Vector Addition

It was explained earlier that an element of \( \mathbb{R}^n \) is an \( n \) tuple of numbers and it was also shown that this can be used to determine a point in three dimensional space in the case where \( n = 3 \) and in two dimensional space, in the case where \( n = 2 \). This point was specified relative to some coordinate axes.

Consider the case where \( n = 3 \) for now. If you draw an arrow from the point in three dimensional space determined by \( (0, 0, 0) \) to the point \( (a, b, c) \) with its tail sitting at the point \( (0, 0, 0) \) and its point at the point \( (a, b, c) \), this arrow is called the position vector of the point determined by \( u \equiv (a, b, c) \). One way to get to this point is to start at \( (0, 0, 0) \) and move in the direction of the \( x_1 \) axis to \( (a, 0, 0) \) and then in the direction of the \( x_2 \) axis to \( (a, b, 0) \) and finally in the direction of the \( x_3 \) axis to \( (a, b, c) \). It is evident that the same arrow (vector) would result if you began at the point, \( v \equiv (d, e, f) \), moved in the direction of the \( x_1 \) axis to \( (d + a, e, f) \), then in the direction of the \( x_2 \) axis to \( (d + a, e + b, f) \), and finally in the direction of the \( x_3 \) direction to \( (d + a, e + b, f + c) \) only this time, the arrow would have its tail sitting at the point determined by \( v \equiv (d, e, f) \) and its point at \( (d + a, e + b, f + c) \). It is said to be the same arrow (vector) because it will point in the same direction and have the same length. It is like you took an actual arrow, the sort of thing you shoot with a bow, and moved it from one location to another keeping it pointing the same direction. This is illustrated in the following picture in which \( v + u \) is illustrated. Note the parallelogram determined in the picture by the vectors \( u \) and \( v \).
Thus the geometric significance of \((d, e, f) + (a, b, c) = (d + a, e + b, f + c)\) is this. You start with the position vector of the point \((d, e, f)\) and at its point, you place the vector determined by \((a, b, c)\) with its tail at \((d, e, f)\). Then the point of this last vector will be \((d + a, e + b, f + c)\). This is the geometric significance of vector addition. Also, as shown in the picture, \(u + v\) is the directed diagonal of the parallelogram determined by the two vectors \(u\) and \(v\). A similar interpretation holds in \(\mathbb{R}^n, n > 3\) but I can’t draw a picture in this case.

Since the convention is that identical arrows pointing in the same direction represent the same vector, the geometric significance of vector addition is as follows in any number of dimensions.

**Procedure 9.3.1** Let \(u\) and \(v\) be two vectors. Slide \(v\) so that the tail of \(v\) is on the point of \(u\). Then draw the arrow which goes from the tail of \(u\) to the point of the slid vector, \(v\). This arrow represents the vector \(u + v\).

Note that \(P + \overrightarrow{PQ} = Q\).

### 9.4 Distance Between Points In \(\mathbb{R}^n\) Length Of A Vector

How is distance between two points in \(\mathbb{R}^n\) defined?

**Definition 9.4.1** Let \(x = (x_1, \cdots, x_n)\) and \(y = (y_1, \cdots, y_n)\) be two points in \(\mathbb{R}^n\). Then \(|x - y|\) to indicates the distance between these points and is defined as

\[
|\mathbf{x} - \mathbf{y}| \equiv \sqrt{\sum_{k=1}^{n} |x_k - y_k|^2}.
\]

This is called the **distance formula**. Thus \(|\mathbf{x}| \equiv |\mathbf{x} - \mathbf{0}|\). The symbol, \(B(\mathbf{a}, r)\) is defined by

\[
B(\mathbf{a}, r) \equiv \{ \mathbf{x} \in \mathbb{R}^n : |\mathbf{x} - \mathbf{a}| < r \}.
\]
This is called an open ball of radius $r$ centered at $a$. It means all points in $\mathbb{R}^n$ which are closer to $a$ than $r$. The length of a vector $x$ is the distance between $x$ and $0$.

First of all note this is a generalization of the notion of distance in $\mathbb{R}$. There the distance between two points, $x$ and $y$ was given by the absolute value of their difference. Thus $|x - y|$ is equal to the distance between these two points on $\mathbb{R}$. Now $|x - y| = \left( (x - y)^2 \right)^{1/2}$ where the square root is always the positive square root. Thus it is the same formula as the above definition except there is only one term in the sum. Geometrically, this is the right way to define distance which is seen from the Pythagorean theorem. Often people use two lines to denote this distance, $||x - y||$. However, I want to emphasize this is really just like the absolute value. Also, the notation I am using is fairly standard.

Consider the following picture in the case that $n = 2$.

There are two points in the plane whose Cartesian coordinates are $(x_1, x_2)$ and $(y_1, y_2)$ respectively. Then the solid line joining these two points is the hypotenuse of a right triangle which is half of the rectangle shown in dotted lines. What is its length? Note the lengths of the sides of this triangle are $|y_1 - x_1|$ and $|y_2 - x_2|$. Therefore, the Pythagorean theorem implies the length of the hypotenuse equals

$$\left( |y_1 - x_1|^2 + |y_2 - x_2|^2 \right)^{1/2} = \left( (y_1 - x_1)^2 + (y_2 - x_2)^2 \right)^{1/2}$$

which is just the formula for the distance given above. In other words, this distance defined above is the same as the distance of plane geometry in which the Pythagorean theorem holds.

Now suppose $n = 3$ and let $(x_1, x_2, x_3)$ and $(y_1, y_2, y_3)$ be two points in $\mathbb{R}^3$. Consider the following picture in which one of the solid lines joins the two points and a dotted line joins the points $(x_1, x_2, x_3)$ and $(y_1, y_2, x_3)$.

By the Pythagorean theorem, the length of the dotted line joining $(x_1, x_2, x_3)$ and $(y_1, y_2, x_3)$ equals

$$\left( (y_1 - x_1)^2 + (y_2 - x_2)^2 \right)^{1/2}$$
9.4. DISTANCE BETWEEN POINTS IN $\mathbb{R}^N$ LENGTH OF A VECTOR

while the length of the line joining $(y_1, y_2, x_3)$ to $(y_1, y_2, y_3)$ is just $|y_3 - x_3|$. Therefore, by the Pythagorean theorem again, the length of the line joining the points $(x_1, x_2, x_3)$ and $(y_1, y_2, y_3)$ equals

$$
\left\{ \left[ (y_1 - x_1)^2 + (y_2 - x_2)^2 \right]^{1/2} + (y_3 - x_3)^2 \right\}^{1/2},
$$

which is again just the distance formula above.

This completes the argument that the above definition is reasonable. Of course you cannot continue drawing pictures in ever higher dimensions but there is no problem with the formula for distance in any number of dimensions. Here is an example.

Example 9.4.2 Find the distance between the points in $\mathbb{R}^4$, $a = (1, 2, -4, 6)$ and $b = (2, 3, -1, 0)$

Use the distance formula and write

$$
|a - b|^2 = (1 - 2)^2 + (2 - 3)^2 + (6 - 0)^2 = 47
$$

Therefore, $|a - b| = \sqrt{47}$.

All this amounts to defining the distance between two points as the length of a straight line joining these two points. However, there is nothing sacred about using straight lines. One could define the distance to be the length of some other sort of line joining these points. It won’t be done in this book but sometimes this sort of thing is done.

Another convention which is usually followed, especially in $\mathbb{R}^2$ and $\mathbb{R}^3$ is to denote the first component of a point in $\mathbb{R}^2$ by $x$ and the second component by $y$. In $\mathbb{R}^3$ it is customary to denote the first and second components as just described while the third component is called $z$.

Example 9.4.3 Describe the points which are at the same distance between $(1, 2, 3)$ and $(0, 1, 2)$.

Let $(x, y, z)$ be such a point. Then

$$
\sqrt{(x - 1)^2 + (y - 2)^2 + (z - 3)^2} = \sqrt{x^2 + (y - 1)^2 + (z - 2)^2}.
$$

Squaring both sides

$$(x - 1)^2 + (y - 2)^2 + (z - 3)^2 = x^2 + (y - 1)^2 + (z - 2)^2$$

and so

$$x^2 - 2x + 14 + y^2 - 4y + z^2 - 6z = x^2 + y^2 - 2y + 5 + z^2 - 4z$$

which implies

$$-2x + 14 - 4y - 6z = -2y + 5 - 4z$$

and so

$$2x + 2y + 2z = -9. \quad (9.11)$$

Since these steps are reversible, the set of points which is at the same distance from the two given points consists of the points, $(x, y, z)$ such that $(9.11)$ holds.

There are certain properties of the distance which are obvious. Two of them which follow directly from the definition are

$$|x - y| = |y - x|,$$
\[ |x - y| \geq 0 \text{ and equals } 0 \text{ only if } y = x. \]

The third fundamental property of distance is known as the triangle inequality. Recall that in any triangle the sum of the lengths of two sides is always at least as large as the third side. I will show you a proof of this later. This is usually stated as

\[ |x + y| \leq |x| + |y|. \]

Here is a picture which illustrates the statement of this inequality in terms of geometry.

---

9.5 Geometric Meaning Of Scalar Multiplication

As discussed earlier, \( x = (x_1, x_2, x_3) \) determines a vector. You draw the line from 0 to \( x \) placing the point of the vector on \( x \). What is the length of this vector? The length of this vector is defined to equal \( |x| \) as in Definition 9.4.1. Thus the length of \( x \) equals \( \sqrt{x_1^2 + x_2^2 + x_3^2} \). When you multiply \( x \) by a scalar, \( \alpha \), you get \((\alpha x_1, \alpha x_2, \alpha x_3)\) and the length of this vector is defined as

\[
\sqrt{(\alpha x_1)^2 + (\alpha x_2)^2 + (\alpha x_3)^2} = |\alpha| \sqrt{x_1^2 + x_2^2 + x_3^2}.
\]

Thus the following holds.

\[ |\alpha x| = |\alpha| |x|. \]

In other words, multiplication by a scalar magnifies the length of the vector. What about the direction? You should convince yourself by drawing a picture that if \( \alpha \) is negative, it causes the resulting vector to point in the opposite direction while if \( \alpha > 0 \) it preserves the direction the vector points.

You can think of vectors as quantities which have direction and magnitude, little arrows. Thus any two little arrows which have the same length and point in the same direction are considered to be the same vector even if their tails are at different points.

You can always slide such an arrow and place its tail at the origin. If the resulting point of the vector is \((a, b, c)\), it is clear the length of the little arrow is \( \sqrt{a^2 + b^2 + c^2} \).

Geometrically, the way you add two geometric vectors is to place the tail of one on the point of the other and then to form the vector which results by starting with the tail of the first and ending with this point as illustrated in the following picture. Also when \((a, b, c)\) is referred to as a vector, you mean any of the arrows which have the same direction and magnitude as the position vector of this point. Geometrically, for \( u = (u_1, u_2, u_3) \), \( \alpha u \) is any of the little arrows which have the same direction and magnitude as \((\alpha u_1, \alpha u_2, \alpha u_3)\).
The following example is art which illustrates these definitions and conventions.

**Exercise 9.5.1** Here is a picture of two vectors, $u$ and $v$.

Sketch a picture of $u + v$, $u - v$, and $u + 2v$.

First here is a picture of $u + v$. You first draw $u$ and then at the point of $u$ you place the tail of $v$ as shown. Then $u + v$ is the vector which results which is drawn in the following pretty picture.

Next consider $u - v$. This means $u + (-v)$. From the above geometric description of vector addition, $-v$ is the vector which has the same length but which points in the opposite direction to $v$. Here is a picture.

Finally consider the vector $u + 2v$. Here is a picture of this one also.

### 9.6 Exercises


2. Compute $5 (1, 2 + 3i, 3, -2) + 6 (2 - i, 1, -2, 7)$. 
3. Draw a picture of the points in $\mathbb{R}^2$ which are determined by the following ordered pairs.

   (a) $(1, 2)$
   (b) $(-2, -2)$
   (c) $(-2, 3)$
   (d) $(2, -5)$

4. Does it make sense to write $(1, 2) + (2, 3, 1)$? Explain.

5. Draw a picture of the points in $\mathbb{R}^3$ which are determined by the following ordered triples.

   (a) $(1, 2, 0)$
   (b) $(-2, -2, 1)$
   (c) $(-2, 3, -2)$

### 9.7 Vectors And Physics

Suppose you push on something. What is important? There are really two things which are important, how hard you push and the direction you push. This illustrates the concept of force.

**Definition 9.7.1** *Force* is a vector. The magnitude of this vector is a measure of how hard it is pushing. It is measured in units such as Newtons or pounds or tons. Its direction is the direction in which the push is taking place.

Vectors are used to model force and other physical vectors like velocity. What was just described would be called a force vector. It has two essential ingredients, its magnitude and its direction. Geometrically think of vectors as directed line segments or arrows as shown in the following picture in which all the directed line segments are considered to be the same vector because they have the same direction, the direction in which the arrows point, and the same magnitude (length).

![Diagram of vectors](image)

Because of this fact that only direction and magnitude are important, it is always possible to put a vector in a certain particularly simple form. Let $\overrightarrow{pq}$ be a directed line segment or vector. Then it follows that $\overrightarrow{pq}$ consists of the points of the form

$$p + t(q - p)$$

where $t \in [0, 1]$. Subtract $p$ from all these points to obtain the directed line segment consisting of the points

$$0 + t(q - p), \ t \in [0, 1].$$

The point in $\mathbb{R}^n, q - p$, will represent the vector.
Geometrically, the arrow, \( \overrightarrow{pq} \), was slid so it points in the same direction and the base is at the origin, \( \mathbf{0} \). For example, see the following picture.

In this way vectors can be identified with points of \( \mathbb{R}^n \).

**Definition 9.7.2** Let \( \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n \). The **position vector** of this point is the vector whose point is at \( \mathbf{x} \) and whose tail is at the origin, \( (0, \ldots, 0) \). If \( \mathbf{x} = (x_1, \ldots, x_n) \) is called a vector, the vector which is meant is this position vector just described. Another term associated with this is **standard position**. A vector is in standard position if the tail is placed at the origin.

It is customary to identify the point in \( \mathbb{R}^n \) with its position vector.

The magnitude of a vector determined by a directed line segment \( \overrightarrow{pq} \) is just the distance between the point \( \mathbf{p} \) and the point \( \mathbf{q} \). By the distance formula this equals

\[
\left( \sum_{k=1}^{n} (q_k - p_k)^2 \right)^{1/2} = |\mathbf{p} - \mathbf{q}|
\]

and for \( \mathbf{v} \) any vector in \( \mathbb{R}^n \) the magnitude of \( \mathbf{v} \) equals \( \left( \sum_{k=1}^{n} v_k^2 \right)^{1/2} = |\mathbf{v}| \).

**Example 9.7.3** Consider the vector, \( \mathbf{v} \equiv (1, 2, 3) \) in \( \mathbb{R}^n \). Find \( |\mathbf{v}| \).

First, the vector is the directed line segment (arrow) which has its base at \( \mathbf{0} \equiv (0, 0, 0) \) and its point at \( (1, 2, 3) \). Therefore,

\[|\mathbf{v}| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}.
\]

What is the geometric significance of scalar multiplication? If \( \mathbf{a} \) represents the vector, \( \mathbf{v} \) in the sense that when it is slid to place its tail at the origin, the element of \( \mathbb{R}^n \) at its point is \( \mathbf{a} \), what is \( r\mathbf{v} \)?

\[
|r\mathbf{v}| = \left( \sum_{k=1}^{n} (ra_k)^2 \right)^{1/2} = \left( \sum_{k=1}^{n} r^2 (a_k)^2 \right)^{1/2} = (r^2)^{1/2} \left( \sum_{k=1}^{n} a_k^2 \right)^{1/2} = |r| |\mathbf{v}|.
\]

Thus the magnitude of \( r\mathbf{v} \) equals \( |r| \) times the magnitude of \( \mathbf{v} \). If \( r \) is positive, then the vector represented by \( r\mathbf{v} \) has the same direction as the vector, \( \mathbf{v} \) because multiplying by the scalar, \( r \), only has the effect of scaling all the distances. Thus the unit distance along any coordinate axis now has length \( r \) and in this rescaled system the vector is represented by \( \mathbf{a} \). If \( r < 0 \) similar considerations apply except in this case all the \( a_i \) also change sign. From now on, \( \mathbf{a} \) will be referred to as a vector instead of an element of \( \mathbb{R}^n \) representing a vector as just described. The following picture illustrates the effect of scalar multiplication.
Note there are $n$ special vectors which point along the coordinate axes. These are

$$e_i \equiv (0, \cdots, 0, 1, 0, \cdots, 0)$$

where the 1 is in the $i^{th}$ slot and there are zeros in all the other spaces. See the picture in the case of $\mathbb{R}^3$.

The direction of $e_i$ is referred to as the $i^{th}$ direction. Given a vector, $v = (a_1, \cdots, a_n)$, $a_i e_i$ is the $i^{th}$ component of the vector. Thus

$$a_i e_i = (0, \cdots, 0, a_i, 0, \cdots, 0)$$

and so this vector gives something possibly nonzero only in the $i^{th}$ direction. Also, knowledge of the $i^{th}$ component of the vector is equivalent to knowledge of the vector because it gives the entry in the $i^{th}$ slot and for $v = (a_1, \cdots, a_n)$,

$$v = \sum_{k=1}^{n} a_k e_k.$$ 

What does addition of vectors mean physically? Suppose two forces are applied to some object. Each of these would be represented by a force vector and the two forces acting together would yield an overall force acting on the object which would also be a force vector known as the resultant. Suppose the two vectors are $a = \sum_{k=1}^{n} a_k e_k$ and $b = \sum_{k=1}^{n} b_k e_k$. Then the vector, $a$ involves a component in the $i^{th}$ direction, $a_i e_i$ while the component in the $i^{th}$ direction of $b$ is $b_i e_i$. Then it seems physically reasonable that the resultant vector should have a component in the $i^{th}$ direction equal to $(a_i + b_i) e_i$.

This is exactly what is obtained when the vectors, $a$ and $b$ are added.

$$a + b = (a_1 + b_1, \cdots, a_n + b_n).$$

$$= \sum_{i=1}^{n} (a_i + b_i) e_i.$$ 

Thus the addition of vectors according to the rules of addition in $\mathbb{R}^n$ which were presented earlier, yields the appropriate vector which duplicates the cumulative effect of all the vectors in the sum.

What is the geometric significance of vector addition? Suppose $u, v$ are vectors,

$$u = (u_1, \cdots, u_n), v = (v_1, \cdots, v_n)$$
Then \( u + v = (u_1 + v_1, \ldots, u_n + v_n) \). How can one obtain this geometrically? Consider the directed line segment, \( \overrightarrow{0u} \) and then, starting at the end of this directed line segment, follow the directed line segment \( \overrightarrow{u(u + v)} \) to its end, \( u + v \). In other words, place the vector \( u \) in standard position with its base at the origin and then slide the vector \( v \) till its base coincides with the point of \( u \). The point of this slid vector, determines \( u + v \).

To illustrate, see the following picture

Note the vector \( u + v \) is the diagonal of a parallelogram determined from the two vectors \( u \) and \( v \) and that identifying \( u + v \) with the directed diagonal of the parallelogram determined by the vectors \( u \) and \( v \) amounts to the same thing as the above procedure.

An item of notation should be mentioned here. In the case of \( \mathbb{R}^n \) where \( n \leq 3 \), it is standard notation to use \( i \) for \( e_1 \), \( j \) for \( e_2 \), and \( k \) for \( e_3 \). Now here are some applications of vector addition to some problems.

**Example 9.7.4** There are three ropes attached to a car and three people pull on these ropes. The first exerts a force of \( 2i + 3j - 2k \) Newtons, the second exerts a force of \( 3i + 5j + k \) Newtons and the third exerts a force of \( 5i - j + 2k \) Newtons. Find the total force in the direction of \( i \).

To find the total force add the vectors as described above. This gives \( 10i + 7j + k \) Newtons. Therefore, the force in the \( i \) direction is 10 Newtons.

As mentioned earlier, the Newton is a unit of force like pounds.

**Example 9.7.5** An airplane flies North East at 100 miles per hour. Write this as a vector.

A picture of this situation follows.

The vector has length 100. Now using that vector as the hypotenuse of a right triangle having equal sides, the sides should be each of length \( 100/\sqrt{2} \). Therefore, the vector would be \( 100/\sqrt{2} i + 100/\sqrt{2} j \).

This example also motivates the concept of velocity.

**Definition 9.7.6** The speed of an object is a measure of how fast it is going. It is measured in units of length per unit time. For example, miles per hour, kilometers per minute, feet per second. The velocity is a vector having the speed as the magnitude but also specifying the direction.
Thus the velocity vector in the above example is $100/\sqrt{2}\mathbf{i} + 100/\sqrt{2}\mathbf{j}$.

**Example 9.7.7** The velocity of an airplane is $100\mathbf{i} + \mathbf{j} + \mathbf{k}$ measured in kilometers per hour and at a certain instant of time its position is $(1, 2, 1)$. Here imagine a Cartesian coordinate system in which the third component is altitude and the first and second components are measured on a line from West to East and a line from South to North. Find the position of this airplane one minute later.

Consider the vector $(1, 2, 1)$, is the initial position vector of the airplane. As it moves, the position vector changes. After one minute the airplane has moved in the $\mathbf{i}$ direction a distance of $100 \times \frac{1}{60} = \frac{5}{3}$ kilometer. In the $\mathbf{j}$ direction it has moved $\frac{1}{60}$ kilometer during this same time, while it moves $\frac{1}{60}$ kilometer in the $\mathbf{k}$ direction. Therefore, the new displacement vector for the airplane is $(1, 2, 1) + \left(\frac{5}{3}, \frac{1}{60}, \frac{1}{60}\right) = \left(\frac{8}{3}, \frac{121}{60}, \frac{121}{60}\right)$

**Example 9.7.8** A certain river is one half mile wide with a current flowing at 4 miles per hour from East to West. A man swims directly toward the opposite shore from the South bank of the river at a speed of 3 miles per hour. How far down the river does he find himself when he has swam across? How far does he end up swimming?

Consider the following picture.

You should write these vectors in terms of components. The velocity of the swimmer in still water would be $3\mathbf{j}$ while the velocity of the river would be $-4\mathbf{i}$. Therefore, the velocity of the swimmer is $-4\mathbf{i} + 3\mathbf{j}$. Since the component of velocity in the direction across the river is 3, it follows the trip takes $1/6$ hour or 10 minutes. The speed at which he travels is $\sqrt{4^2 + 3^2} = 5$ miles per hour and so he travels $5 \times \frac{1}{6} = \frac{5}{6}$ miles. Now to find the distance downstream he finds himself, note that if $x$ is this distance, $x$ and $1/2$ are two legs of a right triangle whose hypotenuse equals 5/6 miles. Therefore, by the Pythagorean theorem the distance downstream is

$$\sqrt{(5/6)^2 - (1/2)^2} = \frac{2}{3}$$ miles.

### 9.8 Exercises

1. The wind blows from West to East at a speed of 50 miles per hour and an airplane which travels at 300 miles per hour in still air is heading North West. What is the velocity of the airplane relative to the ground? What is the component of this velocity in the direction North?

2. In the situation of Problem 1 how many degrees to the West of North should the airplane head in order to fly exactly North. What will be the speed of the airplane relative to the ground?

3. In the situation of 2 suppose the airplane uses 34 gallons of fuel every hour at that air speed and that it needs to fly North a distance of 600 miles. Will the airplane have enough fuel to arrive at its destination given that it has 63 gallons of fuel?
4. An airplane is flying due north at 150 miles per hour. A wind is pushing the airplane due east at 40 miles per hour. After 1 hour, the plane starts flying 30° East of North. Assuming the plane starts at (0, 0), where is it after 2 hours? Let North be the direction of the positive y axis and let East be the direction of the positive x axis.

5. City A is located at the origin while city B is located at (300, 500) where distances are in miles. An airplane flies at 250 miles per hour in still air. This airplane wants to fly from city A to city B but the wind is blowing in the direction of the positive y axis at a speed of 50 miles per hour. Find a unit vector such that if the plane heads in this direction, it will end up at city B having flown the shortest possible distance. How long will it take to get there?

6. A certain river is one half mile wide with a current flowing at 2 miles per hour from East to West. A man swims directly toward the opposite shore from the South bank of the river at a speed of 3 miles per hour. How far down the river does he find himself when he has swam across? How far does he end up swimming?

7. A certain river is one half mile wide with a current flowing at 2 miles per hour from East to West. A man can swim at 3 miles per hour in still water. In what direction should he swim in order to travel directly across the river? What would the answer to this problem be if the river flowed at 3 miles per hour and the man could swim only at the rate of 2 miles per hour?

8. Three forces are applied to a point which does not move. Two of the forces are \(2\mathbf{i} + \mathbf{j} + 3\mathbf{k}\) Newtons and \(\mathbf{i} - 3\mathbf{j} + 2\mathbf{k}\) Newtons. Find the third force.

9. The total force acting on an object is to be \(2\mathbf{i} + \mathbf{j} + \mathbf{k}\) Newtons. A force of \(-\mathbf{i} + \mathbf{j} + \mathbf{k}\) Newtons is being applied. What other force should be applied to achieve the desired total force?

10. A bird flies from its nest 5 km. in the direction 60° north of east where it stops to rest on a tree. It then flies 10 km. in the direction due southeast and lands atop a telephone pole. Place an xy coordinate system so that the origin is the bird’s nest, and the positive x axis points east and the positive y axis points north. Find the displacement vector from the nest to the telephone pole.

11. A car is stuck in the mud. There is a cable stretched tightly from this car to a tree which is 20 feet long. A person grasps the cable in the middle and pulls with a force of 100 pounds perpendicular to the stretched cable. The center of the cable moves two feet and remains still. What is the tension in the cable? The tension in the cable is the force exerted on this point by the part of the cable nearer the car as well as the force exerted on this point by the part of the cable nearer the tree.

9.9 Exercises With Answers

1. The wind blows from West to East at a speed of 30 kilometers per hour and an airplane which travels at 300 Kilometers per hour in still air is heading North West. What is the velocity of the airplane relative to the ground? What is the component of this velocity in the direction North?

   Let the positive y axis point in the direction North and let the positive x axis point in the direction East. The velocity of the wind is \(30\mathbf{i}\). The plane moves in the direction \(\mathbf{i} + \mathbf{j}\). A unit vector in this direction is \(\frac{1}{\sqrt{2}}(\mathbf{i} + \mathbf{j})\). Therefore, the velocity
of the plane relative to the ground is \(30\mathbf{i} + \frac{300}{\sqrt{2}} (\mathbf{i} + \mathbf{j}) = 150\sqrt{2} \mathbf{j} + (30 + 150\sqrt{2}) \mathbf{i}\). The component of velocity in the direction North is 150\sqrt{2}.

2. In the situation of Problem 1 how many degrees to the West of North should the airplane head in order to fly exactly North. What will be the speed of the airplane relative to the ground?

In this case the unit vector will be \(-\sin (\theta) \mathbf{i} + \cos (\theta) \mathbf{j}\). Therefore, the velocity of the plane will be

\[
300 (-\sin (\theta) \mathbf{i} + \cos (\theta) \mathbf{j})
\]

and this is supposed to satisfy

\[
300 (-\sin (\theta) \mathbf{i} + \cos (\theta) \mathbf{j}) + 30\mathbf{i} = 0\mathbf{i} + \mathbf{j}.
\]

Therefore, you need to have \(\sin \theta = 1/10\), which means \(\theta = .100 \text{ 17 radians}\). Therefore, the degrees should be \(\frac{1 \times 180}{\pi} = 5.729 \text{ 6 degrees}\). In this case the velocity vector of the plane relative to the ground is 300 \(\frac{\sqrt{99}}{10}\) \mathbf{j}.

3. In the situation of 2 suppose the airplane uses 34 gallons of fuel every hour at that air speed and that it needs to fly North a distance of 600 miles. Will the airplane have enough fuel to arrive at its destination given that it has 63 gallons of fuel?

The airplane needs to fly 600 miles at a speed of 300 \(\frac{\sqrt{99}}{10}\) \mathbf{j}. Therefore, it takes 2.0101 hours to get there. Therefore, the plane will need to use about 68 gallons of gas. It won’t make it.

4. A certain river is one half mile wide with a current flowing at 3 miles per hour from East to West. A man swims directly toward the opposite shore from the South bank of the river at a speed of 2 miles per hour. How far down the river does he find himself when he has swam across? How far does he end up swimming?

The velocity of the man relative to the earth is then \(-3 \mathbf{i} + 2 \mathbf{j}\). Since the component of \(\mathbf{j}\) equals 2 it follows he takes 1/8 of an hour to get across. During this time he is swept downstream at the rate of 3 miles per hour and so he ends up 3/8 of a mile down stream. He has gone \(\sqrt{(\frac{3}{8})^2 + (\frac{1}{2})^2} = .625 \text{ miles in all}\).

5. Three forces are applied to a point which does not move. Two of the forces are \(2\mathbf{i} - \mathbf{j} + 3\mathbf{k}\) Newtons and \(\mathbf{i} - 3\mathbf{j} - 2\mathbf{k}\) Newtons. Find the third force.

Call it \(a\mathbf{i} + b\mathbf{j} + c\mathbf{k}\). Then you need 

\[
a + 2 + 1 = 0, \quad b - 1 - 3 = 0, \quad \text{and} \quad c + 3 - 2 = 0.
\]

Therefore, the force is \(-3\mathbf{i} + 4\mathbf{j} - \mathbf{k}\).
Vector Products

10.1 The Dot Product

There are two ways of multiplying vectors which are of great importance in applications. The first of these is called the dot product, also called the scalar product and sometimes the inner product.

**Definition 10.1.1** Let \( \mathbf{a}, \mathbf{b} \) be two vectors in \( \mathbb{R}^n \) define \( \mathbf{a} \cdot \mathbf{b} \) as

\[
\mathbf{a} \cdot \mathbf{b} \equiv \sum_{k=1}^{n} a_k b_k.
\]

With this definition, there are several important properties satisfied by the dot product. In the statement of these properties, \( \alpha \) and \( \beta \) will denote scalars and \( \mathbf{a}, \mathbf{b}, \mathbf{c} \) will denote vectors.

**Proposition 10.1.2** The dot product satisfies the following properties.

\[
\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} \tag{10.1}
\]

\[
\mathbf{a} \cdot \mathbf{a} \geq 0 \text{ and equals zero if and only if } \mathbf{a} = \mathbf{0} \tag{10.2}
\]

\[
(\alpha \mathbf{a} + \beta \mathbf{b}) \cdot \mathbf{c} = \alpha (\mathbf{a} \cdot \mathbf{c}) + \beta (\mathbf{b} \cdot \mathbf{c}) \tag{10.3}
\]

\[
\mathbf{c} \cdot (\alpha \mathbf{a} + \beta \mathbf{b}) = \alpha (\mathbf{c} \cdot \mathbf{a}) + \beta (\mathbf{c} \cdot \mathbf{b}) \tag{10.4}
\]

\[
|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a} \tag{10.5}
\]

You should verify these properties. Also be sure you understand that (10.4) follows from the first three and is therefore redundant. It is listed here for the sake of convenience.

**Example 10.1.3** Find \( (1, 2, 0, -1) \cdot (0, 1, 2, 3) \).

This equals \( 0 + 2 + 0 + -3 = -1 \).

**Example 10.1.4** Find the magnitude of \( \mathbf{a} = (2, 1, 4, 2) \). That is, find \( |\mathbf{a}| \).

This is \( \sqrt{(2,1,4,2) \cdot (2,1,4,2)} = 5 \).

The dot product satisfies a fundamental inequality known as the Cauchy Schwarz inequality.
Theorem 10.1.5  The dot product satisfies the inequality
\[ |\mathbf{a} \cdot \mathbf{b}| \leq |\mathbf{a}| |\mathbf{b}|. \]  (10.6)

Furthermore equality is obtained if and only if one of \( \mathbf{a} \) or \( \mathbf{b} \) is a scalar multiple of the other.

Proof: First note that if \( \mathbf{b} = \mathbf{0} \) both sides of (10.6) equal zero and so the inequality holds in this case. Therefore, it will be assumed in what follows that \( \mathbf{b} \neq \mathbf{0} \).

Define a function of \( t \in \mathbb{R} \)
\[ f(t) = (\mathbf{a} + tb) \cdot (\mathbf{a} + tb). \]
Then by (10.2) \( f(t) \geq 0 \) for all \( t \in \mathbb{R} \). Also from (10.3), (10.4), (10.1), and (10.5)
\[ f(t) = \mathbf{a} \cdot (\mathbf{a} + tb) + t \mathbf{b} \cdot (\mathbf{a} + tb) \]
\[ = |\mathbf{a}|^2 + 2|\mathbf{a} \cdot \mathbf{b}| + t^2 |\mathbf{b}|^2. \]
Now this means the graph, \( y = f(t) \) is a polynomial which opens up and either its vertex touches the \( t \) axis or else the entire graph is above the \( x \) axis. In the first case, there exists some \( t \) where \( f(t) = 0 \) and this requires \( \mathbf{a} + tb = \mathbf{0} \) so one vector is a multiple of the other. Then clearly equality holds in (10.6) In the case where \( \mathbf{b} \) is not a multiple of \( \mathbf{a} \), it follows \( f(t) > 0 \) for all \( t \) which says \( f(t) \) has no real zeros and so from the quadratic formula,
\[ (2(\mathbf{a} \cdot \mathbf{b}))^2 - 4|\mathbf{a}|^2 |\mathbf{b}|^2 < 0 \]
which is equivalent to \( |\mathbf{a} \cdot \mathbf{b}| < |\mathbf{a}| |\mathbf{b}| \). This proves the theorem.

You should note that the entire argument was based only on the properties of the dot product listed in (10.1 - 10.5). This means that whenever something satisfies these properties, the Cauchy Schwartz inequality holds. There are many other instances of these properties besides vectors in \( \mathbb{R}^n \).

The Cauchy Schwartz inequality allows a proof of the triangle inequality for distances in \( \mathbb{R}^n \) in much the same way as the triangle inequality for the absolute value.

Theorem 10.1.6  (Triangle inequality) For \( \mathbf{a}, \mathbf{b} \in \mathbb{R}^n \)
\[ |\mathbf{a} + \mathbf{b}| \leq |\mathbf{a}| + |\mathbf{b}| \]  (10.7)

and equality holds if and only if one of the vectors is a nonnegative scalar multiple of the other. Also
\[ ||\mathbf{a} - \mathbf{b}|| \leq |\mathbf{a} - \mathbf{b}| \]  (10.8)

Proof: By properties of the dot product and the Cauchy Schwartz inequality,
\[ |\mathbf{a} + \mathbf{b}|^2 = (\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} + \mathbf{b}) \]
\[ = (\mathbf{a} \cdot \mathbf{a}) + (\mathbf{a} \cdot \mathbf{b}) + (\mathbf{b} \cdot \mathbf{a}) + (\mathbf{b} \cdot \mathbf{b}) \]
\[ = |\mathbf{a}|^2 + 2|\mathbf{a} \cdot \mathbf{b}| + |\mathbf{b}|^2 \]
\[ \leq |\mathbf{a}|^2 + 2|\mathbf{a} \cdot \mathbf{b}| + |\mathbf{b}|^2 \]
\[ \leq |\mathbf{a}|^2 + 2|\mathbf{a}| |\mathbf{b}| + |\mathbf{b}|^2 \]
\[ = (|\mathbf{a}| + |\mathbf{b}|)^2. \]
Taking square roots of both sides you obtain [10.7]

It remains to consider when equality occurs. If either vector equals zero, then that vector equals zero times the other vector and the claim about when equality occurs is verified. Therefore, it can be assumed both vectors are nonzero. To get equality in the second inequality above, Theorem [10.1.5] implies one of the vectors must be a multiple of the other. Say \( \mathbf{b} = \alpha \mathbf{a} \). If \( \alpha < 0 \) then equality cannot occur in the first inequality because in this case

\[
(\mathbf{a} \cdot \mathbf{b}) = \alpha |\mathbf{a}|^2 < 0 < |\alpha||\mathbf{a}|^2 = |\mathbf{a} \cdot \mathbf{b}|
\]

Therefore, \( \alpha \geq 0 \).

To get the other form of the triangle inequality,

\[
\mathbf{a} = \mathbf{a} - \mathbf{b} + \mathbf{b}
\]

so

\[
|\mathbf{a}| = |\mathbf{a} - \mathbf{b} + \mathbf{b}|
\]

\[
\leq |\mathbf{a} - \mathbf{b}| + |\mathbf{b}|.
\]

Therefore,

\[
|\mathbf{a}| - |\mathbf{b}| \leq |\mathbf{a} - \mathbf{b}|
\]

(10.9)

Similarly,

\[
|\mathbf{b}| - |\mathbf{a}| \leq |\mathbf{b} - \mathbf{a}| = |\mathbf{a} - \mathbf{b}|.
\]

(10.10)

It follows from [10.9] and [10.10] that [10.8] holds. This is because \(|\mathbf{a} - |\mathbf{b}|\) equals the left side of either [10.9] or [10.10] and either way, \(|\mathbf{a} - |\mathbf{b}|\) \leq |\mathbf{a} - \mathbf{b}|. This proves the theorem.

10.2 The Geometric Significance Of The Dot Product

10.2.1 The Angle Between Two Vectors

Given two vectors, \( \mathbf{a} \) and \( \mathbf{b} \), the included angle is the angle between these two vectors which is less than or equal to 180 degrees. The dot product can be used to determine the included angle between two vectors. To see how to do this, consider the following picture.

\[
|\mathbf{a} - \mathbf{b}|^2 = |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2 |\mathbf{a}| |\mathbf{b}| \cos \theta.
\]

Also from the properties of the dot product,

\[
|\mathbf{a} - \mathbf{b}|^2 = (\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b})
\]

\[
= |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2 \mathbf{a} \cdot \mathbf{b}
\]
and so comparing the above two formulas,

\[ \mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta. \]  

(10.11)

In words, the dot product of two vectors equals the product of the magnitude of the two vectors multiplied by the cosine of the included angle. Note this gives a geometric description of the dot product which does not depend explicitly on the coordinates of the vectors.

**Example 10.2.1** Find the angle between the vectors \(2\mathbf{i} + \mathbf{j} - \mathbf{k}\) and \(3\mathbf{i} + 4\mathbf{j} + \mathbf{k}\).

The dot product of these two vectors equals \(6 + 4 - 1 = 9\) and the norms are \(\sqrt{4+1+1} = \sqrt{6}\) and \(\sqrt{9+16+1} = \sqrt{26}\). Therefore, from (10.11) the cosine of the included angle equals

\[ \cos \theta = \frac{9}{\sqrt{26}\sqrt{6}} = .72058 \]

Now the cosine is known, the angle can be determined by solving the equation, \(\cos \theta = .72058\). This will involve using a calculator or a table of trigonometric functions. The answer is \(\theta = .76616\) radians or in terms of degrees, \(\theta = 43.898^\circ\). Recall how this last computation is done. Set up a proportion, \(\frac{x}{360} = \frac{\pi}{2\pi}\) because \(360^\circ\) corresponds to \(2\pi\) radians. However, in calculus, you should get used to thinking in terms of radians and not degrees. This is because all the important calculus formulas are defined in terms of radians.

**Example 10.2.2** Let \(\mathbf{u}, \mathbf{v}\) be two vectors whose magnitudes are equal to 3 and 4 respectively and such that if they are placed in standard position with their tails at the origin, the angle between \(\mathbf{u}\) and the positive \(x\) axis equals \(30^\circ\) and the angle between \(\mathbf{v}\) and the positive \(x\) axis is \(-30^\circ\). Find \(\mathbf{u} \cdot \mathbf{v}\).

From the geometric description of the dot product in (10.11)

\[ \mathbf{u} \cdot \mathbf{v} = 3 \times 4 \times \cos (60^\circ) = 3 \times 4 \times 1/2 = 6. \]

**Observation 10.2.3** Two vectors are said to be **perpendicular** if the included angle is \(\pi/2\) radians (\(90^\circ\)). You can tell if two nonzero vectors are perpendicular by simply taking their dot product. If the answer is zero, this means they are perpendicular because \(\cos \theta = 0\).

**Example 10.2.4** Determine whether the two vectors, \(2\mathbf{i} + \mathbf{j} - \mathbf{k}\) and \(\mathbf{i} + 3\mathbf{j} + 5\mathbf{k}\) are perpendicular.

When you take this dot product you get \(2 + 3 - 5 = 0\) and so these two are indeed perpendicular.

**Definition 10.2.5** When two lines intersect, the angle between the two lines is the smaller of the two angles determined.

**Example 10.2.6** Find the angle between the two lines,

\[ (1, 2, 0) + t (1, 2, 3) \]

and

\[ (0, 4, -3) + t (-1, 2, -3). \]
10.2. THE GEOMETRIC SIGNIFICANCE OF THE DOT PRODUCT

These two lines intersect, when \( t = 0 \) in the first and \( t = -1 \) in the second. It is only a matter of finding the angle between the direction vectors. One angle determined is given by

\[
\cos \theta = \frac{-6}{14} = \frac{-3}{7}. \tag{10.12}
\]

We don’t want this angle because it is obtuse. The angle desired is the acute angle given by

\[
\cos \theta = \frac{3}{7}.
\]

It is obtained by using replacing one of the direction vectors with \(-1\) times it.

10.2.2 Work And Projections

Our first application will be to the concept of work. The physical concept of work does not in any way correspond to the notion of work employed in ordinary conversation. For example, if you were to slide a 150 pound weight off a table which is three feet high and shuffle along the floor for 50 yards, sweating profusely and exerting all your strength to keep the weight from falling on your feet, keeping the height always three feet and then deposit this weight on another three foot high table, the physical concept of work would indicate that the force exerted by your arms did no work during this project even though the muscles in your hands and arms would likely be very tired. The reason for such an unusual definition is that even though your arms exerted considerable force on the weight, enough to keep it from falling, the direction of motion was at right angles to the force they exerted. The only part of a force which does work in the sense of physics is the component of the force in the direction of motion (This is made more precise below.). The work is defined to be the magnitude of the component of this force times the distance over which it acts in the case where this component of force points in the direction of motion and \((-1)\) times the magnitude of this component times the distance in case the force tends to impede the motion. Thus the work done by a force on an object as the object moves from one point to another is a measure of the extent to which the force contributes to the motion. This is illustrated in the following picture in the case where the given force contributes to the motion.

In this picture the force, \( \mathbf{F} \) is applied to an object which moves on the straight line from \( \mathbf{p}_1 \) to \( \mathbf{p}_2 \). There are two vectors shown, \( \mathbf{F}_\parallel \) and \( \mathbf{F}_\perp \) and the picture is intended to indicate that when you add these two vectors you get \( \mathbf{F} \) while \( \mathbf{F}_\parallel \) acts in the direction of motion and \( \mathbf{F}_\perp \) acts perpendicular to the direction of motion. Only \( \mathbf{F}_\parallel \) contributes to the work done by \( \mathbf{F} \) on the object as it moves from \( \mathbf{p}_1 \) to \( \mathbf{p}_2 \). \( \mathbf{F}_\parallel \) is called the component of the force in the direction of motion. From trigonometry, you see the magnitude of \( \mathbf{F}_\parallel \) should equal \( |\mathbf{F}| \cos \theta \). Thus, since \( \mathbf{F}_\parallel \) points in the direction of the vector from \( \mathbf{p}_1 \) to \( \mathbf{p}_2 \), the total work done should equal

\[
|\mathbf{F}| \left| \mathbf{p}_1 \mathbf{p}_2 \right| \cos \theta = |\mathbf{F}_\parallel| \left| \mathbf{p}_2 - \mathbf{p}_1 \right| \cos \theta
\]

If the included angle had been obtuse, then the work done by the force, \( \mathbf{F} \) on the object would have been negative because in this case, the force tends to impede the motion from \( \mathbf{p}_1 \) to \( \mathbf{p}_2 \) but in this case, \( \cos \theta \) would also be negative and so it is still the case
that the work done would be given by the above formula. Thus from the geometric description of the dot product given above, the work equals

$$|\mathbf{F}| |\mathbf{p}_2 - \mathbf{p}_1| \cos \theta = \mathbf{F} \cdot (\mathbf{p}_2 - \mathbf{p}_1).$$

This explains the following definition.

**Definition 10.2.7** Let \( \mathbf{F} \) be a force acting on an object which moves from the point, \( \mathbf{p}_1 \) to the point \( \mathbf{p}_2 \). Then the **work** done on the object by the given force equals \( \mathbf{F} \cdot (\mathbf{p}_2 - \mathbf{p}_1) \).

The concept of writing a given vector, \( \mathbf{F} \) in terms of two vectors, one which is parallel to a given vector, \( \mathbf{D} \) and the other which is perpendicular can also be explained with no reliance on trigonometry, completely in terms of the algebraic properties of the dot product. As before, this is mathematically more significant than any approach involving geometry or trigonometry because it extends to more interesting situations. This is done next.

**Theorem 10.2.8** Let \( \mathbf{F} \) and \( \mathbf{D} \) be nonzero vectors. Then there exist unique vectors \( \mathbf{F}_|| \) and \( \mathbf{F}_\perp \) such that

$$\mathbf{F} = \mathbf{F}_|| + \mathbf{F}_\perp \quad (10.13)$$

where \( \mathbf{F}_|| \) is a scalar multiple of \( \mathbf{D} \), also referred to as

$$\text{proj}_\mathbf{D}(\mathbf{F}),$$

and \( \mathbf{F}_\perp \cdot \mathbf{D} = 0 \). The vector \( \text{proj}_\mathbf{D}(\mathbf{F}) \) is called the **projection** of \( \mathbf{F} \) onto \( \mathbf{D} \).

**Proof:** Suppose \( \mathbf{F}_|| = \alpha \mathbf{D} \). Taking the dot product of both sides with \( \mathbf{D} \) and using \( \mathbf{F}_\perp \cdot \mathbf{D} = 0 \), this yields

$$\mathbf{F} \cdot \mathbf{D} = \alpha |\mathbf{D}|^2$$

which requires \( \alpha = \mathbf{F} \cdot \mathbf{D} / |\mathbf{D}|^2 \). Thus there can be no more than one vector, \( \mathbf{F}_|| \). It follows \( \mathbf{F}_\perp \) must equal \( \mathbf{F} - \mathbf{F}_|| \). This verifies there can be no more than one choice for both \( \mathbf{F}_|| \) and \( \mathbf{F}_\perp \).

Now let

$$\mathbf{F}_|| \equiv \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \mathbf{D}$$

and let

$$\mathbf{F}_\perp = \mathbf{F} - \mathbf{F}_|| = \mathbf{F} - \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \mathbf{D}$$

Then \( \mathbf{F}_|| = \alpha \mathbf{D} \) where \( \alpha = \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \). It only remains to verify \( \mathbf{F}_\perp \cdot \mathbf{D} = 0 \). But

$$\mathbf{F}_\perp \cdot \mathbf{D} = \mathbf{F} \cdot \mathbf{D} - \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2} \mathbf{D} \cdot \mathbf{D}$$

$$= \mathbf{F} \cdot \mathbf{D} - \mathbf{F} \cdot \mathbf{D} = 0.$$  

This proves the theorem.

**Example 10.2.9** Let \( \mathbf{F} = 2\mathbf{i} + 7\mathbf{j} - 3\mathbf{k} \) Newtons. Find the work done by this force in moving from the point \((1, 2, 3)\) to the point \((-9, -3, 4)\) along the straight line segment joining these points where distances are measured in meters.
According to the definition, this work is
\[(2i+7j-3k) \cdot (-10i-5j+k) = -20 + (-35) + (-3) = -58 \text{ Newton meters.}\]

Note that if the force had been given in pounds and the distance had been given in feet, the units on the work would have been foot pounds. In general, work has units equal to units of a force times units of a length. Instead of writing Newton meter, people write joule because a joule is by definition a Newton meter. That word is pronounced “jewel” and it is the unit of work in the metric system of units. Also be sure you observe that the work done by the force can be negative as in the above example. In fact, work can be either positive, negative, or zero. You just have to do the computations to find out.

Example 10.2.10 Find \(\text{proj}_u(v)\) if \(u = 2i + 3j - 4k\) and \(v = i - 2j + k\).

From the above discussion in Theorem 10.2.8, this is just
\[
\frac{1}{4 + 9 + 16} \cdot \frac{8}{29} \cdot (2i + 3j - 4k) = \frac{-16}{29} - \frac{24}{29}j + \frac{32}{29}k.
\]

Example 10.2.11 Suppose \(a\) and \(b\) are vectors and \(b_\perp = b - \text{proj}_a(b)\). What is the magnitude of \(b_\perp\) in terms of the included angle?

\[
|b_\perp|^2 = (b - \text{proj}_a(b)) \cdot (b - \text{proj}_a(b))
\]
\[
= (\frac{b \cdot a}{|a|^2}a) \cdot (\frac{b \cdot a}{|a|^2}a)
\]
\[
= |b|^2 - \frac{(b \cdot a)^2}{|a|^2} + \frac{(b \cdot a)^2}{|a|^2} |a|^2
\]
\[
= |b|^2 \left(1 - \frac{(b \cdot a)^2}{|a|^2 |b|^2}\right)
\]
\[
= |b|^2 \left(1 - \cos^2 \theta\right) = |b|^2 \sin^2 (\theta)
\]

where \(\theta\) is the included angle between \(a\) and \(b\) which is less than \(\pi\) radians. Therefore, taking square roots,

\[
|b_\perp| = |b| \sin \theta.
\]

10.2.3 The Dot Product And Distance In \(\mathbb{C}^n\)

It is necessary to give a generalization of the dot product for vectors in \(\mathbb{C}^n\). This definition reduces to the usual one in the case the components of the vector are real.

Definition 10.2.12 Let \(x, y \in \mathbb{C}^n\). Thus \(x = (x_1, \cdots, x_n)\) where each \(x_k \in \mathbb{C}\) and a similar formula holding for \(y\). Then the dot product of these two vectors is defined to be

\[
x \cdot y = \sum_j x_j \overline{y_j} = x_1 \overline{y_1} + \cdots + x_n \overline{y_n}.
\]
Notice how you put the conjugate on the entries of the vector, $y$. It makes no difference if the vectors happen to be real vectors but with complex vectors you must do it this way. The reason for this is that when you take the dot product of a vector with itself, you want to get the square of the length of the vector, a positive number. Placing the conjugate on the components of $y$ in the above definition assures this will take place. Thus

$$\mathbf{x} \cdot \mathbf{x} = \sum_j x_j \overline{x}_j = \sum_j |x_j|^2 \geq 0.$$ 

If you didn't place a conjugate as in the above definition, things wouldn't work out correctly. For example,

$$(1 + i)^2 + 2^2 = 4 + 2i$$

and this is not a positive number.

The following properties of the dot product follow immediately from the definition and you should verify each of them.

**Properties of the dot product:**

1. $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$.
2. If $a, b$ are numbers and $\mathbf{u}, \mathbf{v}, \mathbf{z}$ are vectors then $(a \mathbf{u} + b \mathbf{v}) \cdot \mathbf{z} = a (\mathbf{u} \cdot \mathbf{z}) + b (\mathbf{v} \cdot \mathbf{z})$.
3. $\mathbf{u} \cdot \mathbf{u} \geq 0$ and it equals 0 if and only if $\mathbf{u} = \mathbf{0}$.

Note this implies $(\mathbf{x} \cdot \alpha \mathbf{y}) = \overline{\alpha} (\mathbf{x} \cdot \mathbf{y})$ because

$$(\mathbf{x} \cdot \alpha \mathbf{y}) = (\alpha \mathbf{y} \cdot \mathbf{x}) = \overline{\alpha} (\mathbf{y} \cdot \mathbf{x}) = \overline{\alpha} (\mathbf{x} \cdot \mathbf{y})$$

The norm is defined in the usual way.

**Definition 10.2.13** For $\mathbf{x} \in \mathbb{C}^n$,

$$|\mathbf{x}| \equiv \left( \sum_{k=1}^{n} |x_k|^2 \right)^{1/2} = (\mathbf{x} \cdot \mathbf{x})^{1/2}$$

Here is a fundamental inequality called the **Cauchy Schwarz inequality** which is stated here in $\mathbb{C}^n$. First here is a simple lemma.

**Lemma 10.2.14** If $z \in \mathbb{C}$ there exists $\theta \in \mathbb{C}$ such that $\theta z = |z|$ and $|\theta| = 1$.

**Proof:** Let $\theta = 1$ if $z = 0$ and otherwise, let $\theta = \frac{z}{|z|}$. Recall that for $z = x + iy, \overline{z} = x - iy$ and $\overline{z} z = |z|^2$.

I will give a proof of this important inequality which depends only on the above list of properties of the dot product. It will be slightly different than the earlier proof.

**Theorem 10.2.15** (Cauchy Schwarz) The following inequality holds for $\mathbf{x}$ and $\mathbf{y} \in \mathbb{C}^n$.

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq (\mathbf{x} \cdot \mathbf{x})^{1/2} (\mathbf{y} \cdot \mathbf{y})^{1/2} \quad (10.14)$$

Equality holds in this inequality if and only if one vector is a multiple of the other.

**Proof:** Let $\theta \in \mathbb{C}$ such that $|\theta| = 1$ and

$$\theta (\mathbf{x} \cdot \mathbf{y}) = |\langle \mathbf{x}, \mathbf{y} \rangle|$$
Consider $p(t) \equiv (x + \bar{t}y, x + t\overline{y})$ where $t \in \mathbb{R}$. Then from the above list of properties of the dot product,

$$
0 \leq p(t) = (x \cdot x) + t\theta(x \cdot y) + t\overline{y}(y \cdot x) + t^2(y \cdot y)
$$

and this must hold for all $t \in \mathbb{R}$. Therefore, if $(y \cdot y) = 0$ it must be the case that $|(x \cdot y)| = 0$ also since otherwise the above inequality would be violated. Therefore, in this case,

$$
|(x \cdot y)| \leq (x \cdot x)^{1/2}(y \cdot y)^{1/2}.
$$

On the other hand, if $(y \cdot y) \neq 0$, then $p(t) \geq 0$ for all $t$ means the graph of $y = p(t)$ is a parabola which opens up and it either has exactly one real zero in the case its vertex touches the $t$ axis or it has no real zeros. From the quadratic formula this happens exactly when

$$
4|(x \cdot y)|^2 - 4(x \cdot x)(y \cdot y) \leq 0
$$

which is equivalent to \[10.14]\]

It is clear from a computation that if one vector is a scalar multiple of the other that equality holds in \[10.14]. Conversely, suppose equality does hold. Then this is equivalent to saying $4|(x \cdot y)|^2 - 4(x \cdot x)(y \cdot y) = 0$ and so from the quadratic formula, there exists one real zero to $p(t) = 0$. Call it $t_0$. Then

$$
p(t_0) \equiv (x + \bar{t}_0 y, x + t_0\overline{y}) = |x + \bar{t}_0 y|^2 = 0
$$

and so $x = -\bar{t}_0 y$. This proves the theorem.

Note that I only used part of the above properties of the dot product. It was not necessary to use the one which says that if $(x \cdot x) = 0$ then $x = 0$.

By analogy to the case of $\mathbb{R}^n$, length or magnitude of vectors in $\mathbb{C}^n$ can be defined.

**Definition 10.2.16** Let $z \in \mathbb{C}^n$. Then $|z| \equiv (z \cdot z)^{1/2}$.

**Theorem 10.2.17** For length defined in Definition 10.2.16, the following hold.

$$
|z| \geq 0 \text{ and } |z| = 0 \text{ if and only if } z = 0
$$

(10.16)

If $\alpha$ is a scalar, $|\alpha z| = |\alpha||z|$

(10.17)

$$
|z + w| \leq |z| + |w|.
$$

(10.18)

**Proof:** The first two claims are left as exercises. To establish the third, you use the same argument which was used in $\mathbb{R}^n$.

$$
|z + w|^2 = (z + w, z + w)
$$

$$
= z \cdot z + w \cdot w + w \cdot z + z \cdot w
$$

$$
= |z|^2 + |w|^2 + 2\Re w \cdot z
$$

$$
\leq |z|^2 + |w|^2 + 2|w||z| = (|z| + |w|)^2.
$$
10.3 Exercises

1. Use formula [10.11] to verify the Cauchy Schwartz inequality and to show that equality occurs if and only if one of the vectors is a scalar multiple of the other.

2. For \( \mathbf{u}, \mathbf{v} \) vectors in \( \mathbb{R}^3 \), define the product, \( \mathbf{u} \ast \mathbf{v} \equiv u_1v_1 + 2u_2v_2 + 3u_3v_3 \). Show the axioms for a dot product all hold for this funny product. Prove

\[
|\mathbf{u} \ast \mathbf{v}| \leq (\mathbf{u} \ast \mathbf{u})^{1/2} (\mathbf{v} \ast \mathbf{v})^{1/2}.
\]

Hint: Do not try to do this with methods from trigonometry.

3. Find the angle between the vectors \( 3\mathbf{i} - \mathbf{j} - \mathbf{k} \) and \( \mathbf{i} + 4\mathbf{j} + 2\mathbf{k} \).

4. Find the angle between the vectors \( \mathbf{i} - 2\mathbf{j} + \mathbf{k} \) and \( \mathbf{i} + 2\mathbf{j} - 7\mathbf{k} \).

5. Find \( \text{proj}_u (\mathbf{v}) \) where \( \mathbf{v} = (1, 0, -2) \) and \( \mathbf{u} = (1, 2, 3) \).

6. Find \( \text{proj}_u (\mathbf{v}) \) where \( \mathbf{v} = (1, 2, -2) \) and \( \mathbf{u} = (1, 0, 3) \).

7. Find \( \text{proj}_u (\mathbf{v}) \) where \( \mathbf{v} = (1, 2, -2, 1) \) and \( \mathbf{u} = (1, 2, 3, 0) \).

8. Does it make sense to speak of \( \text{proj}_0 (\mathbf{v}) \)?

9. If \( \mathbf{F} \) is a force and \( \mathbf{D} \) is a vector, show \( \text{proj}_D (\mathbf{F}) = (|\mathbf{F}| \cos \theta) \mathbf{u} \) where \( \mathbf{u} \) is the unit vector in the direction of \( \mathbf{D} \), \( \mathbf{u} = \mathbf{D}/|\mathbf{D}| \) and \( \theta \) is the included angle between the two vectors, \( \mathbf{F} \) and \( \mathbf{D} \). \( |\mathbf{F}| \cos \theta \) is sometimes called the component of the force, \( \mathbf{F} \) in the direction, \( \mathbf{D} \).

10. A boy drags a sled for 100 feet along the ground by pulling on a rope which is 20 degrees from the horizontal with a force of 40 pounds. How much work does this force do?

11. A girl drags a sled for 200 feet along the ground by pulling on a rope which is 30 degrees from the horizontal with a force of 20 pounds. How much work does this force do?

12. A large dog drags a sled for 300 feet along the ground by pulling on a rope which is 45 degrees from the horizontal with a force of 20 pounds. How much work does this force do?

13. How much work in Newton meters does it take to slide a crate 20 meters along a loading dock by pulling on it with a 200 Newton force at an angle of 30° from the horizontal?

14. An object moves 10 meters in the direction of \( \mathbf{j} \). There are two forces acting on this object, \( \mathbf{F}_1 = \mathbf{i} + \mathbf{j} + 2\mathbf{k} \), and \( \mathbf{F}_2 = -5\mathbf{i} + 2\mathbf{j} - 6\mathbf{k} \). Find the total work done on the object by the two forces. Hint: You can take the work done by the resultant of the two forces or you can add the work done by each force. Why?

15. An object moves 10 meters in the direction of \( \mathbf{j} + \mathbf{i} \). There are two forces acting on this object, \( \mathbf{F}_1 = \mathbf{i} + 2\mathbf{j} + 2\mathbf{k} \), and \( \mathbf{F}_2 = 5\mathbf{i} + 2\mathbf{j} - 6\mathbf{k} \). Find the total work done on the object by the two forces. Hint: You can take the work done by the resultant of the two forces or you can add the work done by each force. Why?

16. An object moves 20 meters in the direction of \( \mathbf{k} + \mathbf{j} \). There are two forces acting on this object, \( \mathbf{F}_1 = \mathbf{i} + \mathbf{j} + 2\mathbf{k} \), and \( \mathbf{F}_2 = \mathbf{i} + 2\mathbf{j} - 6\mathbf{k} \). Find the total work done on the object by the two forces. Hint: You can take the work done by the resultant of the two forces or you can add the work done by each force.
17. If \( \mathbf{a} \), \( \mathbf{b} \), and \( \mathbf{c} \) are vectors. Show that \( (\mathbf{b} + \mathbf{c})_\perp = \mathbf{b}_\perp + \mathbf{c}_\perp \) where \( \mathbf{b}_\perp = \mathbf{b} - \text{proj}_\mathbf{a}(\mathbf{b}) \).

18. Find \((1, 2, 3, 4) \cdot (2, 0, 1, 3)\).

19. Show that \((\mathbf{a} \cdot \mathbf{b}) = \frac{1}{4} \left[ |\mathbf{a} + \mathbf{b}|^2 - |\mathbf{a} - \mathbf{b}|^2 \right] \).

20. Prove from the axioms of the dot product the parallelogram identity, \(|\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2\).

### 10.4 Exercises With Answers

1. Find the angle between the vectors \(3\mathbf{i} - \mathbf{j} - \mathbf{k}\) and \(\mathbf{i} + 4\mathbf{j} + 2\mathbf{k}\).

\[
\cos \theta = \frac{3\mathbf{i} - \mathbf{j} - \mathbf{k} \cdot \mathbf{i} + 4\mathbf{j} + 2\mathbf{k}}{\sqrt{3^2 + (-1)^2 + (-1)^2} \cdot \sqrt{1^2 + 4^2 + 2^2}} = -0.19739.
\]

Therefore, you have to solve the equation \(\cos \theta = -0.19739\). Solution is: \(\theta = 1.7695\) radians. You need to use a calculator or table to solve this.

2. Find \(\text{proj}_\mathbf{u}(\mathbf{v})\) where \(\mathbf{v} = (1, 3, -2)\) and \(\mathbf{u} = (1, 2, 3)\).

Remember to find this you take \(\frac{\mathbf{v} \cdot \mathbf{u}}{||\mathbf{u}||^2} \mathbf{u}\). Thus the answer is \(\frac{1}{11}(1, 2, 3)\).

3. If \(\mathbf{F}\) is a force and \(\mathbf{D}\) is a vector, show \(\text{proj}_\mathbf{D}(\mathbf{F}) = (||\mathbf{F}|\cos \theta|) \mathbf{u}\) where \(\mathbf{u}\) is the unit vector in the direction of \(\mathbf{D}\), \(\mathbf{u} = \mathbf{D}/||\mathbf{D}||\) and \(\theta\) is the included angle between the two vectors, \(\mathbf{F}\) and \(\mathbf{D}\). \(\mathbf{F}|\cos \theta\) is sometimes called the component of the force, \(\mathbf{F}\) in the direction, \(\mathbf{D}\).

\[
\text{proj}_\mathbf{D}(\mathbf{F}) = \frac{\mathbf{F} \cdot \mathbf{D}}{||\mathbf{D}||} \mathbf{D} = ||\mathbf{F}|\cos \theta| \frac{\mathbf{D}}{||\mathbf{D}||}.
\]

4. A boy drags a sled for 100 feet along the ground by pulling on a rope which is 40 degrees from the horizontal with a force of 10 pounds. How much work does this force do?

The component of force is \(10 \cos \left(\frac{40}{180}\pi\right)\) and it acts for 100 feet so the work done is

\[
10 \cos \left(\frac{40}{180}\pi\right) \times 100 = 766.04
\]

5. If \(\mathbf{a}\), \(\mathbf{b}\), and \(\mathbf{c}\) are vectors. Show that \((\mathbf{b} + \mathbf{c})_\perp = \mathbf{b}_\perp + \mathbf{c}_\perp\) where \(\mathbf{b}_\perp = \mathbf{b} - \text{proj}_\mathbf{a}(\mathbf{b})\).

6. Find \((1, 0, 3, 4) \cdot (2, 7, 1, 3)\). \((1, 0, 3, 4) \cdot (2, 7, 1, 3) = 17\).

7. Show that \((\mathbf{a} \cdot \mathbf{b}) = \frac{1}{4} \left[ |\mathbf{a} + \mathbf{b}|^2 - |\mathbf{a} - \mathbf{b}|^2 \right] \).

This follows from the axioms of the dot product and the definition of the norm. Thus

\[
|\mathbf{a} + \mathbf{b}|^2 = (\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} + \mathbf{b}) = |\mathbf{a}|^2 + |\mathbf{b}|^2 + 2(\mathbf{a} \cdot \mathbf{b})
\]

Do something similar for \(|\mathbf{a} - \mathbf{b}|^2\).

8. Prove from the axioms of the dot product the parallelogram identity, \(|\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2\).

Start with the left side and write the definitions of \(|\mathbf{a} + \mathbf{b}|^2\), \(|\mathbf{a} - \mathbf{b}|^2\) in terms of the dot product and then simplify.
10.5 The Cross Product

The cross product is the other way of multiplying two vectors in $\mathbb{R}^3$. It is very different from the dot product in many ways. First the geometric meaning is discussed and then a description in terms of coordinates is given. Both descriptions of the cross product are important. The geometric description is essential in order to understand the applications to physics and geometry while the coordinate description is the only way to practically compute the cross product.

**Definition 10.5.1** Three vectors, $\mathbf{a}, \mathbf{b}, \mathbf{c}$ form a right handed system if when you extend the fingers of your right hand along the vector, $\mathbf{a}$ and close them in the direction of $\mathbf{b}$, the thumb points roughly in the direction of $\mathbf{c}$.

For an example of a right handed system of vectors, see the following picture.

In this picture the vector $\mathbf{c}$ points upwards from the plane determined by the other two vectors. You should consider how a right hand system would differ from a left hand system. Try using your left hand and you will see that the vector, $\mathbf{c}$ would need to point in the opposite direction as it would for a right hand system.

From now on, the vectors, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ will always form a right handed system. To repeat, if you extend the fingers of your right hand along $\mathbf{i}$ and close them in the direction $\mathbf{j}$, the thumb points in the direction of $\mathbf{k}$.

The following is the geometric description of the cross product. It gives both the direction and the magnitude and therefore specifies the vector.

**Definition 10.5.2** Let $\mathbf{a}$ and $\mathbf{b}$ be two vectors in $\mathbb{R}^3$. Then $\mathbf{a} \times \mathbf{b}$ is defined by the following two rules.

1. $|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}| |\mathbf{b}| \sin \theta$ where $\theta$ is the included angle.

2. $\mathbf{a} \times \mathbf{b} \cdot \mathbf{a} = 0$, $\mathbf{a} \times \mathbf{b} \cdot \mathbf{b} = 0$, and $\mathbf{a}, \mathbf{b}, \mathbf{a} \times \mathbf{b}$ forms a right hand system.

Note that $|\mathbf{a} \times \mathbf{b}|$ is the area of the parallelogram determined by $\mathbf{a}$ and $\mathbf{b}$. 
The cross product satisfies the following properties.

\[ \mathbf{a} \times \mathbf{b} = - (\mathbf{b} \times \mathbf{a}) \], \quad \mathbf{a} \times \mathbf{a} = \mathbf{0}, \quad (10.19) \]

For \( \alpha \) a scalar,

\[ (\alpha \mathbf{a}) \times \mathbf{b} = \alpha (\mathbf{a} \times \mathbf{b}) = \alpha \times (\mathbf{a} \times \mathbf{b}), \quad (10.20) \]

For \( \mathbf{a}, \mathbf{b}, \) and \( \mathbf{c} \) vectors, one obtains the distributive laws,

\[ \mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}, \quad (10.21) \]

\[ (\mathbf{b} + \mathbf{c}) \times \mathbf{a} = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}. \quad (10.22) \]

Formula (10.19) follows immediately from the definition. The vectors \( \mathbf{a} \times \mathbf{b} \) and \( \mathbf{b} \times \mathbf{a} \) have the same magnitude, \( |\mathbf{a}| |\mathbf{b}| \sin \theta \), and an application of the right hand rule shows they have opposite direction. Formula (10.20) is also fairly clear. If \( \alpha \) is a nonnegative scalar, the direction of \( (\alpha \mathbf{a}) \times \mathbf{b} \) is the same as the direction of \( \mathbf{a} \times \mathbf{b} \), \( \mathbf{a} \times (\alpha \mathbf{b}) \) while the magnitude is just \( \alpha \) times the magnitude of \( \mathbf{a} \times \mathbf{b} \) which is the same as the magnitude of \( \alpha (\mathbf{a} \times \mathbf{b}) \) and \( \mathbf{a} \times (\alpha \mathbf{b}) \). Using this yields equality in (10.20). In the case where \( \alpha < 0 \), everything works the same way except the vectors are all pointing in the opposite direction and you must multiply by \( |\alpha| \) when comparing their magnitudes.

The distributive laws are much harder to establish but the second follows from the first quite easily. Thus, assuming the first, and using (10.19),

\[ (\mathbf{b} + \mathbf{c}) \times \mathbf{a} = - \mathbf{a} \times (\mathbf{b} + \mathbf{c}) \]
\[ = - (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}) \]
\[ = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}. \]

A proof of the distributive law is given in a later section for those who are interested.

Now from the definition of the cross product,

\[ \mathbf{i} \times \mathbf{j} = \mathbf{k} \quad \mathbf{j} \times \mathbf{i} = - \mathbf{k} \]
\[ \mathbf{k} \times \mathbf{i} = \mathbf{j} \quad \mathbf{i} \times \mathbf{k} = - \mathbf{j} \]
\[ \mathbf{j} \times \mathbf{k} = \mathbf{i} \quad \mathbf{k} \times \mathbf{j} = - \mathbf{i} \]

With this information, the following gives the coordinate description of the cross product.

**Proposition 10.5.3** Let \( \mathbf{a} = a_1 \mathbf{i} + a_2 \mathbf{j} + a_3 \mathbf{k} \) and \( \mathbf{b} = b_1 \mathbf{i} + b_2 \mathbf{j} + b_3 \mathbf{k} \) be two vectors. Then

\[ \mathbf{a} \times \mathbf{b} = (a_2 b_3 - a_3 b_2) \mathbf{i} + (a_3 b_1 - a_1 b_3) \mathbf{j} + (a_1 b_2 - a_2 b_1) \mathbf{k}. \quad (10.23) \]
**Proof:** From the above table and the properties of the cross product listed,

\[
(a_1i + a_2j + a_3k) \times (b_1i + b_2j + b_3k) = \\
= a_1b_2i \times j + a_3b_1i \times k + a_2b_3j \times k + a_3b_1i \times j + a_3b_2k \times j \\
= a_1b_2k - a_1b_3j - a_2b_1k + a_2b_3i + a_3b_1j - a_3b_2i \\
= (a_2b_3 - a_3b_2)i + (a_3b_1 - a_1b_3)j + (a_1b_2 - a_2b_1)k \\
\text{(10.24)}
\]

This proves the proposition.

It is probably impossible for most people to remember \[10.23\]. Fortunately, there is a somewhat easier way to remember it. Define the determinant of a $2 \times 2$ matrix as follows

\[
\begin{vmatrix}
  a & b \\
  c & d
\end{vmatrix} \equiv ad - bc
\]

Then

\[
a \times b = \begin{vmatrix}
  i & j & k \\
  a_1 & a_2 & a_3 \\
  b_1 & b_2 & b_3
\end{vmatrix}
\]

where you expand the determinant along the top row. This yields

\[
i (-1)^{1+1} \begin{vmatrix}
  a_2 & a_3 \\
  b_2 & b_3
\end{vmatrix} + j (-1)^{2+1} \begin{vmatrix}
  a_1 & a_3 \\
  b_1 & b_3
\end{vmatrix} + k (-1)^{3+1} \begin{vmatrix}
  a_1 & a_2 \\
  b_1 & b_2
\end{vmatrix}
\]

\[
= i \begin{vmatrix}
  a_2 & a_3 \\
  b_2 & b_3
\end{vmatrix} - j \begin{vmatrix}
  a_1 & a_3 \\
  b_1 & b_3
\end{vmatrix} + k \begin{vmatrix}
  a_1 & a_2 \\
  b_1 & b_2
\end{vmatrix}
\]

Note that to get the scalar which multiplies $i$ you take the determinant of what is left after deleting the first row and the first column and multiply by $(-1)^{1+1}$ because $i$ is in the first row and the first column. Then you do the same thing for the $j$ and $k$. In the case of the $j$ there is a minus sign because $j$ is in the first row and the second column and so $(-1)^{1+2} = -1$ while the $k$ is multiplied by $(-1)^{3+1} = 1$. The above equals

\[
(a_2b_3 - a_3b_2)i - (a_1b_3 - a_3b_1)j + (a_1b_2 - a_2b_1)k \\
\text{(10.26)}
\]

which is the same as \[10.24\]. There will be much more presented on determinants later. For now, consider this an introduction if you have not seen this topic.

**Example 10.5.4** Find $(i - j + 2k) \times (3i - 2j + k)$.

Use \[10.25\] to compute this.

\[
\begin{vmatrix}
  i & j & k \\
  1 & -1 & 2 \\
  3 & -2 & 1
\end{vmatrix} = \begin{vmatrix}
  -1 & 2 \\
  -2 & 1 \\
  1 & 2
\end{vmatrix} i - \begin{vmatrix}
  1 & 3 \\
  -2 & 1 \\
  1 & 3
\end{vmatrix} j + \begin{vmatrix}
  1 & -1 \\
  -2 & 1 \\
  1 & -2
\end{vmatrix} k
\]

\[
= 3i + 5j + k.
\]

**Example 10.5.5** Find the area of the parallelogram determined by the vectors, $(i - j + 2k)$ and $(3i - 2j + k)$.

These are the same two vectors in Example \[10.5.4\].
10.5. THE CROSS PRODUCT

From Example 10.5.4 and the geometric description of the cross product, the area is just the norm of the vector obtained in Example 10.5.4. Thus the area is \( \sqrt{9 + 25 + 1} = \sqrt{35} \).

**Example 10.5.6** Find the area of the triangle determined by \((1, 2, 3), (0, 2, 5), \) and \((5, 1, 2)\).

This triangle is obtained by connecting the three points with lines. Picking \((1, 2, 3)\) as a starting point, there are two displacement vectors, \((-1, 0, 2)\) and \((4, -1, -1)\) such that the given vector added to these displacement vectors gives the other two vectors.

The area of the triangle is half the area of the parallelogram determined by \((-1, 0, 2)\) and \((4, -1, -1)\). Thus \((-1, 0, 2) \times (4, -1, -1) = (2, 7, 1)\) and so the area of the triangle is \(\frac{1}{2} \sqrt{4 + 49 + 1} = \frac{1}{2} \sqrt{6} \).

**Observation 10.5.7** In general, if you have three points (vectors) in \(\mathbb{R}^3\), \(P, Q, R\) the area of the triangle is given by

\[
\frac{1}{2} |(Q - P) \times (R - P)|.
\]

![Diagram of triangle]

10.5.1 The Distributive Law For The Cross Product

This section gives a proof for 10.21, a fairly difficult topic. It is included here for the interested student. If you are satisfied with taking the distributive law on faith, it is not necessary to read this section. The proof given here is quite clever and follows the one given in [6]. Another approach, based on volumes of parallelepipeds is found in [15] and is discussed a little later.

**Lemma 10.5.8** Let \(b\) and \(c\) be two vectors. Then \(b \times c = b \times c_\perp\) where \(c_\parallel + c_\perp = c\) and \(c_\perp \cdot b = 0\).

**Proof:** Consider the following picture.

![Diagram of vectors]

Now \(c_\perp = c - c_\parallel = c - \frac{b}{|b|} b\) and so \(c_\perp\) is in the plane determined by \(c\) and \(b\). Therefore, from the geometric definition of the cross product, \(b \times c\) and \(b \times c_\perp\) have the same direction. Now, referring to the picture,

\[
|b \times c_\perp| = |b| |c_\perp| = |b| |c| \sin \theta = |b \times c|.
\]

Therefore, \(b \times c\) and \(b \times c_\perp\) also have the same magnitude and so they are the same vector.

With this, the proof of the distributive law is in the following theorem.
Theorem 10.5.9 Let \( \mathbf{a}, \mathbf{b}, \) and \( \mathbf{c} \) be vectors in \( \mathbb{R}^3 \). Then
\[
\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}
\] (10.27)

**Proof:** Suppose first that \( \mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{c} = 0 \). Now imagine \( \mathbf{a} \) is a vector coming out of the page and let \( \mathbf{b}, \mathbf{c} \) and \( \mathbf{b} + \mathbf{c} \) be as shown in the following picture.

Then \( \mathbf{a} \times \mathbf{b}, \mathbf{a} \times (\mathbf{b} + \mathbf{c}), \) and \( \mathbf{a} \times \mathbf{c} \) are each vectors in the same plane, perpendicular to \( \mathbf{a} \) as shown. Thus \( \mathbf{a} \times \mathbf{c} \cdot \mathbf{c} = 0 \), \( \mathbf{a} \times (\mathbf{b} + \mathbf{c}) \cdot (\mathbf{b} + \mathbf{c}) = 0 \), and \( \mathbf{a} \times \mathbf{b} \cdot \mathbf{b} = 0 \). This implies that to get \( \mathbf{a} \times \mathbf{b} \) you move counterclockwise through an angle of \( \pi/2 \) radians from the vector, \( \mathbf{b} \). Similar relationships exist between the vectors \( \mathbf{a} \times (\mathbf{b} + \mathbf{c}) \) and \( \mathbf{b} + \mathbf{c} \) and the vectors \( \mathbf{a} \times \mathbf{c} \) and \( \mathbf{c} \). Thus the angle between \( \mathbf{a} \times \mathbf{b} \) and \( \mathbf{a} \times (\mathbf{b} + \mathbf{c}) \) is the same as the angle between \( \mathbf{b} + \mathbf{c} \) and \( \mathbf{b} \) and the angle between \( \mathbf{a} \times \mathbf{c} \) and \( \mathbf{a} \times (\mathbf{b} + \mathbf{c}) \) is the same as the angle between \( \mathbf{c} \) and \( \mathbf{b} + \mathbf{c} \). In addition to this, since \( \mathbf{a} \) is perpendicular to these vectors,
\[
|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}| \cdot |\mathbf{b}|, |\mathbf{a} \times (\mathbf{b} + \mathbf{c})| = |\mathbf{a}| \cdot |\mathbf{b} + \mathbf{c}|, \text{ and}
|\mathbf{a} \times \mathbf{c}| = |\mathbf{a}| \cdot |\mathbf{c}|.
\]

Therefore,
\[
\frac{|\mathbf{a} \times (\mathbf{b} + \mathbf{c})|}{|\mathbf{b} + \mathbf{c}|} = \frac{|\mathbf{a} \times \mathbf{c}|}{|\mathbf{c}|} = \frac{|\mathbf{a} \times \mathbf{b}|}{|\mathbf{b}|} = \frac{|\mathbf{a}|}{|\mathbf{b}|}
\]
and so
\[
\frac{|\mathbf{a} \times (\mathbf{b} + \mathbf{c})|}{|\mathbf{a} \times \mathbf{c}|} = \frac{|\mathbf{b} + \mathbf{c}|}{|\mathbf{c}|} = \frac{|\mathbf{a} \times (\mathbf{b} + \mathbf{c})|}{|\mathbf{a} \times \mathbf{b}|} = \frac{|\mathbf{b} + \mathbf{c}|}{|\mathbf{b}|}
\]
showing the triangles making up the parallelogram on the right and the four sided figure on the left in the above picture are similar. It follows the four sided figure on the left is in fact a parallelogram and this implies the diagonal is the vector sum of the vectors on the sides, yielding \( 10.27 \).

Now suppose it is not necessarily the case that \( \mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{c} = 0 \). Then write \( \mathbf{b} = \mathbf{b}_{||} + \mathbf{b}_{\perp} \) where \( \mathbf{b}_{\perp} \cdot \mathbf{a} = 0 \). Similarly \( \mathbf{c} = \mathbf{c}_{||} + \mathbf{c}_{\perp} \). By the above lemma and what was just shown,
\[
\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times (\mathbf{b}_{\perp} + \mathbf{c}_{\perp})
\]
\[
= \mathbf{a} \times (\mathbf{b}_{\perp} + \mathbf{c}_{\perp})
\]
\[
= \mathbf{a} \times \mathbf{b}_{\perp} + \mathbf{a} \times \mathbf{c}_{\perp}
\]
\[
= \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}.
\]

This proves the theorem.

The result of Problem 17 of the exercises 10.3 is used to go from the first to the second line.
10.5.2 The Box Product

**Definition 10.5.10** A parallelepiped determined by the three vectors, \( \mathbf{a}, \mathbf{b}, \) and \( \mathbf{c} \) consists of
\[
\{ r\mathbf{a} + s\mathbf{b} + t\mathbf{c} : r, s, t \in [0, 1] \}.
\]
That is, if you pick three numbers, \( r, s, \) and \( t \) each in \([0, 1]\) and form \( r\mathbf{a} + s\mathbf{b} + t\mathbf{c} \), then the collection of all such points is what is meant by the parallelepiped determined by these three vectors.

The following is a picture of such a thing.

![Diagram of a parallelepiped](image)

You notice the area of the base of the parallelepiped, the parallelogram determined by the vectors, \( \mathbf{a} \) and \( \mathbf{b} \) has area equal to \( |\mathbf{a} \times \mathbf{b}| \) while the altitude of the parallelepiped is \( |\mathbf{c}| \cos \theta \) where \( \theta \) is the angle shown in the picture between \( \mathbf{c} \) and \( \mathbf{a} \times \mathbf{b} \). Therefore, the volume of this parallelepiped is the area of the base times the altitude which is just
\[
|\mathbf{a} \times \mathbf{b}| |\mathbf{c}| \cos \theta = \mathbf{a} \times \mathbf{b} \cdot \mathbf{c}.
\]

This expression is known as the box product and is sometimes written as \([\mathbf{a}, \mathbf{b}, \mathbf{c}]\). You should consider what happens if you interchange the \( \mathbf{b} \) with the \( \mathbf{c} \) or the \( \mathbf{a} \) with the \( \mathbf{c} \). You can see geometrically from drawing pictures that this merely introduces a minus sign. In any case the box product of three vectors always equals either the volume of the parallelepiped determined by the three vectors or else minus this volume.

**Example 10.5.11** Find the volume of the parallelepiped determined by the vectors, \( \mathbf{i} + 2\mathbf{j} - 5\mathbf{k}, \mathbf{i} + 3\mathbf{j} - 6\mathbf{k}, 3\mathbf{i} + 2\mathbf{j} + 3\mathbf{k} \).

According to the above discussion, pick any two of these, take the cross product and then take the dot product of this with the third of these vectors. The result will be either the desired volume or minus the desired volume.

\[
(i + 2j - 5k) \times (i + 3j - 6k) = \begin{vmatrix}
\mathbf{i} & \mathbf{j} & \mathbf{k} \\
1 & 2 & -5 \\
1 & 3 & -6
\end{vmatrix} = 3\mathbf{i} + \mathbf{j} + \mathbf{k}
\]

Now take the dot product of this vector with the third which yields
\[
(3\mathbf{i} + \mathbf{j} + \mathbf{k}) \cdot (3\mathbf{i} + 2\mathbf{j} + 3\mathbf{k}) = 9 + 2 + 3 = 14.
\]

This shows the volume of this parallelepiped is 14 cubic units.

There is a fundamental observation which comes directly from the geometric definitions of the cross product and the dot product.
Lemma 10.5.12 Let \( \mathbf{a}, \mathbf{b}, \) and \( \mathbf{c} \) be vectors. Then \( (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) \).

Proof: This follows from observing that either \((\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}\) and \(\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})\) both give the volume of the parallelepiped or they both give \(-1\) times the volume.

10.5.3 A Proof Of The Distributive Law

Here is another proof of the distributive law for the cross product. Let \( \mathbf{x} \) be a vector.

From the above observation,

\[
\mathbf{x} \cdot (\mathbf{a} \times (\mathbf{b} + \mathbf{c})) = (\mathbf{x} \times \mathbf{a}) \cdot (\mathbf{b} + \mathbf{c})
\]

\[
= (\mathbf{x} \times \mathbf{a}) \cdot \mathbf{b} + (\mathbf{x} \times \mathbf{a}) \cdot \mathbf{c}
\]

\[
= \mathbf{x} \cdot (\mathbf{a} \times (\mathbf{b} + \mathbf{a} \times \mathbf{c})).
\]

Therefore,

\[
\mathbf{x} \cdot [\mathbf{a} \times (\mathbf{b} + \mathbf{c}) - (\mathbf{a} \times (\mathbf{b} + \mathbf{a} \times \mathbf{c}))] = 0
\]

for all \( \mathbf{x} \). In particular, this holds for \( \mathbf{x} = \mathbf{a} \times (\mathbf{b} + \mathbf{c}) - (\mathbf{a} \times (\mathbf{b} + \mathbf{a} \times \mathbf{c})) \) showing that \( \mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c} \) and this proves the distributive law for the cross product another way.

10.6 Exercises

1. Show that if \( \mathbf{a} \times \mathbf{u} = 0 \) for all unit vectors, \( \mathbf{u} \), then \( \mathbf{a} = 0 \).

2. Find the area of the triangle determined by the three points, \((1, 2, 3), (4, 2, 0)\) and \((-3, 2, 1)\).

3. Find the area of the triangle determined by the three points, \((1, 0, 3), (4, 1, 0)\) and \((-3, 1, 1)\).

4. Find the area of the triangle determined by the three points, \((1, 2, 3), (2, 3, 1)\) and \((0, 1, 2)\). Did something interesting happen here? What does it mean geometrically?

5. Find the area of the parallelogram determined by the vectors, \((1, 2, 3)\) and \((3, -2, 1)\).

6. Find the area of the parallelogram determined by the vectors, \((1, 0, 3)\) and \((4, -2, 1)\).

7. Find the area of the parallelogram determined by the vectors, \((1, -2, 2)\) and \((3, 1, 1)\).

8. Find the volume of the parallelepiped determined by the vectors, \(\mathbf{i} - 7\mathbf{j} - 5\mathbf{k}, \mathbf{i} - 2\mathbf{j} - 6\mathbf{k}, 3\mathbf{i} + 2\mathbf{j} + 3\mathbf{k}\).

9. Find the volume of the parallelepiped determined by the vectors, \(\mathbf{i} + \mathbf{j} - 5\mathbf{k}, \mathbf{i} + 5\mathbf{j} - 6\mathbf{k}, 3\mathbf{i} + \mathbf{j} + 3\mathbf{k}\).

10. Find the volume of the parallelepiped determined by the vectors, \(\mathbf{i} + 6\mathbf{j} + 5\mathbf{k}, \mathbf{i} + 5\mathbf{j} - 6\mathbf{k}, 3\mathbf{i} + \mathbf{j} + \mathbf{k}\).

11. Suppose \( \mathbf{a}, \mathbf{b}, \) and \( \mathbf{c} \) are three vectors whose components are all integers. Can you conclude the volume of the parallelepiped determined from these three vectors will always be an integer?

12. What does it mean geometrically if the box product of three vectors gives zero?
13. Using Problem 12, find an equation of a plane containing the two position vectors, \( \mathbf{a} \) and \( \mathbf{b} \) and the point \( \mathbf{0} \). **Hint:** If \((x, y, z)\) is a point on this plane the volume of the parallelepiped determined by \((x, y, z)\) and the vectors \( \mathbf{a}, \mathbf{b} \) equals 0.

14. Using the notion of the box product yielding either plus or minus the volume of the parallelepiped determined by the given three vectors, show that

\[
(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})
\]

In other words, the dot and the cross can be switched as long as the order of the vectors remains the same. **Hint:** There are two ways to do this, by the coordinate description of the dot and cross product and by geometric reasoning. It is better if you use geometric reasoning.

15. Is \( \mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \times \mathbf{c} \)? What is the meaning of \( \mathbf{a} \times \mathbf{b} \times \mathbf{c} \)? Explain. **Hint:** Try \((\mathbf{i} \times \mathbf{j}) \times \mathbf{j}\) and \(\mathbf{i} \times (\mathbf{j} \times \mathbf{j})\).

16. Verify directly that the coordinate description of the cross product, \( \mathbf{a} \times \mathbf{b} \) has the property that it is perpendicular to both \( \mathbf{a} \) and \( \mathbf{b} \). Then show by direct computation that this coordinate description satisfies

\[
|\mathbf{a} \times \mathbf{b}|^2 = |\mathbf{a}|^2 |\mathbf{b}|^2 - (\mathbf{a} \cdot \mathbf{b})^2
= |\mathbf{a}|^2 |\mathbf{b}|^2 (1 - \cos^2(\theta))
\]

where \( \theta \) is the angle included between the two vectors. Explain why \( |\mathbf{a} \times \mathbf{b}| \) has the correct magnitude. All that is missing is the material about the right hand rule. Verify directly from the coordinate description of the cross product that the right thing happens with regards to the vectors \( \mathbf{i}, \mathbf{j}, \mathbf{k} \). Next verify that the distributive law holds for the coordinate description of the cross product. This gives another way to approach the cross product. First define it in terms of coordinates and then get the geometric properties from this. However, this approach does not yield the right hand rule property very easily.
11.1 Systems Of Equations, Geometric Interpretations

As you know, equations like $2x + 3y = 6$ can be graphed as straight lines in $\mathbb{R}^2$. To find the solution to two such equations, you could graph the two straight lines and the ordered pairs identifying the point (or points) of intersection would give the $x$ and $y$ values of the solution to the two equations because such an ordered pair satisfies both equations. The following picture illustrates what can occur with two equations involving two variables.

In the first example of the above picture, there is a unique point of intersection. In the second, there are no points of intersection. The other thing which can occur is that the two lines are really the same line. For example, $x + y = 1$ and $2x + 2y = 2$ are relations which when graphed yield the same line. In this case there are infinitely many points in the simultaneous solution of these two equations, every ordered pair which is on the graph of the line. It is always this way when considering linear systems of equations. There is either no solution, exactly one or infinitely many although the reasons for this are not completely comprehended by considering a simple picture in two dimensions, $\mathbb{R}^2$.

**Example 11.1.1** Find the solution to the system $x + y = 3$, $y - x = 5$.

You can verify the solution is $(x, y) = (-1, 4)$. You can see this geometrically by graphing the equations of the two lines. If you do so correctly, you should obtain a graph which looks something like the following in which the point of intersection represents the solution of the two equations.
Example 11.1.2 You can also imagine other situations such as the case of three intersecting lines having no common point of intersection or three intersecting lines which do intersect at a single point as illustrated in the following picture.

In the case of the first picture above, there would be no solution to the three equations whose graphs are the given lines. In the case of the second picture there is a solution to the three equations whose graphs are the given lines.

The points, \((x, y, z)\) satisfying an equation in three variables like \(2x + 4y - 5z = 8\) form a plane and geometrically, when you solve systems of equations involving three variables, you are taking intersections of planes. Consider the following picture involving two planes.

Notice how these two planes intersect in a line. It could also happen the two planes could fail to intersect.

Now imagine a third plane. One thing that could happen is this third plane could have an intersection with one of the first planes which results in a line which fails to intersect the first line as illustrated in the following picture.

\(^1\)Don’t worry about why this is at this time. It is not important. The following discussion is intended to show you that geometric considerations like this don’t take you anywhere. It is the algebraic procedures which are important and lead to important applications.
Thus there is no point which lies in all three planes. The picture illustrates the situation in which the line of intersection of the new plane with one of the original planes forms a line parallel to the line of intersection of the first two planes. However, in three dimensions, it is possible for two lines to fail to intersect even though they are not parallel. Such lines are called **skew lines**. You might consider whether there exist two skew lines, each of which is the intersection of a pair of planes selected from a set of exactly three planes such that there is no point of intersection between the three planes. You can also see that if you tilt one of the planes you could obtain every pair of planes having a nonempty intersection in a line and yet there may be no point in the intersection of all three.

It could happen also that the three planes could intersect in a single point as shown in the following picture.

In this case, the three planes have a single point of intersection. The three planes could also intersect in a line.
Thus in the case of three equations having three variables, the planes determined by these equations could intersect in a single point, a line, or even fail to intersect at all. You see that in three dimensions there are many possibilities. If you want to waste some time, you can try to imagine all the things which could happen but this will not help for more variables than 3 which is where many of the important applications lie.

Relations like $x + y - 2z + 4w = 8$ are often called hyper-planes. However, it is impossible to draw pictures of such things. The only rational and useful way to deal with this subject is through the use of algebra not art. Mathematics exists partly to free us from having to always draw pictures in order to draw conclusions.

11.2 Systems Of Equations, Algebraic Procedures

11.2.1 Elementary Operations

Consider the following example.

Example 11.2.1 Find $x$ and $y$ such that

$$x + y = 7 \quad \text{and} \quad 2x - y = 8. \quad (11.1)$$

The set of ordered pairs, $(x, y)$ which solve both equations is called the solution set.

You can verify that $(x, y) = (5, 2)$ is a solution to the above system. The interesting question is this: If you were not given this information to verify, how could you determine the solution? You can do this by using the following basic operations on the equations, none of which change the set of solutions of the system of equations.

Definition 11.2.2 Elementary operations are those operations consisting of the following.

1. Interchange the order in which the equations are listed.
2. Multiply any equation by a nonzero number.
3. Replace any equation with itself added to a multiple of another equation.

The evocative semi word, “hyper” conveys absolutely no meaning but is traditional usage which makes the terminology sound more impressive than something like long wide flat thing. Later we will discuss some terms which are not just evocative but yield real understanding.
Example 11.2.3 To illustrate the third of these operations on this particular system, consider the following.

\[ \begin{align*}
  x + y &= 7 \\
  2x - y &= 8 
\end{align*} \]

The system has the same solution set as the system

\[ \begin{align*}
  x + y &= 7 \\
  -3y &= -6 
\end{align*} \]

To obtain the second system, take the second equation of the first system and add -2 times the first equation to obtain

\[-3y = -6.\]

Now, this clearly shows that \( y = 2 \) and so it follows from the other equation that \( x + 2 = 7 \) and so \( x = 5 \).

Of course a linear system may involve many equations and many variables. The solution set is still the collection of solutions to the equations. In every case, the above operations of Definition 11.2.2 do not change the set of solutions to the system of linear equations.

Theorem 11.2.4 Suppose you have two equations, involving the variables,

\[ (x_1, \ldots, x_n) \]

\[ E_1 = f_1, \quad E_2 = f_2 \]  \hspace{1cm} (11.2)

where \( E_1 \) and \( E_2 \) are expressions involving the variables and \( f_1 \) and \( f_2 \) are constants.

(In the above example there are only two variables, \( x \) and \( y \) and \( E_1 = x + y \) while \( E_2 = 2x - y \).) Then the system \( E_1 = f_1, E_2 = f_2 \) has the same solution set as

\[ E_1 = f_1, \quad E_2 + aE_1 = f_2 + af_1. \]  \hspace{1cm} (11.3)

Also the system \( E_1 = f_1, E_2 = f_2 \) has the same solutions as the system, \( E_2 = f_2, E_1 = f_1 \). The system \( E_1 = f_1, E_2 = f_2 \) has the same solution as the system \( E_1 = f_1, aE_2 = af_2 \) provided \( a \neq 0 \).

Proof: If \((x_1, \ldots, x_n)\) solves \( E_1 = f_1, E_2 = f_2 \), then it solves the first equation in \( E_1 = f_1, \quad E_2 + aE_1 = f_2 + af_1. \) Also, it satisfies \( aE_1 = af_1 \) and so, since it also solves \( E_2 = f_2 \) it must solve \( E_2 + aE_1 = f_2 + af_1 \). Therefore, if \((x_1, \ldots, x_n)\) solves \( E_1 = f_1, E_2 = f_2 \), it must also solve \( E_2 + aE_1 = f_2 + af_1. \) On the other hand, if it solves the system \( E_1 = f_1 \) and \( E_2 + aE_1 = f_2 + af_1 \) then \( aE_1 = af_1 \) and so you can subtract these equal quantities from both sides of \( E_2 + aE_1 = f_2 + af_1 \) to obtain \( E_2 = f_2 \) showing that it satisfies \( E_1 = f_1, E_2 = f_2 \).

The second assertion of the theorem which says that the system \( E_1 = f_1, E_2 = f_2 \) has the same solution as the system, \( E_2 = f_2, E_1 = f_1 \), is seen to be true because it involves nothing more than listing the two equations in a different order. They are the same equations.

The third assertion of the theorem which says \( E_1 = f_1, E_2 = f_2 \) has the same solution as the system \( E_1 = f_1, aE_2 = af_2 \) provided \( a \neq 0 \) is verified as follows: If \((x_1, \ldots, x_n)\) is a solution of \( E_1 = f_1, E_2 = f_2 \), then it is a solution to \( E_1 = f_1, aE_2 = af_2 \) because the second system only involves multiplying the equation, \( E_2 = f_2 \) by \( a \). If \((x_1, \ldots, x_n)\) is a solution of \( E_1 = f_1, aE_2 = af_2 \), then upon multiplying \( aE_2 = af_2 \) by the number, \( 1/a \), you find that \( E_2 = f_2 \).

Stated simply, the above theorem shows that the elementary operations do not change the solution set of a system of equations.
Here is an example in which there are three equations and three variables. You want to find values for $x, y, z$ such that each of the given equations are satisfied when these values are plugged in to the equations.

**Example 11.2.5** Find the solutions to the system,

\[
\begin{align*}
    x + 3y + 6z &= 25 \\
    2x + 7y + 14z &= 58 \\
    2y + 5z &= 19
\end{align*}
\]  

(11.4)

To solve this system replace the second equation by $(-2)$ times the first equation added to the second. This yields the system

\[
\begin{align*}
    x + 3y + 6z &= 25 \\
    y + 2z &= 8 \\
    2y + 5z &= 19
\end{align*}
\]  

(11.5)

Now take $(-2)$ times the second and add to the third. More precisely, replace the third equation with $(-2)$ times the second added to the third. This yields the system

\[
\begin{align*}
    x + 3y + 6z &= 25 \\
    y + 2z &= 8 \\
    z &= 3
\end{align*}
\]  

(11.6)

At this point, you can tell what the solution is. This system has the same solution as the original system and in the above, $z = 3$. Then using this in the second equation, it follows $y + 6 = 8$ and so $y = 2$. Now using this in the top equation yields $x + 6 + 18 = 25$ and so $x = 1$. This process is called **back substitution**.

Alternatively, in (11.6) you could have continued as follows. Add $(-2)$ times the bottom equation to the middle and then add $(-6)$ times the bottom to the top. This yields

\[
\begin{align*}
    x + 3y &= 7 \\
    y &= 2 \\
    z &= 3
\end{align*}
\]

Now add $(-3)$ times the second to the top. This yields

\[
\begin{align*}
    x &= 1 \\
    y &= 2 \\
    z &= 3
\end{align*}
\]

a system which has the same solution set as the original system. This avoided back substitution and led to the same solution set.

**11.2.2 Gauss Elimination**

A less cumbersome way to represent a linear system is to write it as an **augmented matrix**. For example the linear system, (11.4) can be written as

\[
\begin{pmatrix}
1 & 3 & 6 & | & 25 \\
2 & 7 & 14 & | & 58 \\
0 & 2 & 5 & | & 19
\end{pmatrix}
\]

It has exactly the same information as the original system but here it is understood there is an $x$ column, \[
\begin{pmatrix}
1 \\
2 \\
0
\end{pmatrix}
\], a $y$ column, \[
\begin{pmatrix}
3 \\
7 \\
2
\end{pmatrix}
\] and a $z$ column, \[
\begin{pmatrix}
6 \\
14 \\
5
\end{pmatrix}
\]. The rows
correspond to the equations in the system. Thus the top row in the augmented matrix corresponds to the equation,

\[ x + 3y + 6z = 25. \]

Now when you replace an equation with a multiple of another equation added to itself, you are just taking a row of this augmented matrix and replacing it with a multiple of another row added to it. Thus the first step in solving [11.4] would be to take \((-2)\) times the first row of the augmented matrix above and add it to the second row,

\[
\begin{pmatrix}
1 & 3 & 6 & | & 25 \\
0 & 1 & 2 & | & 8 \\
0 & 2 & 5 & | & 19
\end{pmatrix}
\]

Note how this corresponds to [11.5]. Next take \((-2)\) times the second row and add to the third,

\[
\begin{pmatrix}
1 & 3 & 6 & | & 25 \\
0 & 1 & 2 & | & 8 \\
0 & 0 & 1 & | & 3
\end{pmatrix}
\]

This augmented matrix corresponds to the system

\[
\begin{align*}
x + 3y + 6z &= 25 \\
y + 2z &= 8 \\
z &= 3
\end{align*}
\]

which is the same as [11.6]. By back substitution you obtain the solution \(x = 1, y = 6, \) and \(z = 3.\)

In general a linear system is of the form

\[
a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\
\vdots \\
a_{m1}x_1 + \cdots + a_{mn}x_n = b_m
\]

where the \(x_i\) are variables and the \(a_{ij}\) and \(b_i\) are constants. This system can be represented by the augmented matrix,

\[
\begin{pmatrix}
a_{11} & \cdots & a_{1n} & | & b_1 \\
\vdots & \ddots & \vdots & | & \vdots \\
a_{m1} & \cdots & a_{mn} & | & b_m
\end{pmatrix}
\]

Changes to the system of equations in [11.7] as a result of an elementary operations translate into changes of the augmented matrix resulting from a row operation. Note that Theorem [11.2.4] implies that the row operations deliver an augmented matrix for a system of equations which has the same solution set as the original system.

**Definition 11.2.6** The row operations consist of the following

1. Switch two rows.
2. Multiply a row by a nonzero number.
3. Replace a row by a multiple of another row added to it.

**Gauss elimination** is a systematic procedure to simplify an augmented matrix to a reduced form. In the following definition, the term “leading entry” refers to the first nonzero entry of a row when scanning the row from left to right.
Definition 11.2.7  An augmented matrix is in **echelon form** if

1. All nonzero rows are above any rows of zeros.
2. Each leading entry of a row is in a column to the right of the leading entries of any rows above it.

Definition 11.2.8  An augmented matrix is in **row reduced echelon form** if

1. All nonzero rows are above any rows of zeros.
2. Each leading entry of a row is in a column to the right of the leading entries of any rows above it.
3. All entries in a column above and below a leading entry are zero.
4. Each leading entry is a 1, the only nonzero entry in its column.

Example 11.2.9  Here are some augmented matrices which are in row reduced echelon form.

\[
\begin{pmatrix}
1 & 0 & 0 & 5 & 8 & | & 0 \\
0 & 0 & 1 & 2 & 7 & | & 0 \\
0 & 0 & 0 & 0 & 0 & | & 1 \\
0 & 0 & 0 & 0 & 0 & | & 0
\end{pmatrix}
, \quad
\begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]

Example 11.2.10  Here are augmented matrices in echelon form which are not in row reduced echelon form but which are in echelon form.

\[
\begin{pmatrix}
1 & 0 & 6 & 5 & 8 & | & 2 \\
0 & 0 & 2 & 2 & 7 & | & 3 \\
0 & 0 & 0 & 0 & 0 & | & 1 \\
0 & 0 & 0 & 0 & 0 & | & 0
\end{pmatrix}
, \quad
\begin{pmatrix}
1 & 3 & 5 & 4 \\
0 & 2 & 0 & 7 \\
0 & 0 & 3 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]

Example 11.2.11  Here are some augmented matrices which are not in echelon form.

\[
\begin{pmatrix}
0 & 0 & 0 & 0 \\
1 & 2 & 3 & 3 \\
0 & 1 & 0 & 2 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix}
, \quad
\begin{pmatrix}
1 & 2 & 3 \\
2 & 4 & -6 \\
4 & 0 & 7 \\
0 & 0 & 1 \\
0 & 0 & 0
\end{pmatrix}
, \quad
\begin{pmatrix}
0 & 2 & 3 & 3 \\
1 & 5 & 0 & 2 \\
7 & 5 & 0 & 1 \\
0 & 0 & 1 & 0
\end{pmatrix}
\]

Definition 11.2.12  A **pivot position** in a matrix is the location of a leading entry in an echelon form resulting from the application of row operations to the matrix. A **pivot column** is a column that contains a pivot position.

For example consider the following.

Example 11.2.13  Suppose

\[
A = \begin{pmatrix}
1 & 2 & 3 & | & 4 \\
3 & 2 & 1 & | & 6 \\
4 & 4 & 4 & | & 10
\end{pmatrix}
\]

Where are the pivot positions and pivot columns?
Replace the second row by \(-3\) times the first added to the second. This yields
\[
\begin{pmatrix}
1 & 2 & 3 & | & 4 \\
0 & -4 & -8 & | & -6 \\
4 & 4 & 4 & | & 10 \\
\end{pmatrix}.
\]
This is not in reduced echelon form so replace the bottom row by \(-4\) times the top row added to the bottom. This yields
\[
\begin{pmatrix}
1 & 2 & 3 & | & 4 \\
0 & -4 & -8 & | & -6 \\
0 & -4 & -8 & | & -6 \\
\end{pmatrix}.
\]
This is still not in reduced echelon form. Replace the bottom row by \(-1\) times the middle row added to the bottom. This yields
\[
\begin{pmatrix}
1 & 2 & 3 & | & 4 \\
0 & -4 & -8 & | & -6 \\
0 & 0 & 0 & | & 0 \\
\end{pmatrix}
\]
which is in echelon form, although not in reduced echelon form. Therefore, the pivot positions in the original matrix are the locations corresponding to the first row and first column and the second row and second columns as shown in the following:
\[
\begin{pmatrix}
1 & 2 & 3 & | & 4 \\
3 & 2 & 1 & | & 6 \\
4 & 4 & 4 & | & 10 \\
\end{pmatrix}
\]
Thus the pivot columns in the matrix are the first two columns.

The following is the algorithm for obtaining a matrix which is in row reduced echelon form.

**Algorithm 11.2.14**

This algorithm tells how to start with a matrix and do row operations on it in such a way as to end up with a matrix in row reduced echelon form.

1. Find the first nonzero column from the left. This is the first pivot column. The position at the top of the first pivot column is the first pivot position. Switch rows if necessary to place a nonzero number in the first pivot position.

2. Use row operations to zero out the entries below the first pivot position.

3. Ignore the row containing the most recent pivot position identified and the rows above it. Repeat steps 1 and 2 to the remaining sub-matrix, the rectangular array of numbers obtained from the original matrix by deleting the rows you just ignored. Repeat the process until there are no more rows to modify. The matrix will then be in echelon form.

4. Moving from right to left, use the nonzero elements in the pivot positions to zero out the elements in the pivot columns which are above the pivots.

5. Divide each nonzero row by the value of the leading entry. The result will be a matrix in row reduced echelon form.

This row reduction procedure applies to both augmented matrices and non augmented matrices. There is nothing special about the augmented column with respect to the row reduction procedure.
Example 11.2.15 Here is a matrix.

\[
\begin{bmatrix}
0 & 0 & 2 & 3 & 2 \\
0 & 1 & 1 & 4 & 3 \\
0 & 0 & 1 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & 1
\end{bmatrix}
\]

Do row reductions till you obtain a matrix in echelon form. Then complete the process by producing one in row reduced echelon form.

The pivot column is the second. Hence the pivot position is the one in the first row and second column. Switch the first two rows to obtain a nonzero entry in this pivot position.

\[
\begin{bmatrix}
0 & 1 & 1 & 4 & 3 \\
0 & 0 & 2 & 3 & 2 \\
0 & 0 & 1 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & 1
\end{bmatrix}
\]

Step two is not necessary because all the entries below the first pivot position in the resulting matrix are zero. Now ignore the top row and the columns to the left of this first pivot position. Thus you apply the same operations to the smaller matrix,

\[
\begin{bmatrix}
2 & 3 & 2 \\
1 & 2 & 2 \\
0 & 0 & 0 \\
0 & 2 & 1
\end{bmatrix}
\]

The next pivot column is the third corresponding to the first in this smaller matrix and the second pivot position is therefore, the one which is in the second row and third column. In this case it is not necessary to switch any rows to place a nonzero entry in this position because there is already a nonzero entry there. Multiply the third row of the original matrix by \(-2\) and then add the second row to it. This yields

\[
\begin{bmatrix}
0 & 1 & 1 & 4 & 3 \\
0 & 0 & 2 & 3 & 2 \\
0 & 0 & 0 & -1 & -2 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & 1
\end{bmatrix}
\]

The next matrix the steps in the algorithm are applied to is

\[
\begin{bmatrix}
-1 & -2 \\
0 & 0 \\
2 & 1
\end{bmatrix}
\]

The first pivot column is the first column in this case and no switching of rows is necessary because there is a nonzero entry in the first pivot position. Therefore, the algorithm yields for the next step

\[
\begin{bmatrix}
0 & 1 & 1 & 4 & 3 \\
0 & 0 & 2 & 3 & 2 \\
0 & 0 & 0 & -1 & -2 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -3
\end{bmatrix}
\]
Now the algorithm will be applied to the matrix,
\[
\begin{pmatrix}
0 & -3
\end{pmatrix}
\]
There is only one column and it is nonzero so this single column is the pivot column. Therefore, the algorithm yields the following matrix for the echelon form.
\[
\begin{pmatrix}
0 & 1 & 1 & 4 & 3 \\
0 & 0 & 2 & 3 & 2 \\
0 & 0 & 0 & -1 & -2 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
To complete placing the matrix in reduced echelon form, multiply the third row by 3 and add $-2$ times the fourth row to it. This yields
\[
\begin{pmatrix}
0 & 1 & 1 & 4 & 0 \\
0 & 0 & 6 & 9 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Next multiply the second row by 3 and take $2$ times the fourth row and add to it. Then add the fourth row to the first.
\[
\begin{pmatrix}
0 & 3 & 3 & 0 & 0 \\
0 & 0 & 6 & 0 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Next work on the fourth column in the same way.
\[
\begin{pmatrix}
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 6 & 0 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Take $-1/2$ times the second row and add to the first.
\[
\begin{pmatrix}
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 6 & 0 & 0 \\
0 & 0 & 0 & -3 & 0 \\
0 & 0 & 0 & 0 & -3 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
Finally, divide by the value of the leading entries in the nonzero rows.
\[
\begin{pmatrix}
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\]
The above algorithm is the way a computer would obtain a reduced echelon form for a given matrix. It is not necessary for you to pretend you are a computer but if
you like to do so, the algorithm described above will work. The main idea is to do row operations in such a way as to end up with a matrix in echelon form or row reduced echelon form because when this has been done, the resulting augmented matrix will allow you to describe the solutions to the linear system of equations in a meaningful way.

**Example 11.2.16** Give the complete solution to the system of equations, $5x + 10y - 7z = -2$, $2x + 4y - 3z = -1$, and $3x + 6y + 5z = 9$.

The augmented matrix for this system is

$$
\begin{pmatrix}
2 & 4 & -3 & | & -1 \\
5 & 10 & -7 & | & -2 \\
3 & 6 & 5 & | & 9
\end{pmatrix}
$$

Multiply the second row by $2$, the first row by $5$, and then take $(-1)$ times the first row and add to the second. Then multiply the first row by $1/5$. This yields

$$
\begin{pmatrix}
2 & 4 & -3 & | & -1 \\
0 & 0 & 1 & | & 1 \\
3 & 6 & 5 & | & 9
\end{pmatrix}
$$

Now, combining some row operations, take $(-3)$ times the first row and add this to $2$ times the last row and replace the last row with this. This yields.

$$
\begin{pmatrix}
2 & 4 & -3 & | & -1 \\
0 & 0 & 1 & | & 1 \\
0 & 0 & 1 & | & 21
\end{pmatrix}
$$

One more row operation, taking $(-1)$ times the second row and adding to the bottom yields.

$$
\begin{pmatrix}
2 & 4 & -3 & | & -1 \\
0 & 0 & 1 & | & 1 \\
0 & 0 & 0 & | & 20
\end{pmatrix}
$$

This is impossible because the last row indicates the need for a solution to the equation

$$0x + 0y + 0z = 20$$

and there is no such thing because $0 \neq 20$. This shows there is no solution to the three given equations. When this happens, the system is called **inconsistent**. In this case it is very easy to describe the solution set. The system has no solution.

Here is another example based on the use of row operations.

**Example 11.2.17** Give the complete solution to the system of equations, $3x - y - 5z = 9$, $y - 10z = 0$, and $-2x + y = -6$.

The augmented matrix of this system is

$$
\begin{pmatrix}
3 & -1 & -5 & | & 9 \\
0 & 1 & -10 & | & 0 \\
-2 & 1 & 0 & | & -6
\end{pmatrix}
$$

Replace the last row with $2$ times the top row added to $3$ times the bottom row. This gives

$$
\begin{pmatrix}
3 & -1 & -5 & | & 9 \\
0 & 1 & -10 & | & 0 \\
0 & 1 & -10 & | & 0
\end{pmatrix}
$$
The entry, 3 in this sequence of row operations is called the *pivot*. It is used to create zeros in the other places of the column. Next take $-1$ times the middle row and add to the bottom. Here the 1 in the second row is the pivot.

\[
\begin{pmatrix}
3 & -1 & -5 & | & 9 \\
0 & 1 & -10 & | & 0 \\
0 & 0 & 0 & | & 0
\end{pmatrix}
\]

Take the middle row and add to the top and then divide the top row which results by 3.

\[
\begin{pmatrix}
1 & 0 & -5 & | & 3 \\
0 & 1 & -10 & | & 0 \\
0 & 0 & 0 & | & 0
\end{pmatrix}
\]

This is in reduced echelon form. The equations corresponding to this reduced echelon form are $y = 10t$ and $x = 3 + 5z$. Apparently $z$ can equal any number. Let’s call this number, $t$. Therefore, the solution set of this system is $x = 3 + 5t$, $y = 10t$, and $z = t$ where $t$ is completely arbitrary. The system has an infinite set of solutions which are given in the above simple way. This is what it is all about, finding the solutions to the system.

There is some terminology connected to this which is useful. Recall how each column corresponds to a variable in the original system of equations. The variables corresponding to a pivot column are called *basic variables*. The other variables are called *free variables*. In Example 11.2.17 there was one free variable, $z$, and two basic variables, $x$ and $y$. In describing the solution to the system of equations, the free variables are assigned a parameter. In Example 11.2.17 this parameter was $t$. Sometimes there are many free variables and in these cases, you need to use many parameters. Here is another example.

**Example 11.2.18** Find the solution to the system

\[
\begin{align*}
x + 2y - z + w &= 3 \\
x + y - z + w &= 1 \\
x + 3y - z + w &= 5
\end{align*}
\]

The augmented matrix is

\[
\begin{pmatrix}
1 & 2 & -1 & 1 & | & 3 \\
1 & 1 & -1 & 1 & | & 1 \\
1 & 3 & -1 & 1 & | & 5
\end{pmatrix}
\]

Take $-1$ times the first row and add to the second. Then take $-1$ times the first row and add to the third. This yields

\[
\begin{pmatrix}
1 & 2 & -1 & 1 & | & 3 \\
0 & -1 & 0 & 0 & | & -2 \\
0 & 1 & 0 & 0 & | & 2
\end{pmatrix}
\]

Now add the second row to the bottom row

\[
\begin{pmatrix}
1 & 2 & -1 & 1 & | & 3 \\
0 & -1 & 0 & 0 & | & -2 \\
0 & 0 & 0 & 0 & | & 0
\end{pmatrix}
\]

This matrix is in echelon form and you see the basic variables are $x$ and $y$ while the free variables are $z$ and $w$. Assign $s$ to $z$ and $t$ to $w$. Then the second row yields the
equation, \( y = 2 \) while the top equation yields the equation, \( x + 2y - s + t = 3 \) and so since \( y = 2 \), this gives \( x + 4 - s + t = 3 \) showing that \( x = -1 + s - t, y = 2, z = s, \) and \( w = t \). It is customary to write this in the form

\[
\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} -1 + s - t \\ 2 \\ s \\ t \end{pmatrix}.
\] (11.10)

This is another example of a system which has an infinite solution set but this time the solution set depends on two parameters, not one. Most people find it less confusing in the case of an infinite solution set to first place the augmented matrix in row reduced echelon form rather than just echelon form before seeking to write down the description of the solution. In the above, this means we don’t stop with the echelon form [11.9]. Instead we first place it in reduced echelon form as follows.

\[
\begin{pmatrix}
1 & 0 & -1 & 1 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\end{pmatrix}
\begin{pmatrix}
y \\
z \\
w \\
\end{pmatrix}
\] .

Then the solution is \( y = 2 \) from the second row and \( x = -1 + z - w \) from the first. Thus letting \( z = s \) and \( w = t \), the solution is given in [11.10].

The number of free variables is always equal to the number of different parameters used to describe the solution. If there are no free variables, then either there is no solution as in the case where row operations yield an echelon form like

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

or there is a unique solution as in the case where row operations yield an echelon form like

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

Also, sometimes there are free variables and no solution as in the following:

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

There are a lot of cases to consider but it is not necessary to make a major production of this. Do row operations till you obtain a matrix in echelon form or reduced echelon form and determine whether there is a solution. If there is, see if there are free variables. In this case, there will be infinitely many solutions. Find them by assigning different parameters to the free variables and obtain the solution. If there are no free variables, then there will be a unique solution which is easily determined once the augmented matrix is in echelon or row reduced echelon form. In every case, the process yields a straightforward way to describe the solutions to the linear system. As indicated above, you are probably less likely to become confused if you place the augmented matrix in row reduced echelon form rather than just echelon form.

In summary,
Definition 11.2.19 A system of linear equations is a list of equations,

\[ a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \]
\[ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \]
\[ \vdots \]
\[ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \]

where \( a_{ij} \) are numbers, and \( b_j \) is a number. The above is a system of \( m \) equations in the \( n \) variables, \( x_1, x_2, \ldots, x_n \). Nothing is said about the relative size of \( m \) and \( n \). Written more simply in terms of summation notation, the above can be written in the form

\[ \sum_{j=1}^{n} a_{ij}x_j = f_j, \ i = 1, 2, 3, \ldots, m \]

It is desired to find \((x_1, \ldots, x_n)\) solving each of the equations listed.

As illustrated above, such a system of linear equations may have a unique solution, no solution, or infinitely many solutions and these are the only three cases which can occur for any linear system. Furthermore, you do exactly the same things to solve any linear system. You write the augmented matrix and do row operations until you get a simpler system in which it is possible to see the solution, usually obtaining a matrix in echelon or reduced echelon form. All is based on the observation that the row operations do not change the solution set. You can have more equations than variables, fewer equations than variables, etc. It doesn’t matter. You always set up the augmented matrix and go to work on it.

Definition 11.2.20 A system of linear equations is called consistent if there exists a solution. It is called inconsistent if there is no solution.

These are reasonable words to describe the situations of having or not having a solution. If you think of each equation as a condition which must be satisfied by the variables, consistent would mean there is some choice of variables which can satisfy all the conditions. Inconsistent would mean there is no choice of the variables which can satisfy each of the conditions.

11.3 Exercises

1. Find the point, \((x_1, y_1)\) which lies on both lines, \( x + 3y = 1 \) and \( 4x - y = 3 \).
2. Solve Problem 1 graphically. That is, graph each line and see where they intersect.
3. Find the point of intersection of the two lines \( 3x + y = 3 \) and \( x + 2y = 1 \).
4. Solve Problem 3 graphically. That is, graph each line and see where they intersect.
5. Do the three lines, \( x + 2y = 1, 2x - y = 1, \) and \( 4x + 3y = 3 \) have a common point of intersection? If so, find the point and if not, tell why they don’t have such a common point of intersection.
6. Do the three planes, \( x + y - 3z = 2, 2x + y + z = 1, \) and \( 3x + 2y - 2z = 0 \) have a common point of intersection? If so, find one and if not, tell why there is no such point.
7. You have a system of \( k \) equations in two variables, \( k \geq 2 \). Explain the geometric significance of
(a) No solution.
(b) A unique solution.
(c) An infinite number of solutions.

8. Here is an augmented matrix in which \( * \) denotes an arbitrary number and \( \Box \) denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

\[
\begin{pmatrix}
\Box & * & * & * & | & * \\
0 & \Box & * & 0 & | & * \\
0 & 0 & \Box & * & * & | & * \\
0 & 0 & 0 & 0 & \Box & | & * \\
\end{pmatrix}
\]

9. Here is an augmented matrix in which \( * \) denotes an arbitrary number and \( \Box \) denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

\[
\begin{pmatrix}
\Box & * & | & * \\
0 & \Box & * & | & * \\
0 & 0 & \Box & | & * \\
\end{pmatrix}
\]

10. Here is an augmented matrix in which \( * \) denotes an arbitrary number and \( \Box \) denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

\[
\begin{pmatrix}
\Box & * & * & * & | & * \\
0 & \Box & 0 & 0 & | & * \\
0 & 0 & 0 & \Box & * & | & * \\
0 & 0 & 0 & 0 & \Box & | & * \\
\end{pmatrix}
\]

11. Here is an augmented matrix in which \( * \) denotes an arbitrary number and \( \Box \) denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

\[
\begin{pmatrix}
\Box & * & * & * & | & * \\
0 & \Box & * & 0 & | & * \\
0 & 0 & 0 & \Box & 0 & | & 0 \\
0 & 0 & 0 & 0 & \Box & | & 0 \\
\end{pmatrix}
\]

12. Suppose a system of equations has fewer equations than variables. Must such a system be consistent? If so, explain why and if not, give an example which is not consistent.

13. If a system of equations has more equations than variables, can it have a solution? If so, give an example and if not, tell why not.

14. Find \( h \) such that

\[
\begin{pmatrix}
2 & h & | & 4 \\
3 & 6 & | & 7 \\
\end{pmatrix}
\]

is the augmented matrix of an inconsistent matrix.

15. Find \( h \) such that

\[
\begin{pmatrix}
1 & h & | & 3 \\
2 & 4 & | & 6 \\
\end{pmatrix}
\]

is the augmented matrix of a consistent matrix.
16. Find $h$ such that
\[
\begin{pmatrix}
1 & 1 & | & 4 \\
3 & h & | & 12
\end{pmatrix}
\]
is the augmented matrix of a consistent matrix.

17. Choose $h$ and $k$ such that the augmented matrix shown has one solution. Then choose $h$ and $k$ such that the system has no solutions. Finally, choose $h$ and $k$ such that the system has infinitely many solutions.
\[
\begin{pmatrix}
1 & h & | & 2 \\
2 & 4 & | & k
\end{pmatrix}
\]

18. Choose $h$ and $k$ such that the augmented matrix shown has one solution. Then choose $h$ and $k$ such that the system has no solutions. Finally, choose $h$ and $k$ such that the system has infinitely many solutions.
\[
\begin{pmatrix}
1 & 2 & | & 2 \\
2 & h & | & k
\end{pmatrix}
\]

19. Determine if the system is consistent. If so, is the solution unique?
\[
x + 2y + z - w = 2 \\
x - y + z + w = 1 \\
2x + y - z = 1 \\
4x + 2y + z = 5
\]

20. Determine if the system is consistent. If so, is the solution unique?
\[
x + 2y + z - w = 2 \\
x - y + z + w = 0 \\
2x + y - z = 1 \\
4x + 2y + z = 3
\]

21. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 2 & 0 & | & 2 \\
1 & 3 & 4 & | & 2 \\
1 & 0 & 2 & | & 1
\end{pmatrix}
\]

22. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 2 & 0 & | & 2 \\
2 & 0 & 1 & | & 1 \\
3 & 2 & 1 & | & 3
\end{pmatrix}
\]

23. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 1 & 0 & | & 1 \\
1 & 0 & 4 & | & 2
\end{pmatrix}
\]

24. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 2 & 1 & 1 & | & 2 \\
0 & 1 & 0 & 1 & 2 & | & 1 \\
1 & 2 & 0 & 0 & 1 & | & 3 \\
1 & 0 & 1 & 0 & 2 & | & 2
\end{pmatrix}
\]
25. Find the general solution of the system whose augmented matrix is
\[
\begin{pmatrix}
1 & 0 & 2 & 1 & 1 & | & 2 \\
0 & 1 & 0 & 1 & 2 & | & 1 \\
0 & 2 & 0 & 0 & 1 & | & 3 \\
1 & -1 & 2 & 2 & 2 & | & 0
\end{pmatrix}
\]

26. Give the complete solution to the system of equations, \(7x + 14y + 15z = 22\), \(2x + 4y + 3z = 5\), and \(3x + 6y + 10z = 13\).

27. Give the complete solution to the system of equations, \(3x - y + 4z = 6\), \(y + 8z = 0\), and \(-2x + y = -4\).

28. Give the complete solution to the system of equations, \(9x - 2y + 4z = -17\), \(13x - 3y + 6z = -25\), and \(-2x - z = 3\).

29. Give the complete solution to the system of equations, \(65x + 84y + 16z = 546\), \(81x + 105y + 20z = 682\), and \(84x + 110y + 21z = 713\).

30. Give the complete solution to the system of equations, \(-8x + 2y + 5z = 18\), \(-8x + 3y + 5z = 13\), and \(-4x + y + 5z = 19\).

31. Give the complete solution to the system of equations, \(3x - y - 2z = 3\), \(y - 4z = 0\), and \(-2x + y = -2\).

32. Give the complete solution to the system of equations, \(-9x + 15y = 66\), \(-11x + 18y = 79\), \(-x + y = 4\), and \(z = 3\).

33. Give the complete solution to the system of equations, \(-19x + 8y = -108\), \(-71x + 30y = -404\), \(-2x + y = -12\), \(4x + z = 14\).

34. Give the complete solution to the system of equations, \(-5x + 2y - z = 0\) and \(-5x - 2y - z = 0\). Both equations equal zero and so \(-5x + 2y - z = -5x - 2y - z\) which is equivalent to \(y = 0\). Thus \(x\) and \(z\) can equal anything. But when \(x = 1\), \(z = -4\), and \(y = 0\) are plugged in to the equations, it doesn’t work. Why?

35. Four times the weight of Gaston is 150 pounds more than the weight of Ichabod.

36. Four times the weight of Ichabod is 660 pounds less than seventeen times the weight of Gaston.

37. Four times the weight of Gaston plus the weight of Siegfried equals 290 pounds. Brunhilde would balance all three of the others. Find the weights of the four people.

38. The steady state temperature, \(u\) in a plate solves Laplace’s equation, \(\Delta u = 0\). One way to approximate the solution which is often used is to divide the plate into a square mesh and require the temperature at each node to equal the average of the temperature at the four adjacent nodes. This procedure is justified by the mean value property of harmonic functions. In the following picture, the numbers represent the observed temperature at the indicated nodes. Your task is to find the temperature at the interior nodes, indicated by \(x, y, z,\) and \(w\). One of the equations is \(z = \frac{1}{4}(10 + 0 + w + x)\).
Matrices

There is a lot more to matrices than as a shortcut for solving systems of equations. The interesting and useful parts have to do with the consideration of matrices as concrete realizations of certain kinds of functions which take vectors and make them into other vectors.

12.1 Addition And Scalar Multiplication Of Matrices

You have now solved systems of equations by writing them in terms of an augmented matrix and then doing row operations on this augmented matrix. It turns out such rectangular arrays of numbers are important from many other different points of view. Numbers are also called scalars. I will refer to the set of numbers as \( \mathbb{F} \) sometimes when it is not important to worry about whether the number is real or complex. In fact, the scalars can come from other fields besides the field of real or complex number. However, if you wish, let \( \mathbb{F} \) be either the real numbers, \( \mathbb{R} \) or the complex numbers, \( \mathbb{C} \).

A matrix is a rectangular array of numbers. Several of them are referred to as matrices. For example, here is a matrix.

\[
\begin{pmatrix}
1 & 2 & 3 & 4 \\
5 & 2 & 8 & 7 \\
6 & -9 & 1 & 2
\end{pmatrix}
\]

The size or dimension of a matrix is defined as \( m \times n \) where \( m \) is the number of rows and \( n \) is the number of columns. The above matrix is a \( 3 \times 4 \) matrix because there are three rows and four columns. The first row is \( (1 \ 2 \ 3 \ 4) \), the second row is \( (5 \ 2 \ 8 \ 7) \) and so forth. The first column is \( \left( \begin{array}{c}
1 \\
5 \\
6
\end{array} \right) \). When specifying the size of a matrix, you always list the number of rows before the number of columns. Also, you can remember the columns are like columns in a Greek temple. They stand upright while the rows just lay there like rows made by a tractor in a plowed field. Elements of the matrix are identified according to position in the matrix. For example, 8 is in position 2,3 because it is in the second row and the third column. You might remember that you always list the rows before the columns by using the phrase Rowman Catholic. The symbol, \( (a_{ij}) \) refers to a matrix. The entry in the \( i^{th} \) row and the \( j^{th} \) column of this matrix is denoted by \( a_{ij} \). Using this notation on the above matrix, \( a_{23} = 8, a_{32} = -9, a_{12} = 2 \), etc.

There are various operations which are done on matrices. Matrices can be added multiplied by a scalar, and multiplied by other matrices. To illustrate scalar multiplication, consider the following example in which a matrix is being multiplied by the scalar,
3. 
\[
\begin{pmatrix}
1 & 2 & 3 & 4 \\
5 & 8 & 7 \\
6 & -9 & 1 & 2
\end{pmatrix}
\begin{pmatrix}
3 & 6 & 9 & 12 \\
15 & 6 & 24 & 21 \\
18 & -27 & 3 & 6
\end{pmatrix}
\]

The new matrix is obtained by multiplying every entry of the original matrix by the given scalar. If \( A \) is an \( m \times n \) matrix, \(-A\) is defined to equal \((-1)A\).

Two matrices must be the same size to be added. The sum of two matrices is a matrix which is obtained by adding the corresponding entries. Thus
\[
\begin{pmatrix}
1 & 2 \\
3 & 4 \\
5 & 2
\end{pmatrix}
+ 
\begin{pmatrix}
-1 & 4 \\
2 & 8 \\
6 & -4
\end{pmatrix}
= 
\begin{pmatrix}
0 & 6 \\
5 & 12 \\
11 & -2
\end{pmatrix}
\]

Two matrices are equal exactly when they are the same size and the corresponding entries are identical. Thus
\[
\begin{pmatrix}
0 & 0 \\
0 & 0 \\
0 & 0
\end{pmatrix}
\neq 
\begin{pmatrix}
0 & 0 \\
0 & 0 
\end{pmatrix}
\]

because they are different sizes. As noted above, you write \((c_{ij})\) for the matrix \(C\) whose \(ij^{th}\) entry is \(c_{ij}\). In doing arithmetic with matrices you must define what happens in terms of the \(c_{ij}\) sometimes called the \textit{entries} of the matrix or the \textit{components} of the matrix.

The above discussion stated for general matrices is given in the following definition.

\textbf{Definition 12.1.1 (Scalar Multiplication)} If \( A = (a_{ij}) \) and \( k \) is a scalar, then \( kA = (ka_{ij}) \).

\textbf{Example 12.1.2} \( 7 \begin{pmatrix} 2 & 0 \\ 1 & -4 \end{pmatrix} = \begin{pmatrix} 14 & 0 \\ 7 & -28 \end{pmatrix} \).

\textbf{Definition 12.1.3 (Addition)} If \( A = (a_{ij}) \) and \( B = (b_{ij}) \) are two \( m \times n \) matrices. Then \( A + B = C \) where
\[
C = (c_{ij})
\]

for \( c_{ij} = a_{ij} + b_{ij} \).

\textbf{Example 12.1.4}
\[
\begin{pmatrix}
1 & 2 & 3 \\
1 & 0 & 4
\end{pmatrix}
+ 
\begin{pmatrix}
5 & 2 & 3 \\
-6 & 2 & 1
\end{pmatrix}
= 
\begin{pmatrix}
6 & 4 & 6 \\
-5 & 2 & 5
\end{pmatrix}
\]

To save on notation, we will often use \( A_{ij} \) to refer to the \( ij^{th} \) entry of the matrix, \( A \).

\textbf{Definition 12.1.5 (The zero matrix)} The \( m \times n \) zero matrix is the \( m \times n \) matrix having every entry equal to zero. It is denoted by \( 0 \).

\textbf{Example 12.1.6} The \( 2 \times 3 \) zero matrix is \( \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \).

Note there are \( 2 \times 3 \) zero matrices, \( 3 \times 4 \) zero matrices, etc. In fact there is a zero matrix for every size.

\textbf{Definition 12.1.7 (Equality of matrices)} Let \( A \) and \( B \) be two matrices. Then \( A = B \) means that the two matrices are of the same size and for \( A = (a_{ij}) \) and \( B = (b_{ij}) \), \( a_{ij} = b_{ij} \) for all \( 1 \leq i \leq m \) and \( 1 \leq j \leq n \).
The following properties of matrices can be easily verified. You should do so.

- **Commutative Law Of Addition.**
  \[ A + B = B + A, \quad (12.1) \]

- **Associative Law for Addition.**
  \[ (A + B) + C = A + (B + C), \quad (12.2) \]

- **Existence of an Additive Identity**
  \[ A + 0 = A, \quad (12.3) \]

- **Existence of an Additive Inverse**
  \[ A + (-A) = 0, \quad (12.4) \]

Also for \( \alpha, \beta \) scalars, the following additional properties hold.

- **Distributive law over Matrix Addition.**
  \[ \alpha (A + B) = \alpha A + \alpha B, \quad (12.5) \]

- **Distributive law over Scalar Addition**
  \[ (\alpha + \beta) A = \alpha A + \beta A, \quad (12.6) \]

- **Associative law for Scalar Multiplication**
  \[ \alpha (\beta A) = \alpha \beta (A), \quad (12.7) \]

- **Rule for Multiplication by 1.**
  \[ 1A = A. \quad (12.8) \]

As an example, consider the Commutative Law of Addition. Let \( A + B = C \) and \( B + A = D \). Why is \( D = C \)?

\[ C_{ij} = A_{ij} + B_{ij} = B_{ij} + A_{ij} = D_{ij}. \]

Therefore, \( C = D \) because the \( ij^{th} \) entries are the same. Note that the conclusion follows from the commutative law of addition of numbers.

### 12.2 Multiplication Of Matrices

**Definition 12.2.1** Matrices which are \( n \times 1 \) or \( 1 \times n \) are called *vectors* and are often denoted by a bold letter. Thus the \( n \times 1 \) matrix

\[ x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \]

is also called a *column vector*. The \( 1 \times n \) matrix

\[ \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \]

is called a *row vector*. 
Although the following description of matrix multiplication may seem strange, it is in fact the most important and useful of the matrix operations. To begin with consider the case where a matrix is multiplied by a column vector. First consider a special case.

\[
\begin{pmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
\end{pmatrix}
\begin{pmatrix}
7 \\
8 \\
9 \\
\end{pmatrix}
= ?
\]

One way to remember this is as follows. Slide the vector, placing it on top the two rows as shown and then do the indicated operation.

\[
\begin{pmatrix}
7 & 8 & 9 \\
1 & 2 & 3 \\
4 & 5 & 6 \\
\end{pmatrix}
\Rightarrow
\begin{pmatrix}
7 \times 1 + 8 \times 2 + 9 \times 3 \\
7 \times 4 + 8 \times 5 + 9 \times 6 \\
\end{pmatrix}
= \begin{pmatrix}50 \\
122 \end{pmatrix}.
\]

multiply the numbers on the top by the numbers on the bottom and add them up to get a single number for each row of the matrix as shown above.

In more general terms,

\[
\begin{pmatrix}
a_{11} & a_{12} & a_{13} \\
a_{21} & a_{22} & a_{23} \\
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
\end{pmatrix}
= \begin{pmatrix}a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\
\end{pmatrix}.
\]

Another way to think of this is

\[
x_1 \begin{pmatrix}a_{11} \\
a_{21} \end{pmatrix} + x_2 \begin{pmatrix}a_{12} \\
a_{22} \end{pmatrix} + x_3 \begin{pmatrix}a_{13} \\
a_{23} \end{pmatrix}
\]

Thus you take \(x_1\) times the first column, add to \(x_2\) times the second column, and finally \(x_3\) times the third column. In general, here is the definition of how to multiply an \((m \times n)\) matrix times a \((n \times 1)\) matrix.

**Definition 12.2.2** Let \(A = A_{ij}\) be an \(m \times n\) matrix and let \(v\) be an \(n \times 1\) matrix,

\[
v = \begin{pmatrix}v_1 \\
\vdots \\
v_n \end{pmatrix}
\]

Then \(Av\) is an \(m \times 1\) matrix and the \(i^{th}\) component of this matrix is

\[
(Av)_i = A_{i1}v_1 + A_{i2}v_2 + \cdots + A_{in}v_n = \sum_{j=1}^{n} A_{ij}v_j.
\]

Thus

\[
Av = \begin{pmatrix}
\sum_{j=1}^{n} A_{1j}v_j \\
\vdots \\
\sum_{j=1}^{n} A_{mj}v_j
\end{pmatrix}.
\]

(12.9)

In other words, if

\[
A = (a_1, \cdots, a_n)
\]

where the \(a_k\) are the columns,

\[
Av = \sum_{k=1}^{n} v_k a_k
\]
This follows from 12.9 and the observation that the $j^{th}$ column of $A$ is

$$
\begin{bmatrix}
A_{1j} \\
A_{2j} \\
\vdots \\
A_{mj}
\end{bmatrix}
$$

so 12.9 reduces to

$$
v_1 \begin{bmatrix}
A_{11} \\
A_{21} \\
\vdots \\
A_{m1}
\end{bmatrix} + v_2 \begin{bmatrix}
A_{12} \\
A_{22} \\
\vdots \\
A_{m2}
\end{bmatrix} + \cdots + v_n \begin{bmatrix}
A_{1n} \\
A_{2n} \\
\vdots \\
A_{mn}
\end{bmatrix}
$$

Note also that multiplication by an $m \times n$ matrix takes an $n \times 1$ matrix, and produces an $m \times 1$ matrix.

Here is another example.

**Example 12.2.3** Compute

$$
\begin{bmatrix}
1 & 2 & 1 & 3 \\
0 & 2 & 1 & -2 \\
2 & 1 & 4 & 1
\end{bmatrix}
\begin{bmatrix}
1 \\
2 \\
0 \\
1
\end{bmatrix}.
$$

First of all this is of the form $(3 \times 4)(4 \times 1)$ and so the result should be a $(3 \times 1)$. Note how the inside numbers cancel. To get the element in the second row and first and only column, compute

$$
\sum_{k=1}^{4} a_{2k}v_k = a_{21}v_1 + a_{22}v_2 + a_{23}v_3 + a_{24}v_4 = 0 \times 1 + 2 \times 2 + 1 \times 0 + (-2) \times 1 = 2.
$$

You should do the rest of the problem and verify

$$
\begin{bmatrix}
1 & 2 & 1 & 3 \\
0 & 2 & 1 & -2 \\
2 & 1 & 4 & 1
\end{bmatrix}
\begin{bmatrix}
1 \\
2 \\
0 \\
1
\end{bmatrix} = \begin{bmatrix}
8 \\
2 \\
5
\end{bmatrix}.
$$

The next task is to multiply an $m \times n$ matrix times an $n \times p$ matrix. Before doing so, the following may be helpful.

For $A$ and $B$ matrices, in order to form the product, $AB$ the number of columns of $A$ must equal the number of rows of $B$.

These must match! $(m \times \overline{n}) (n \times p) = m \times p$

Note the two outside numbers give the size of the product. Remember:

**If the two middle numbers don't match,**

Then you can't multiply the matrices!
Definition 12.2.4 When the number of columns of \( A \) equals the number of rows of \( B \) the two matrices are said to be \textit{conformable} and the product, \( AB \) is obtained as follows. Let \( A \) be an \( m \times n \) matrix and let \( B \) be an \( n \times p \) matrix. Then \( B \) is of the form

\[
B = (b_1, \cdots, b_p)
\]

where \( b_k \) is an \( n \times 1 \) matrix or column vector. Then the \( m \times p \) matrix, \( AB \) is defined as follows:

\[
AB \equiv (Ab_1, \cdots, Ab_p) \quad (12.10)
\]

where \( Ab_k \) is an \( m \times 1 \) matrix or column vector which gives the \( k^{th} \) column of \( AB \).

Example 12.2.5 Multiply the following.

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

The first thing you need to check before doing anything else is whether it is possible to do the multiplication. The first matrix is a \( 2 \times 3 \) and the second matrix is a \( 3 \times 3 \). Therefore, is it possible to multiply these matrices. According to the above discussion it should be a \( 2 \times 3 \) matrix of the form

\[
\begin{pmatrix}
\text{First column} & \text{Second column} & \text{Third column}
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 0 \\
0 & 3 & 1 \\
-2 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 1 \\
0 & 2 & 1
\end{pmatrix}
\begin{pmatrix}
2 & 3 & 1 \\
0 & 2 & 1 \\
1 & 1
\end{pmatrix}
\]

You know how to multiply a matrix times a vector and so you do so to obtain each of the three columns. Thus

\[
\begin{pmatrix}
1 & 2 & 1 \\
0 & 2 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 2 & 0 \\
0 & 3 & 1 \\
-2 & 1 & 1
\end{pmatrix}
= \begin{pmatrix}
-1 & 9 & 3 \\
-2 & 7 & 3
\end{pmatrix}.
\]

Example 12.2.6 Multiply the following.

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

First check if it is possible. This is of the form \( (3 \times 3) \cdot (2 \times 3) \). The inside numbers do not match and so you can’t do this multiplication. This means that anything you write will be absolute nonsense because it is impossible to multiply these matrices in this order. Aren’t they the same two matrices considered in the previous example? Yes they are. It is just that here they are in a different order. This shows something you must always remember about matrix multiplication.

\[\text{Order Matters!}\]

\[\text{Matrix Multiplication Is Not Commutative!}\]

This is very different than multiplication of numbers!
12.3 The \( ij \)th Entry Of A Product

It is important to describe matrix multiplication in terms of entries of the matrices. What is the \( ij \)th entry of \( AB \)? It would be the \( i \)th entry of the \( j \)th column of \( AB \). Thus it would be the \( i \)th entry of \( Ab_j \). Now

\[
\mathbf{b}_j = \begin{pmatrix} B_{1j} \\ \vdots \\ B_{nj} \end{pmatrix}
\]

and from the above definition, the \( i \)th entry is

\[
\sum_{k=1}^{n} A_{ik}B_{kj}.
\]

(12.11)

In terms of pictures of the matrix, you are doing

\[
\begin{pmatrix}
A_{11} & A_{12} & \cdots & A_{1n} \\
A_{21} & A_{22} & \cdots & A_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
A_{m1} & A_{m2} & \cdots & A_{mn}
\end{pmatrix}
\begin{pmatrix}
B_{11} & B_{12} & \cdots & B_{1p} \\
B_{21} & B_{22} & \cdots & B_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
B_{n1} & B_{n2} & \cdots & B_{np}
\end{pmatrix}
\]

Then as explained above, the \( j \)th column is of the form

\[
\begin{pmatrix}
A_{11} & A_{12} & \cdots & A_{1n} \\
A_{21} & A_{22} & \cdots & A_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
A_{m1} & A_{m2} & \cdots & A_{mn}
\end{pmatrix}
\begin{pmatrix}
B_{1j} \\
B_{2j} \\
\vdots \\
B_{nj}
\end{pmatrix}
\]

which is a \( m \times 1 \) matrix or column vector which equals

\[
\begin{pmatrix}
A_{11} \\
A_{21} \\
\vdots \\
A_{m1}
\end{pmatrix}B_{1j} + \begin{pmatrix}
A_{12} \\
A_{22} \\
\vdots \\
A_{m2}
\end{pmatrix}B_{2j} + \cdots + \begin{pmatrix}
A_{1n} \\
A_{2n} \\
\vdots \\
A_{mn}
\end{pmatrix}B_{nj}.
\]

The second entry of this \( m \times 1 \) matrix is

\[
A_{21}B_{1j} + A_{22}B_{2j} + \cdots + A_{2n}B_{nj} = \sum_{k=1}^{m} A_{2k}B_{kj}.
\]

Similarly, the \( i \)th entry of this \( m \times 1 \) matrix is

\[
A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{in}B_{nj} = \sum_{k=1}^{m} A_{ik}B_{kj}.
\]

This shows the following definition for matrix multiplication in terms of the \( ij \)th entries of the product coincides with Definition [12.2.4].

**Definition 12.3.1** Let \( A = (A_{ij}) \) be an \( m \times n \) matrix and let \( B = (B_{ij}) \) be an \( n \times p \) matrix. Then \( AB \) is an \( m \times p \) matrix and

\[
(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}.
\]

(12.12)
Another way to write this is

\[(AB)_{ij} = (A_{i1} A_{i2} \cdots A_{in}) \begin{pmatrix} B_{1j} \\ B_{2j} \\ \vdots \\ B_{nj} \end{pmatrix} \]

Note that to get \((AB)_{ij}\) you involve the \(i^{th}\) row of \(A\) and the \(j^{th}\) column of \(B\).

**Example 12.3.2** Multiply if possible

\[
\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \end{pmatrix}
\]

First check to see if this is possible. It is of the form \((3 \times 2) (2 \times 3)\) and since the inside numbers match, the two matrices are conformable and it is possible to do the multiplication. The result should be a \(3 \times 3\) matrix. The answer is of the form

\[
\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 \\ 7 \\ 6 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 3 \\ 6 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix}
\]

where the commas separate the columns in the resulting product. Thus the above product equals

\[
\begin{pmatrix} 16 & 15 & 5 \\ 13 & 15 & 5 \\ 46 & 42 & 14 \end{pmatrix}
\]

a \(3 \times 3\) matrix as desired. In terms of the \(ij^{th}\) entries and the above definition, the entry in the third row and second column of the product should equal

\[
\sum_j a_{3k} b_{kj} = a_{31} b_{12} + a_{32} b_{22}
\]

\[= 2 \times 3 + 6 \times 6 = 42.\]

You should try a few more such examples to verify the above definition in terms of the \(ij^{th}\) entries works for other entries.

**Example 12.3.3** Multiply if possible

\[
\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \end{pmatrix}
\]

This is not possible because it is of the form \((3 \times 2) (3 \times 3)\) and the middle numbers don’t match. In other words the two matrices are not conformable in the indicated order.

**Example 12.3.4** Multiply if possible

\[
\begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}
\]

This is possible because in this case it is of the form \((3 \times 3) (3 \times 2)\) and the middle numbers do match so the matrices are conformable. When the multiplication is done it equals

\[
\begin{pmatrix} 13 & 13 \\ 29 & 32 \\ 0 & 0 \end{pmatrix}
\]

Check this and be sure you come up with the same answer.
Example 12.3.5 Multiply if possible \[
\begin{pmatrix}
1 & 2 \\
1 & 1
\end{pmatrix} \cdot 
\begin{pmatrix}
1 & 2 & 1 & 0 \\
1 & 2 & 1 & 0
\end{pmatrix}.
\]

In this case you are trying to do \((3 \times 1) (1 \times 4)\). The inside numbers match so you can do it. Verify 
\[
\begin{pmatrix}
1 \\
2 \\
1
\end{pmatrix} \cdot 
\begin{pmatrix}
1 & 2 & 1 & 0 \\
1 & 2 & 1 & 0
\end{pmatrix} = 
\begin{pmatrix}
2 & 4 & 2 & 0 \\
1 & 2 & 1 & 0
\end{pmatrix}
\]

12.4 Directed Graphs

Consider the following graph illustrated in the picture.

![Directed Graph Diagram]

There are three locations in this graph, labelled 1, 2, and 3. The directed lines represent a way of going from one location to another. Thus there is one way to go from location 1 to location 1. There is one way to go from location 1 to location 3. It is not possible to go from location 2 to location 3 although it is possible to go from location 3 to location 2. Let’s refer to moving along one of these directed lines as a step. The following \(3 \times 3\) matrix is a numerical way of writing the above graph. This is sometimes called a digraph.

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

Thus \(a_{ij}\), the entry in the \(i^{th}\) row and \(j^{th}\) column represents the number of ways to go from location \(i\) to location \(j\) in one step.

**Problem:** Find the number of ways to go from \(i\) to \(j\) using exactly \(k\) steps.

Denote the answer to the above problem by \(a_{ij}^k\). We don’t know what it is right now unless \(k = 1\) when it equals \(a_{ij}\) described above. However, if we did know what it was, we could find \(a_{ij}^{k+1}\) as follows.

\[a_{ij}^{k+1} = \sum_r a_{ir}^k a_{rj}\]

This is because if you go from \(i\) to \(j\) in \(k + 1\) steps, you first go from \(i\) to \(r\) in \(k\) steps and then for each of these ways there are \(a_{rj}\) ways to go from there to \(j\). Thus \(a_{ir}^k a_{rj}\) gives the number of ways to go from \(i\) to \(j\) in \(k + 1\) steps such that the \(k^{th}\) step leaves you at location \(r\). Adding these gives the above sum. Now you recognize this as the \(ij^{th}\) entry of the product of two matrices. Thus

\[a_{ij}^2 = \sum_r a_{ir} a_{rj}\]
and so forth. From the above definition of matrix multiplication, this shows that if $A$ is the matrix associated with the directed graph as above, then $a^k_{ij}$ is just the $ij^{th}$ entry of $A^k$ where $A^k$ is just what you would think it should be, $A$ multiplied by itself $k$ times.

Thus in the above example, to find the number of ways of going from 1 to 3 in two steps you would take that matrix and multiply it by itself and then take the entry in the first row and third column. Thus

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

and you see there is exactly one way to go from 1 to 3 in two steps. You can easily see this is true from looking at the graph also. Note there are three ways to go from 1 to 1 in 2 steps. Can you find them from the graph? What would you do if you wanted to consider 5 steps?

\[
\begin{pmatrix}
1 & 1 & 1 \\
1 & 0 & 0 \\
1 & 1 & 0
\end{pmatrix}^5 = \begin{pmatrix}
28 & 19 & 13 \\
13 & 9 & 6 \\
19 & 13 & 9
\end{pmatrix}
\]

There are 19 ways to go from 1 to 2 in five steps. Do you think you could list them all by looking at the graph? I don’t think you could do it without wasting a lot of time.

Of course there is nothing sacred about having only three locations. Everything works just as well with any number of locations. In general if you have $n$ locations, you would need to use a $n \times n$ matrix.

Example 12.4.1 Consider the following directed graph.

Write the matrix which is associated with this directed graph and find the number of ways to go from 2 to 4 in three steps.

Here you need to use a $4 \times 4$ matrix. The one you need is

\[
\begin{pmatrix}
0 & 1 & 1 & 0 \\
1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 \\
0 & 1 & 0 & 1
\end{pmatrix}
\]
Then to find the answer, you just need to multiply this matrix by itself three times and look at the entry in the second row and fourth column.

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

There is exactly one way to go from 2 to 4 in three steps.

How many ways would there be of going from 2 to 4 in five steps?

\[
\begin{pmatrix}
0 & 1 & 1 & 0 \\
1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 \\
0 & 1 & 0 & 1
\end{pmatrix}^5 = \begin{pmatrix}
5 & 9 & 5 & 4 \\
5 & 4 & 1 & 3 \\
9 & 10 & 4 & 6 \\
4 & 6 & 3 & 3
\end{pmatrix}
\]

There are three ways. Note there are 10 ways to go from 3 to 2 in five steps.

This is an interesting application of the concept of the \(ij\)th entry of the product matrices.

### 12.5 Properties Of Matrix Multiplication

As pointed out above, sometimes it is possible to multiply matrices in one order but not in the other order. What if it makes sense to multiply them in either order? Will the two products be equal then?

**Example 12.5.1** Compare \(\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\) \(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\) and \(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\) \(\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\).

The first product is

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

The second product is

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

You see these are not equal. Again you cannot conclude that \(AB = BA\) for matrix multiplication even when multiplication is defined in both orders. However, there are some properties which do hold.

**Proposition 12.5.2** If all multiplications and additions make sense, the following hold for matrices, \(A, B, C\) and \(a, b\) scalars.

\[
A(ab + bc) = a(AB) + b(AC)
\]  
(12.13)

\[
(B + C)A = BA + CA
\]  
(12.14)

\[
A(BC) = (AB)C
\]  
(12.15)
Proof: Using Definition [12.3.1],

\[(A(aB + bC))_{ij} = \sum_{k} A_{ik} (aB + bC)_{kj}\]

\[= \sum_{k} A_{ik} (aB_{kj} + bC_{kj})\]

\[= a \sum_{k} A_{ik} B_{kj} + b \sum_{k} A_{ik} C_{kj}\]

\[= a (AB)_{ij} + b (AC)_{ij}\]

\[= (a (AB) + b (AC))_{ij} .\]

Thus \(A (B + C) = AB + AC\) as claimed. Formula [12.14] is entirely similar.

Formula [12.15] is the associative law of multiplication. Using Definition [12.3.1],

\[(A(BC))_{ij} = \sum_{k} A_{ik} (BC)_{kj}\]

\[= \sum_{k} A_{ik} \sum_{l} B_{kl} C_{lj}\]

\[= \sum_{l} (AB)_{il} C_{lj}\]

\[= ((AB)C)_{ij} .\]

This proves [12.15].

12.6 The Transpose

Another important operation on matrices is that of taking the transpose. The following example shows what is meant by this operation, denoted by placing a \(T\) as an exponent on the matrix.

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

What happened? The first column became the first row and the second column became the second row. Thus the \(3 \times 2\) matrix became a \(2 \times 3\) matrix. The number 3 was in the second row and the first column and it ended up in the first row and second column. Here is the definition.

Definition 12.6.1 Let \(A\) be an \(m \times n\) matrix. Then \(A^T\) denotes the \(n \times m\) matrix which is defined as follows.

\[(A^T)_{ij} = A_{ji}\]

Example 12.6.2

\[
\begin{pmatrix}
1 & 2 & -6 \\
3 & 5 & 4
\end{pmatrix}^T = 
\begin{pmatrix}
1 & 3 \\
2 & 5 \\
-6 & 4
\end{pmatrix}
\]

The transpose of a matrix has the following important properties.

Lemma 12.6.3 Let \(A\) be an \(m \times n\) matrix and let \(B\) be a \(n \times p\) matrix. Then

\[(AB)^T = B^T A^T\] \hspace{1cm} (12.16)

and if \(\alpha\) and \(\beta\) are scalars,

\[(\alpha A + \beta B)^T = \alpha A^T + \beta B^T\] \hspace{1cm} (12.17)
**Proof:** From the definition,

\[
\left((AB)^T\right)_{ij} = (AB)_{ji} = \sum_k A_{jk}B_{ki} = \sum_k \left(B^T\right)_{ik} (A^T)_{kj} = \left(B^T A^T\right)_{ij}
\]

The proof of Formula 12.17 is left as an exercise and this proves the lemma.

**Definition 12.6.4** An \( n \times n \) matrix, \( A \) is said to be **symmetric** if \( A = A^T \). It is said to be **skew symmetric** if \( A = -A^T \).

**Example 12.6.5** Let

\[
A = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 5 & -3 \\ 3 & -3 & 7 \end{pmatrix}.
\]

Then \( A \) is symmetric.

**Example 12.6.6** Let

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

Then \( A \) is skew symmetric.

### 12.7 The Identity And Inverses

There is a special matrix called \( I \) and referred to as the identity matrix. It is always a square matrix, meaning the number of rows equals the number of columns and it has the property that there are ones down the main diagonal and zeroes elsewhere. Here are some identity matrices of various sizes:

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

The first is the \( 1 \times 1 \) identity matrix, the second is the \( 2 \times 2 \) identity matrix, the third is the \( 3 \times 3 \) identity matrix, and the fourth is the \( 4 \times 4 \) identity matrix. By extension, you can likely see what the \( n \times n \) identity matrix would be. It is so important that there is a special symbol to denote the \( ij \)th entry of the identity matrix

\[ I_{ij} = \delta_{ij} \]

where \( \delta_{ij} \) is the **Kroneker symbol** defined by

\[
\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}
\]

It is called the **identity matrix** because it is a **multiplicative identity** in the following sense.
Lemma 12.7.1 Suppose \( A \) is an \( m \times n \) matrix and \( I_n \) is the \( n \times n \) identity matrix. Then \( AI_n = A \). If \( I_m \) is the \( m \times m \) identity matrix, it also follows that \( I_mA = A \).

Proof:

\[
(AI_n)_{ij} = \sum_k A_{ik} \delta_{kj} = A_{ij}
\]

and so \( AI_n = A \). The other case is left as an exercise for you.

Definition 12.7.2 An \( n \times n \) matrix, \( A \) has an inverse, \( A^{-1} \) if and only if \( AA^{-1} = A^{-1}A = I \). Such a matrix is called invertible.

It is very important to observe that the inverse of a matrix, if it exists, is unique. Another way to think of this is that if it acts like the inverse, then it is the inverse.

Theorem 12.7.3 Suppose \( A^{-1} \) exists and \( AB = BA = I \). Then \( B = A^{-1} \).

Proof:

\[
A^{-1} = A^{-1}I = A^{-1} (AB) = (A^{-1}A) B = IB = B.
\]

Unlike ordinary multiplication of numbers, it can happen that \( A \neq 0 \) but \( A \) may fail to have an inverse. This is illustrated in the following example.

Example 12.7.4 Let \( A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \). Does \( A \) have an inverse?

One might think \( A \) would have an inverse because it does not equal zero. However,

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

and if \( A^{-1} \) existed, this could not happen because you could write

\[
\begin{pmatrix} 0 \\ 0 \end{pmatrix} = A^{-1} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = A^{-1} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = (A^{-1}A) \begin{pmatrix} -1 \\ 1 \end{pmatrix} = I \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix},
\]

a contradiction. Thus the answer is that \( A \) does not have an inverse.

Example 12.7.5 Let \( A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \). Show \( \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \) is the inverse of \( A \).

To check this, multiply

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

and

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

showing that this matrix is indeed the inverse of \( A \).
12.8 Finding The Inverse Of A Matrix

In the last example, how would you find $A^{-1}$? You wish to find a matrix,
\[
\begin{pmatrix}
x & z \\
y & w
\end{pmatrix}
\]
such that
\[
\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}
\begin{pmatrix} x & z \\ y & w \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.
\]
This requires the solution of the systems of equations,
\[
x + y = 1, \quad x + 2y = 0
\]
and
\[
z + w = 0, \quad z + 2w = 1.
\]
Writing the augmented matrix for these two systems gives
\[
\begin{pmatrix}
1 & 1 & | & 1 \\
1 & 2 & | & 0
\end{pmatrix} \quad (12.18)
\]
for the first system and
\[
\begin{pmatrix}
1 & 1 & | & 0 \\
1 & 2 & | & 1
\end{pmatrix} \quad (12.19)
\]
for the second. Let’s solve the first system. Take $(−1)$ times the first row and add to the second to get
\[
\begin{pmatrix}
1 & 1 & | & 1 \\
0 & 1 & | & −1
\end{pmatrix}
\]
Now take $(−1)$ times the second row and add to the first to get
\[
\begin{pmatrix}
1 & 0 & | & 2 \\
0 & 1 & | & −1
\end{pmatrix}.
\]
Putting in the variables, this says $x = 2$ and $y = −1$.

Now solve the second system, \(12.19\) to find $z$ and $w$. Take $(−1)$ times the first row and add to the second to get
\[
\begin{pmatrix}
1 & 1 & | & 0 \\
0 & 1 & | & 1
\end{pmatrix}.
\]
Now take $(−1)$ times the second row and add to the first to get
\[
\begin{pmatrix}
1 & 0 & | & −1 \\
0 & 1 & | & 1
\end{pmatrix}.
\]
Putting in the variables, this says $z = −1$ and $w = 1$. Therefore, the inverse is
\[
\begin{pmatrix}
2 & −1 \\
−1 & 1
\end{pmatrix}.
\]

Didn’t the above seem rather repetitive? Note that exactly the same row operations were used in both systems. In each case, the end result was something of the form $(I|v)$ where $I$ is the identity and $v$ gave a column of the inverse. In the above, \(x y\), the first column of the inverse was obtained first and then the second column \(z w\).

To simplify this procedure, you could have written
\[
\begin{pmatrix}
1 & 1 & | & 1 & 0 \\
1 & 2 & | & 0 & 1
\end{pmatrix}
and row reduced till you obtained

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

and read off the inverse as the 2 × 2 matrix on the right side.

This is the reason for the following simple procedure for finding the inverse of a matrix. This procedure is called the Gauss-Jordan procedure.

**Procedure 12.8.1** Suppose \( A \) is an \( n \times n \) matrix. To find \( A^{-1} \) if it exists, form the augmented \( n \times 2n \) matrix,

\[
(A|I)
\]

and then, if possible do row operations until you obtain an \( n \times 2n \) matrix of the form

\[
(I|B).
\]

When this has been done, \( B = A^{-1} \). If it is impossible to row reduce to a matrix of the form \((I|B)\), then \( A \) has no inverse.

**Example 12.8.2** Let \( A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 3 & 1 & -1 \end{pmatrix} \). Find \( A^{-1} \) if it exists.

Set up the augmented matrix, \((A|I)\)

\[
\begin{pmatrix}
1 & 2 & 2 | & 1 & 0 & 0 \\
1 & 0 & 2 | & 0 & 1 & 0 \\
3 & 1 & -1 | & 0 & 0 & 1
\end{pmatrix}
\]

Next take \((-1)\) times the first row and add to the second followed by \((-3)\) times the first row added to the last. This yields

\[
\begin{pmatrix}
1 & 2 & 2 | & 1 & 0 & 0 \\
0 & -2 & 0 | & -1 & 1 & 0 \\
0 & -5 & -7 | & -3 & 0 & 1
\end{pmatrix}
\]

Then take 5 times the second row and add to -2 times the last row.

\[
\begin{pmatrix}
1 & 2 & 2 | & 1 & 0 & 0 \\
0 & -10 & 0 | & -5 & 5 & 0 \\
0 & 0 & 14 | & 1 & 5 & -2
\end{pmatrix}
\]

Next take the last row and add to \((-7)\) times the top row. This yields

\[
\begin{pmatrix}
-7 & -14 & 0 | & -6 & 5 & -2 \\
0 & -10 & 0 | & -5 & 5 & 0 \\
0 & 0 & 14 | & 1 & 5 & -2
\end{pmatrix}
\]

Now take \((-7/5)\) times the second row and add to the top.

\[
\begin{pmatrix}
-7 & 0 & 0 | & 1 & -2 & -2 \\
0 & -10 & 0 | & -5 & 5 & 0 \\
0 & 0 & 14 | & 1 & 5 & -2
\end{pmatrix}
\]
Finally divide the top row by -7, the second row by -10 and the bottom row by 14 which yields
\[
\begin{pmatrix}
1 & 0 & 0 & -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\
0 & 1 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\
0 & 0 & 1 & \frac{1}{14} & \frac{5}{14} & -\frac{1}{7}
\end{pmatrix}
\]

Therefore, the inverse is
\[
\begin{pmatrix}
-\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\
\frac{1}{2} & -\frac{1}{2} & 0 \\
\frac{1}{14} & \frac{5}{14} & -\frac{1}{7}
\end{pmatrix}
\]

**Example 12.8.3** Let \( A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 4 \end{pmatrix} \). Find \( A^{-1} \) if it exists.

Write the augmented matrix, \((A|I)\)
\[
\begin{pmatrix}
1 & 2 & 2 & 1 & 0 & 0 \\
1 & 0 & 2 & 0 & 1 & 0 \\
2 & 2 & 4 & 0 & 0 & 1
\end{pmatrix}
\]

and proceed to do row operations attempting to obtain \((I|A^{-1})\). Take \((-1)\) times the top row and add to the second. Then take \((-2)\) times the top row and add to the bottom.
\[
\begin{pmatrix}
1 & 2 & 2 & 1 & 0 & 0 \\
0 & -2 & 0 & -1 & 1 & 0 \\
0 & -2 & 0 & -2 & 0 & 1
\end{pmatrix}
\]

Next add \((-1)\) times the second row to the bottom row.
\[
\begin{pmatrix}
1 & 2 & 2 & 1 & 0 & 0 \\
0 & -2 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & -1 & -1 & 1
\end{pmatrix}
\]

At this point, you can see there will be no inverse because you have obtained a row of zeros in the left half of the augmented matrix, \((A|I)\). Thus there will be no way to obtain \( I \) on the left.

**Example 12.8.4** Let \( A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \). Find \( A^{-1} \) if it exists.

Form the augmented matrix,
\[
\begin{pmatrix}
1 & 0 & 1 & 1 & 0 & 0 \\
1 & -1 & 1 & 0 & 1 & 0 \\
1 & 1 & -1 & 0 & 0 & 1
\end{pmatrix}
\]

Now do row operations until the \( n \times n \) matrix on the left becomes the identity matrix.
This yields after some computations,
\[
\begin{pmatrix}
1 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\
0 & 1 & 0 & 1 & -1 & 0 \\
0 & 0 & 1 & 1 & -\frac{1}{2} & -\frac{1}{2}
\end{pmatrix}
\]
and so the inverse of \( A \) is the matrix on the right,

\[
\begin{pmatrix}
0 & \frac{1}{2} & \frac{1}{2} \\
1 & -1 & 0 \\
1 & -\frac{1}{2} & -\frac{1}{2}
\end{pmatrix}.
\]

Checking the answer is easy. Just multiply the matrices and see if it works.

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

Always check your answer because if you are like some of us, you will usually have made a mistake.

**Example 12.8.5** In this example, it is shown how to use the inverse of a matrix to find the solution to a system of equations. Consider the following system of equations. Use the inverse of a suitable matrix to give the solutions to this system.

\[
\begin{align*}
x + z &= 1 \\
x - y + z &= 3 \\
x + y - z &= 2
\end{align*}
\]

The system of equations can be written in terms of matrices as

\[
\begin{pmatrix}
1 & 0 & 1 \\
1 & -1 & 1 \\
1 & 1 & -1
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix} = 
\begin{pmatrix}
1 \\
3 \\
2
\end{pmatrix}.
\] (12.21)

More simply, this is of the form \( Ax = b \). Suppose you find the inverse of the matrix, \( A^{-1} \). Then you could multiply both sides of this equation by \( A^{-1} \) to obtain

\[x = (A^{-1}A)x = A^{-1}(Ax) = A^{-1}b.\]

This gives the solution as \( x = A^{-1}b \). Note that once you have found the inverse, you can easily get the solution for different right hand sides without any effort. It is always just \( A^{-1}b \). In the given example, the inverse of the matrix is

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

This was shown in Example 12.8.4. Therefore, from what was just explained the solution to the given system is

\[
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix} = 
\begin{pmatrix}
0 & \frac{1}{2} & \frac{1}{2} \\
1 & -1 & 0 \\
1 & -\frac{1}{2} & -\frac{1}{2}
\end{pmatrix}
\begin{pmatrix}
1 \\
3 \\
2
\end{pmatrix} = 
\begin{pmatrix}
\frac{5}{2} \\
-2 \\
-\frac{3}{2}
\end{pmatrix}.
\]

What if the right side of (12.21) had been

\[
\begin{pmatrix}
0 \\
1 \\
3
\end{pmatrix}?
\]
What would be the solution to
\[
\begin{pmatrix}
1 & 0 & 1 \\
1 & -1 & 1 \\
1 & 1 & -1
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= 
\begin{pmatrix}
0 \\
1 \\
3
\end{pmatrix}.
\]

By the above discussion, it is just
\[
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= 
\begin{pmatrix}
0 & \frac{1}{2} & \frac{1}{2} \\
1 & -1 & 0 \\
1 & -\frac{1}{2} & -\frac{1}{2}
\end{pmatrix}
\begin{pmatrix}
0 \\
1 \\
3
\end{pmatrix}
= 
\begin{pmatrix}
2 \\
-1 \\
-2
\end{pmatrix}.
\]

This illustrates why once you have found the inverse of a given matrix, you can use it to solve many different systems easily.

### 12.9 Exercises

1. Here are some matrices:

   \[
   A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 7 \end{pmatrix}, B = \begin{pmatrix} 3 & -1 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\
   C = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}, D = \begin{pmatrix} -1 & 2 \\ 2 & -3 \end{pmatrix}, E = \begin{pmatrix} 2 \\ 3 \end{pmatrix}.
   \]

   Find if possible \(-3A, 3B - A, AC, CB, AE, EA\). If it is not possible explain why.

2. Here are some matrices:

   \[
   A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\
   C = \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 & 1 \\ 4 & -3 \end{pmatrix}, E = \begin{pmatrix} 1 \end{pmatrix}.
   \]

   Find if possible \(-3A, 3B - A, AC, CA, AE, EA, BE, DE, DE\). If it is not possible explain why.

3. Here are some matrices:

   \[
   A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\
   C = \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 & 1 \\ 4 & -3 \end{pmatrix}, E = \begin{pmatrix} 1 \end{pmatrix}.
   \]

   Find if possible \(-3A^T, 3B - A^T, AC, CA, AE, E^TB, BE, DE, EE^T, E^T E\). If it is not possible explain why.

4. Here are some matrices:

   \[
   A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\
   C = \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 \\ 4 \end{pmatrix}, E = \begin{pmatrix} 1 \end{pmatrix}.
   \]

   Find the following if possible and explain why it is not possible if this is the case. 
   \(AD, DA, D^TB, D^TBE, E^TD, DE^T\).
5. Let \( A = \begin{pmatrix} 1 & 1 \\ -2 & -1 \\ 1 & 2 \end{pmatrix} \), \( B = \begin{pmatrix} 1 & -1 & -2 \\ 2 & 1 & -2 \\ -3 & -1 & 0 \end{pmatrix} \), and \( C = \begin{pmatrix} 1 & 1 & -3 \\ -1 & 2 & 0 \\ -3 & -1 & 0 \end{pmatrix} \). Find if possible.

   (a) \( AB \)
   (b) \( BA \)
   (c) \( AC \)
   (d) \( CA \)
   (e) \( CB \)
   (f) \( BC \)

6. Suppose \( A \) and \( B \) are square matrices of the same size. Which of the following are correct?

   (a) \((A - B)^2 = A^2 - 2AB + B^2\)
   (b) \((AB)^2 = A^2B^2\)
   (c) \((A + B)^2 = A^2 + 2AB + B^2\)
   (d) \((A + B)^3 = A^3 + 3A^2B + 3AB^2 + B^3\)
   (e) \(A^2B^2 = A(AB)B\)
   (f) \(A + B)^3 = A^3 + 3A^2B + 3AB^2 + B^3\)
   (g) \((A + B)(A - B) = A^2 - B^2\)

7. Let \( A = \begin{pmatrix} -1 & -1 \\ 3 & 3 \end{pmatrix} \). Find all \( 2 \times 2 \) matrices, \( B \) such that \( AB = 0 \).

8. Let \( x = (-1, -1, 1) \) and \( y = (0, 1, 2) \). Find \( x^T y \) and \( xy^T \) if possible.

9. Let \( A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \), \( B = \begin{pmatrix} 1 & 2 \\ 3 & k \end{pmatrix} \). Is it possible to choose \( k \) such that \( AB = BA \)? If so, what should \( k \) equal?

10. Let \( A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \), \( B = \begin{pmatrix} 1 & 2 \\ 1 & k \end{pmatrix} \). Is it possible to choose \( k \) such that \( AB = BA \)? If so, what should \( k \) equal?

11. In [12.1]-[12.8] describe \(-A\) and \(0\).

12. Let \( A \) be an \( n \times n \) matrix. Show \( A \) equals the sum of a symmetric and a skew symmetric matrix. **Hint:** Show that \( \frac{1}{2} (A^T + A) \) is symmetric and then consider using this as one of the matrices.

13. Show every skew symmetric matrix has all zeros down the main diagonal. The main diagonal consists of every entry of the matrix which is of the form \( a_{ii} \). It runs from the upper left down to the lower right.

14. Using only the properties [12.1]-[12.8] show \(-A\) is unique.

15. Using only the properties [12.1]-[12.8] show \(0\) is unique.

16. Using only the properties [12.1]-[12.8] show \(0A = 0\). Here the \(0\) on the left is the scalar \(0\) and the \(0\) on the right is the zero for \(m \times n\) matrices.
17. Using only the properties [12.1]-[12.8] and previous problems show \((-1)A = -A\).
18. Prove [12.17].
19. Prove that \(I_mA = A\) where \(A\) is an \(m \times n\) matrix.
20. Give an example of matrices, \(A, B, C\) such that \(B \neq C, A \neq 0\), and yet \(AB = AC\).
21. Suppose \(AB = AC\) and \(A\) is an invertible \(n \times n\) matrix. Does it follow that \(B = C\)? Explain why or why not. What if \(A\) were a non invertible \(n \times n\) matrix?
22. Find your own examples:
   (a) \(2 \times 2\) matrices, \(A\) and \(B\) such that \(A \neq 0, B \neq 0\) with \(AB \neq BA\).
   (b) \(2 \times 2\) matrices, \(A\) and \(B\) such that \(A \neq 0, B \neq 0\), but \(AB = 0\).
   (c) \(2 \times 2\) matrices, \(A, D,\) and \(C\) such that \(A \neq 0, C \neq D\), but \(AC = AD\).
23. Explain why if \(AB = AC\) and \(A^{-1}\) exists, then \(B = C\).
24. Give an example of a matrix, \(A\) such that \(A^2 = I\) and yet \(A \neq I\) and \(A \neq -I\).
25. Give an example of matrices, \(A, B\) such that neither \(A\) nor \(B\) equals zero and yet \(AB = 0\).
26. Give another example other than the one given in this section of two square matrices, \(A\) and \(B\) such that \(AB \neq BA\).
27. Let \(A = \begin{pmatrix} 2 & 1 \\ -1 & 3 \end{pmatrix}\).
   Find \(A^{-1}\) if possible. If \(A^{-1}\) does not exist, determine why.
28. Let \(A = \begin{pmatrix} 0 & 1 \\ 5 & 3 \end{pmatrix}\).
   Find \(A^{-1}\) if possible. If \(A^{-1}\) does not exist, determine why.
29. Let \(A = \begin{pmatrix} 2 & 1 \\ 3 & 0 \end{pmatrix}\).
   Find \(A^{-1}\) if possible. If \(A^{-1}\) does not exist, determine why.
30. Let \(A = \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}\).
   Find \(A^{-1}\) if possible. If \(A^{-1}\) does not exist, determine why.
31. Let \(A\) be a \(2 \times 2\) matrix which has an inverse. Say \(A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\). Find a formula for \(A^{-1}\) in terms of \(a, b, c, d\).
32. Let \(A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 1 & 0 & 2 \end{pmatrix}\).
   Find \(A^{-1}\) if possible. If \(A^{-1}\) does not exist, determine why.
33. Let
\[ A = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}. \]
Find \( A^{-1} \) if possible. If \( A^{-1} \) does not exist, determine why.

34. Let
\[ A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 4 & 5 & 10 \end{pmatrix}. \]
Find \( A^{-1} \) if possible. If \( A^{-1} \) does not exist, determine why.

35. Let
\[ A = \begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 1 & 2 & 0 \\ 2 & 1 & -3 & 2 \\ 1 & 2 & 1 & 2 \end{pmatrix}. \]
Find \( A^{-1} \) if possible. If \( A^{-1} \) does not exist, determine why.

36. Write
\[ \begin{pmatrix} x_1 - x_2 + 2x_3 \\ 2x_3 + x_1 \\ 3x_3 \\ 3x_4 + 3x_2 + x_1 \end{pmatrix} \]
in the form \( A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \) where \( A \) is an appropriate matrix.

37. Write
\[ \begin{pmatrix} x_1 + 3x_2 + 2x_3 \\ 2x_3 + x_1 \\ 6x_3 \\ x_4 + 3x_2 + x_1 \end{pmatrix} \]
in the form \( A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \) where \( A \) is an appropriate matrix.

38. Write
\[ \begin{pmatrix} x_1 + x_2 + x_3 \\ 2x_3 + x_1 + x_2 \\ x_3 - x_1 \\ 3x_4 + x_1 \end{pmatrix} \]
in the form \( A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \) where \( A \) is an appropriate matrix.

39. Using the inverse of the matrix, find the solution to the systems
\[ \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \\ -2 \end{pmatrix}. \]
Now give the solution in terms of \( a, b, \) and \( c \) to
\[ \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}. \]
40. Using the inverse of the matrix, find the solution to the systems
\[
\begin{pmatrix}
1 & 0 & 3 \\
2 & 3 & 4 \\
1 & 0 & 2
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= \begin{pmatrix}
1 \\
2 \\
3
\end{pmatrix},
\begin{pmatrix}
1 & 0 & 3 \\
2 & 3 & 4 \\
1 & 0 & 2
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= \begin{pmatrix}
2 \\
1 \\
0
\end{pmatrix}.
\]

Now give the solution in terms of \(a, b,\) and \(c\) to
\[
\begin{pmatrix}
1 & 0 & 3 \\
2 & 3 & 4 \\
1 & 0 & 2
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= \begin{pmatrix}
a \\
b \\
c
\end{pmatrix}.
\]

41. Using the inverse of the matrix, find the solution to the system
\[
\begin{pmatrix}
-1 & \frac{1}{2} & \frac{1}{2} \\
3 & \frac{1}{2} & -\frac{1}{2} \\
-1 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= \begin{pmatrix}
a \\
b \\
c
\end{pmatrix}.
\]

42. Show that if \(A\) is an \(n \times n\) invertible matrix and \(x\) is a \(n \times 1\) matrix such that \(Ax = b\) for \(b\) an \(n \times 1\) matrix, then \(x = A^{-1}b\).

43. Prove that if \(A^{-1}\) exists and \(Ax = 0\) then \(x = 0\).

44. Show that if \(A^{-1}\) exists for an \(n \times n\) matrix, then it is unique. That is, if \(BA = I\) and \(AB = I,\) then \(B = A^{-1}\).

45. Show that if \(A\) is an invertible \(n \times n\) matrix, then so is \(A^T\) and \((A^T)^{-1} = (A^{-1})^T\).

46. Show \((AB)^{-1} = B^{-1}A^{-1}\) by verifying that \(AB \left(B^{-1}A^{-1}\right) = I\) and \(B^{-1}A^{-1} \left(AB\right) = I\). \textbf{Hint:} Use Problem 44.

47. Show that \((ABC)^{-1} = C^{-1}B^{-1}A^{-1}\) by verifying that \((ABC) \left(C^{-1}B^{-1}A^{-1}\right) = I\) and \((C^{-1}B^{-1}A^{-1}) \left(ABC\right) = I\). \textbf{Hint:} Use Problem 44.

48. If \(A\) is invertible, show \((A^T)^{-1} = (A^{-1})^T\). \textbf{Hint:} Use Problem 44.

49. If \(A\) is invertible, show \((A^2)^{-1} = (A^{-1})^2\). \textbf{Hint:} Use Problem 44.

50. If \(A\) is invertible, show \((A^{-1})^{-1} = A\). \textbf{Hint:} Use Problem 44.

12.10 Linear Transformations

In calculus, you have to study functions which take vectors in \(\mathbb{R}^n\) and produce other vectors in \(\mathbb{R}^m\). The easiest functions which do this are the linear transformations. If you want the stuff in calculus to make any sense, you should first study the easy case and from there go to the harder nonlinear material found in calculus. As explained above, an \(m \times n\) matrix can be used to transform vectors in \(\mathbb{R}^n\) to vectors in \(\mathbb{R}^m\) through the use of matrix multiplication. I will show below that the linear transformations are exactly those functions which can be obtained by multiplication by a matrix.
Example 12.10.1 Consider the matrix, \( \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \). Think of it as a function which takes vectors in \( \mathbb{F}^3 \) and makes them into vectors in \( \mathbb{F}^2 \) as follows. For \( \begin{pmatrix} x \\ y \\ z \end{pmatrix} \) a vector in \( \mathbb{F}^3 \), multiply on the left by the given matrix to obtain the vector in \( \mathbb{F}^2 \). Here are some numerical examples.

\[
\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 5 \\ -3 \end{pmatrix},
\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 10 \\ 5 \\ -3 \end{pmatrix} = \begin{pmatrix} 20 \\ 14 \end{pmatrix}.
\]

More generally,

\[
\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + 2y \\ 2x + y \end{pmatrix}.
\]

The idea is to define a function which takes vectors in \( \mathbb{F}^3 \) and delivers new vectors in \( \mathbb{F}^2 \).

This is an example of something called a linear transformation.

Definition 12.10.2 Let \( T : \mathbb{F}^n \rightarrow \mathbb{F}^m \) be a function. Thus for each \( x \in \mathbb{F}^n \), \( Tx \in \mathbb{F}^m \). Then \( T \) is a linear transformation if whenever \( \alpha, \beta \) are scalars and \( x_1 \) and \( x_2 \) are vectors in \( \mathbb{F}^n \),

\[
T(\alpha x_1 + \beta x_2) = \alpha Tx_1 + \beta Tx_2.
\]

In words, linear transformations distribute across + and allow you to factor out scalars. At this point, recall the properties of matrix multiplication. The pertinent property is [12.14 on Page 235]. Recall it states that for \( a \) and \( b \) scalars,

\[
A(aB + bC) = aAB + bAC.
\]

In particular, for \( A \) an \( m \times n \) matrix and \( B \) and \( C \), \( n \times 1 \) matrices (column vectors) the above formula holds which is nothing more than the statement that matrix multiplication gives an example of a linear transformation.

Definition 12.10.3 A linear transformation is called one to one (often written as \( 1-1 \)) if it never takes two different vectors to the same vector. Thus \( T \) is one to one if whenever \( x \neq y \)

\[
Tx \neq Ty.
\]

Equivalently, if \( T(x) = T(y) \), then \( x = y \).

In the case that a linear transformation comes from matrix multiplication, it is common usage to refer to the matrix as a one to one matrix when the linear transformation it determines is one to one.

Definition 12.10.4 A linear transformation mapping \( \mathbb{F}^n \) to \( \mathbb{F}^m \) is called onto if whenever \( y \in \mathbb{F}^m \) there exists \( x \in \mathbb{F}^n \) such that \( T(x) = y \).
Thus \( T \) is onto if everything in \( \mathbb{F}^m \) gets hit. In the case that a linear transformation comes from matrix multiplication, it is common to refer to the matrix as onto when the linear transformation it determines is onto. Also it is common usage to write \( T \mathbb{F}^n \), \( T(\mathbb{F}^n) \), or \( \text{Im}(T) \) as the set of vectors of \( \mathbb{F}^m \) which are of the form \( T\mathbf{x} \) for some \( \mathbf{x} \in \mathbb{F}^n \). In the case that \( T \) is obtained from multiplication by an \( m \times n \) matrix, \( A \), it is standard to simply write \( A(\mathbb{F}^n) \mathbb{F}^n \), or \( \text{Im}(A) \) to denote those vectors in \( \mathbb{F}^m \) which are obtained in the form \( A\mathbf{x} \) for some \( \mathbf{x} \in \mathbb{F}^n \).

### 12.11 Constructing The Matrix Of A Linear Transformation

It turns out that if \( T \) is any linear transformation which maps \( \mathbb{F}^n \) to \( \mathbb{F}^m \), there is always an \( m \times n \) matrix, \( A \) with the property that

\[
A\mathbf{x} = T\mathbf{x}
\]  

(12.22)

for all \( \mathbf{x} \in \mathbb{F}^n \). Here is why. Suppose \( T : \mathbb{F}^n \to \mathbb{F}^m \) is a linear transformation and you want to find the matrix defined by this linear transformation as described in [12.22]. Then if \( \mathbf{x} \in \mathbb{F}^n \) it follows

\[
\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i
\]

where \( \mathbf{e}_i \) is the vector which has zeros in every slot but the \( i^{th} \) and a 1 in this slot. Then since \( T \) is linear,

\[
T\mathbf{x} = \sum_{i=1}^{n} x_i T(\mathbf{e}_i)
\]

= \[
\begin{pmatrix}
T(\mathbf{e}_1) & \cdots & T(\mathbf{e}_n)
\end{pmatrix}
\begin{pmatrix}
x_1 \\
\vdots \\
x_n
\end{pmatrix}
\]

≡ \[
A
\begin{pmatrix}
x_1 \\
\vdots \\
x_n
\end{pmatrix}
\]

and so you see that the matrix desired is obtained from letting the \( i^{th} \) column equal \( T(\mathbf{e}_i) \). We state this as the following theorem.

**Theorem 12.11.1** Let \( T \) be a linear transformation from \( \mathbb{F}^n \) to \( \mathbb{F}^m \). Then the matrix, \( A \) satisfying [12.22] is given by

\[
\begin{pmatrix}
T(\mathbf{e}_1) & \cdots & T(\mathbf{e}_n)
\end{pmatrix}
\]

where \( T\mathbf{e}_i \) is the \( i^{th} \) column of \( A \).

### 12.12 Rotations in \( \mathbb{R}^2 \)

Sometimes you need to find a matrix which represents a given linear transformation which is described in geometrical terms. The idea is to produce a matrix which you can
multiply a vector by to get the same thing as some geometrical description. A good example of this is the problem of rotation of vectors. For example, I may want a matrix which will rotate every vector through an angle of 45 degrees. Such a rotation would achieve something like the following if applied to each vector corresponding to points on the picture which is standing upright.

More generally, consider the problem of rotating through an angle of \( \theta \).

**Example 12.12.1** Determine the matrix which represents the linear transformation defined by rotating every vector through an angle of \( \theta \).

Let \( \mathbf{e}_1 \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix} \) and \( \mathbf{e}_2 \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix} \). These identify the geometric vectors which point along the positive \( x \) axis and positive \( y \) axis as shown.

From the above, you only need to find \( T\mathbf{e}_1 \) and \( T\mathbf{e}_2 \), the first being the first column of the desired matrix, \( A \) and the second being the second column. From the definition of the \( \cos, \sin \) the coordinates of \( T(\mathbf{e}_1) \) are as shown in the picture. The coordinates of \( T(\mathbf{e}_2) \) also follow from simple trigonometry. Thus

\[
T\mathbf{e}_1 = \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}, \quad T\mathbf{e}_2 = \begin{pmatrix} -\sin \theta \\ \cos \theta \end{pmatrix}.
\]

Therefore, from Theorem 12.11.1,

\[
A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}
\]

**Example 12.12.2** Find the matrix of the linear transformation which is obtained by first rotating all vectors through an angle of \( \phi \) and then through an angle \( \theta \). Thus you want the linear transformation which rotates all angles through an angle of \( \theta + \phi \).
12.12. ROTATIONS IN $\mathbb{R}^2$

Let $T_{\theta+\phi}$ denote the linear transformation which rotates every vector through an angle of $\theta + \phi$. Then to get $T_{\theta+\phi}$, you could first do $T_{\phi}$ and then do $T_{\theta}$ where $T_{\phi}$ is the linear transformation which rotates through an angle of $\phi$ and $T_{\theta}$ is the linear transformation which rotates through an angle of $\theta$. Denoting the corresponding matrices by $A_{\theta+\phi}$, $A_{\phi}$, and $A_{\theta}$, you must have for every $x$

$$A_{\theta+\phi}x = T_{\theta+\phi}x = T_{\theta}T_{\phi}x = A_{\theta}A_{\phi}x.$$  

Consequently, you must have

$$A_{\theta+\phi} = \begin{pmatrix} \cos (\theta + \phi) & -\sin (\theta + \phi) \\ \sin (\theta + \phi) & \cos (\theta + \phi) \end{pmatrix} = A_{\theta}A_{\phi}$$

$$= \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}.$$  

You know how to multiply matrices. Do so to the pair on the right. This yields

$$\begin{pmatrix} \cos (\theta + \phi) & -\sin (\theta + \phi) \\ \sin (\theta + \phi) & \cos (\theta + \phi) \end{pmatrix} = \begin{pmatrix} \cos \theta \cos \phi - \sin \theta \sin \phi & -\cos \theta \sin \phi - \sin \theta \cos \phi \\ \sin \theta \cos \phi + \cos \theta \sin \phi & \cos \theta \cos \phi - \sin \theta \sin \phi \end{pmatrix}.$$  

Don’t these look familiar? They are the usual trig. identities for the sum of two angles derived here using linear algebra concepts.

You do not have to stop with two dimensions. You can consider rotations and other geometric concepts in any number of dimensions. This is one of the major advantages of linear algebra. You can break down a difficult geometrical procedure into small steps, each corresponding to multiplication by an appropriate matrix. Then by multiplying the matrices, you can obtain a single matrix which can give you numerical information on the results of applying the given sequence of simple procedures. That which you could never visualize can still be understood to the extent of finding exact numerical answers. Another example follows.

**Example 12.12.3** Find the matrix of the linear transformation which is obtained by first rotating all vectors through an angle of $\pi/6$ and then reflecting through the $x$ axis.

As shown in Example 12.12.2, the matrix of the transformation which involves rotating through an angle of $\pi/6$ is

$$\begin{pmatrix} \cos (\pi/6) & -\sin (\pi/6) \\ \sin (\pi/6) & \cos (\pi/6) \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \sqrt{3} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \sqrt{3} \end{pmatrix}.$$  

The matrix for the transformation which reflects all vectors through the $x$ axis is

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

Therefore, the matrix of the linear transformation which first rotates through $\pi/6$ and then reflects through the $x$ axis is

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{2} \sqrt{3} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \sqrt{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \sqrt{3} & -\frac{1}{2} \sqrt{3} \\ -\frac{1}{2} & -\frac{1}{2} \sqrt{3} \end{pmatrix}.$$
12.13 Exercises

1. Study the definition of a linear transformation. State it from memory.

2. Show from the definition and from memory that for the map \( T : \mathbb{R}^n \rightarrow \mathbb{R}^m \) defined by \( T(x) = Ax \) where \( A \) is an \( m \times n \) matrix and \( x \) is an \( m \times 1 \) column vector, this is a linear transformation.

3. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/3 \).

4. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/4 \).

5. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( -\pi/3 \).

6. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( 2\pi/3 \).

7. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/12 \). **Hint:** Note that \( \pi/12 = \pi/3 - \pi/4 \).

8. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( 2\pi/3 \) and then reflects across the \( x \) axis.

9. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/3 \) and then reflects across the \( x \) axis.

10. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/4 \) and then reflects across the \( x \) axis.

11. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( \pi/6 \) and then reflects across the \( x \) axis followed by a reflection across the \( y \) axis.

12. Find the matrix for the linear transformation which reflects every vector in \( \mathbb{R}^2 \) across the \( x \) axis and then rotates every vector through an angle of \( \pi/4 \).

13. Find the matrix for the linear transformation which reflects every vector in \( \mathbb{R}^2 \) across the \( y \) axis and then rotates every vector through an angle of \( \pi/4 \).

14. Find the matrix for the linear transformation which reflects every vector in \( \mathbb{R}^2 \) across the \( x \) axis and then rotates every vector through an angle of \( \pi/6 \).

15. Find the matrix for the linear transformation which reflects every vector in \( \mathbb{R}^2 \) across the \( y \) axis and then rotates every vector through an angle of \( \pi/6 \).

16. Find the matrix for the linear transformation which rotates every vector in \( \mathbb{R}^2 \) through an angle of \( 5\pi/12 \). **Hint:** Note that \( 5\pi/12 = 2\pi/3 - \pi/4 \).
13.1 Basic Techniques And Properties

13.1.1 Cofactors And $2 \times 2$ Determinants

Let $A$ be an $n \times n$ matrix. The determinant of $A$, denoted as $\det(A)$, is a number. If the matrix is a $2\times2$ matrix, this number is very easy to find.

**Definition 13.1.1** Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then

$$\det(A) \equiv ad - cb.$$  

The determinant is also often denoted by enclosing the matrix with two vertical lines.

Thus

$$\det\left( \begin{array}{cc} a & b \\ c & d \end{array} \right) = \left| \begin{array}{cc} a & b \\ c & d \end{array} \right|.$$  

**Example 13.1.2** Find $\det\left( \begin{array}{cc} 2 & 4 \\ -1 & 6 \end{array} \right)$.

From the definition this is just $(2)(6) - (-1)(4) = 16$.

Having defined what is meant by the determinant of a $2 \times 2$ matrix, what about a $3 \times 3$ matrix?

**Definition 13.1.3** Suppose $A$ is a $3 \times 3$ matrix. The $ij$th minor, denoted as $\text{minor}(A)_{ij}$, is the determinant of the $2 \times 2$ matrix which results from deleting the $i$th row and the $j$th column.

**Example 13.1.4** Consider the matrix, 

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 3 & 2 \\ 3 & 2 & 1 \end{pmatrix}.$$  

The $(1,2)$ minor is the determinant of the $2 \times 2$ matrix which results when you delete the first row and the second column. This minor is therefore

$$\det\left( \begin{array}{cc} 4 & 2 \\ 3 & 1 \end{array} \right) = -2.$$  

The $(2,3)$ minor is the determinant of the $2 \times 2$ matrix which results when you delete the second row and the third column. This minor is therefore

$$\det\left( \begin{array}{cc} 1 & 2 \\ 3 & 2 \end{array} \right) = -4.$$
**Definition 13.1.5** Suppose $A$ is a $3 \times 3$ matrix. The $ij^{th}$ **cofactor** is defined to be $(-1)^{i+j} \times (ij^{th \text{ minor}})$. In words, you multiply $(-1)^{i+j}$ times the $ij^{th}$ minor to get the $ij^{th}$ cofactor. The cofactors of a matrix are so important that special notation is appropriate when referring to them. The $ij^{th}$ cofactor of a matrix, $A$ will be denoted by $\text{cof}(A)_{ij}$. It is also convenient to refer to the cofactor of an entry of a matrix as follows. For $a_{ij}$ an entry of the matrix, its cofactor is just $\text{cof}(A)_{ij}$. Thus the cofactor of the $ij^{th}$ entry is just the $ij^{th}$ cofactor.

**Example 13.1.6** Consider the matrix,

$$
A = \begin{pmatrix}
1 & 2 & 3 \\
4 & 3 & 2 \\
3 & 2 & 1
\end{pmatrix}.
$$

The $(1,2)$ minor is the determinant of the $2 \times 2$ matrix which results when you delete the first row and the second column. This minor is therefore

$$
\det \begin{pmatrix}
4 & 2 \\
3 & 1
\end{pmatrix} = -2.
$$

It follows

$$
\text{cof}(A)_{12} = (-1)^{1+2} \det \begin{pmatrix}
4 & 2 \\
3 & 1
\end{pmatrix} = (-1)^{1+2} (-2) = 2
$$

The $(2,3)$ minor is the determinant of the $2 \times 2$ matrix which results when you delete the second row and the third column. This minor is therefore

$$
\det \begin{pmatrix}
1 & 2 \\
3 & 2
\end{pmatrix} = -4.
$$

Therefore,

$$
\text{cof}(A)_{23} = (-1)^{2+3} \det \begin{pmatrix}
1 & 2 \\
3 & 2
\end{pmatrix} = (-1)^{2+3} (-4) = 4.
$$

Similarly,

$$
\text{cof}(A)_{22} = (-1)^{2+2} \det \begin{pmatrix}
1 & 3 \\
3 & 1
\end{pmatrix} = -8.
$$

**Definition 13.1.7** The determinant of a $3 \times 3$ matrix, $A$, is obtained by picking a row (column) and taking the product of each entry in that row (column) with its cofactor and adding these up. This process when applied to the $i^{th}$ row (column) is known as expanding the determinant along the $i^{th}$ row (column).

**Example 13.1.8** Find the determinant of

$$
A = \begin{pmatrix}
1 & 2 & 3 \\
4 & 3 & 2 \\
3 & 2 & 1
\end{pmatrix}.
$$

Here is how it is done by “expanding along the first column”.

$$
1(-1)^{1+1} \begin{vmatrix}
3 & 2 \\
2 & 1
\end{vmatrix} + 4(-1)^{2+1} \begin{vmatrix}
2 & 3 \\
2 & 1
\end{vmatrix} + 3(-1)^{3+1} \begin{vmatrix}
2 & 3 \\
3 & 2
\end{vmatrix} = 0.
$$

You see, we just followed the rule in the above definition. We took the 1 in the first column and multiplied it by its cofactor, the 4 in the first column and multiplied it by
its cofactor, and the 3 in the first column and multiplied it by its cofactor. Then we added these numbers together.

You could also expand the determinant along the second row as follows.

\[
\begin{vmatrix}
4(-1)^{2+1} & 2 & 3 \\
2 & 3 & 1
\end{vmatrix}
+ \begin{vmatrix}
3(-1)^{2+2} & 1 & 3 \\
3 & 1 & 1
\end{vmatrix}
+ 2(-1)^{2+3}
\begin{vmatrix}
1 & 2 \\
3 & 2
\end{vmatrix}
= 0.
\]

Observe this gives the same number. You should try expanding along other rows and columns. If you don’t make any mistakes, you will always get the same answer.

What about a 4 × 4 matrix? You know now how to find the determinant of a 3 × 3 matrix. The pattern is the same.

**Definition 13.1.9** Suppose \( A \) is a 4 × 4 matrix. The \( ij \)th **minor** is the determinant of the 3 × 3 matrix you obtain when you delete the \( i \)th row and the \( j \)th column. The \( ij \)th **cofactor**, \( \text{cof}(A)_{ij} \) is defined to be \((-1)^{i+j} \times (ij \)th minor\). In words, you multiply \((-1)^{i+j} \) times the \( ij \)th minor to get the \( ij \)th cofactor.

**Definition 13.1.10** The determinant of a 4 × 4 matrix, \( A \), is obtained by picking a row (column) and taking the product of each entry in that row (column) with its cofactor and adding these up. This process when applied to the \( i \)th row (column) is known as expanding the determinant along the \( i \)th row (column).

**Example 13.1.11** Find \( \det(A) \) where

\[
A = \begin{pmatrix}
1 & 2 & 3 & 4 \\
5 & 4 & 2 & 3 \\
1 & 3 & 4 & 5 \\
3 & 4 & 3 & 2
\end{pmatrix}
\]

As in the case of a 3 × 3 matrix, you can expand this along any row or column. Let’s pick the third column. \( \det(A) = \)

\[
3(-1)^{1+3} \begin{vmatrix}
5 & 4 & 3 \\
3 & 4 & 2
\end{vmatrix}
+ 2(-1)^{2+3} \begin{vmatrix}
1 & 2 & 4 \\
3 & 4 & 2
\end{vmatrix}
+ 4(-1)^{3+3} \begin{vmatrix}
1 & 2 & 4 \\
3 & 4 & 2
\end{vmatrix}
\]

Now you know how to expand each of these 3 × 3 matrices along a row or a column. If you do so, you will get \(-12\) assuming you make no mistakes. You could expand this matrix along any row or any column and assuming you make no mistakes, you will always get the same thing which is defined to be the determinant of the matrix, \( A \). This method of evaluating a determinant by expanding along a row or a column is called the method of **Laplace expansion**.

Note that each of the four terms above involves three terms consisting of determinants of 2 × 2 matrices and each of these will need 2 terms. Therefore, there will be \(4 \times 3 \times 2 = 24\) terms to evaluate in order to find the determinant using the method of Laplace expansion. Suppose now you have a 10 × 10 matrix and you follow the above pattern for evaluating determinants. By analogy to the above, there will be \(10! = 3,628,800\) terms involved in the evaluation of such a determinant by Laplace expansion along a row or column. This is a lot of terms.
In addition to the difficulties just discussed, you should regard the above claim that you always get the same answer by picking any row or column with considerable skepticism. It is incredible and not at all obvious. However, it requires a little effort to establish it. This is done in the section on the theory of the determinant.

**Definition 13.1.12** Let \( A = (a_{ij}) \) be an \( n \times n \) matrix and suppose the determinant of a \((n-1) \times (n-1)\) matrix has been defined. Then a new matrix called the **cofactor matrix**, \( \text{cof} \left( A \right) \) is defined by \( \text{cof} \left( A \right) = (c_{ij}) \) where to obtain \( c_{ij} \) delete the \( i^{th} \) row and the \( j^{th} \) column of \( A \), take the determinant of the \((n-1) \times (n-1)\) matrix which results, (This is called the \( ij^{th} \) **minor** of \( A \). ) and then multiply this number by \((-1)^{i+j} \). Thus \((-1)^{i+j} \times (the \ ij^{th} \ minor) \) equals the \( ij^{th} \) cofactor. To make the formulas easier to remember, \( \text{cof} \left( A \right)_{ij} \) will denote the \( ij^{th} \) entry of the cofactor matrix.

With this definition of the cofactor matrix, here is how to define the determinant of an \( n \times n \) matrix.

**Definition 13.1.13** Let \( A \) be an \( n \times n \) matrix where \( n \geq 2 \) and suppose the determinant of an \((n-1) \times (n-1)\) has been defined. Then

\[
\det \left( A \right) = \sum_{j=1}^{n} a_{ij} \text{cof} \left( A \right)_{ij} = \sum_{i=1}^{n} a_{ij} \text{cof} \left( A \right)_{ij} . \quad (13.1)
\]

The first formula consists of expanding the determinant along the \( i^{th} \) row and the second expands the determinant along the \( j^{th} \) column.

**Theorem 13.1.14** Expanding the \( n \times n \) matrix along any row or column always gives the same answer so the above definition is a good definition.

13.1.2 The Determinant Of A Triangular Matrix

Notwithstanding the difficulties involved in using the method of Laplace expansion, certain types of matrices are very easy to deal with.

**Definition 13.1.15** A matrix \( M \), is upper triangular if \( M_{ij} = 0 \) whenever \( i > j \). Thus such a matrix equals zero below the main diagonal, the entries of the form \( M_{ii} \), as shown.

\[
\begin{pmatrix}
* & * & \cdots & * \\
0 & * & \ddots & \vdots \\
\vdots & \ddots & \ddots & * \\
0 & \cdots & 0 & *
\end{pmatrix}
\]

A lower triangular matrix is defined similarly as a matrix for which all entries above the main diagonal are equal to zero.

You should verify the following using the above theorem on Laplace expansion.

**Corollary 13.1.16** Let \( M \) be an upper (lower) triangular matrix. Then \( \det \left( M \right) \) is obtained by taking the product of the entries on the main diagonal.

**Example 13.1.17** Let

\[
A = \begin{pmatrix}
1 & 2 & 3 & 77 \\
0 & 2 & 6 & 7 \\
0 & 0 & 3 & 3.7 \\
0 & 0 & 0 & -1
\end{pmatrix}
\]

Find \( \det \left( A \right) \).
From the above corollary, it suffices to take the product of the diagonal elements. Thus $\det (A) = 1 \times 2 \times 3 \times (-1) = -6$. Without using the corollary, you could expand along the first column. This gives

$$
\begin{vmatrix}
1 & 2 & 6 & 7 \\
0 & 0 & 3 & 33.7 \\
0 & 0 & -1 & \\
0 & 0 & -1 & \\
\end{vmatrix} + 0 (-1)^{2+1} \begin{vmatrix}
2 & 3 & 77 \\
0 & 0 & 33.7 \\
0 & 0 & -1 & \\
0 & 0 & 33.7 & \\
\end{vmatrix} + 0 (-1)^{4+1} \begin{vmatrix}
2 & 3 & 77 \\
0 & 0 & 33.7 & \\
0 & 0 & -1 & \\
0 & 0 & 33.7 & \\
\end{vmatrix}
$$

and the only nonzero term in the expansion is

$$
\begin{vmatrix}
1 & 2 & 6 & 7 \\
0 & 0 & 3 & 33.7 \\
0 & 0 & -1 & \\
\end{vmatrix}
$$

Now expand this along the first column to obtain

$$
1 \times 2 \times \begin{vmatrix}
3 & 33.7 \\
0 & -1 & \\
\end{vmatrix} + 0 (-1)^{2+1} \begin{vmatrix}
6 & 7 \\
0 & -1 & \\
\end{vmatrix} + 0 (-1)^{3+1} \begin{vmatrix}
6 & 7 \\
3 & 33.7 & \\
\end{vmatrix}
$$

$$
= 1 \times 2 \times \begin{vmatrix}
3 & 33.7 \\
0 & -1 & \\
\end{vmatrix}
$$

Next expand this last determinant along the first column to obtain the above equals

$$
1 \times 2 \times 3 \times (-1) = -6
$$

which is just the product of the entries down the main diagonal of the original matrix.

### 13.1.3 Properties Of Determinants

There are many properties satisfied by determinants. Some of these properties have to do with row operations. Recall the row operations.

**Definition 13.1.18** The row operations consist of the following

1. Switch two rows.
2. Multiply a row by a nonzero number.
3. Replace a row by a multiple of another row added to itself.

**Theorem 13.1.19** Let $A$ be an $n \times n$ matrix and let $A_1$ be a matrix which results from multiplying some row of $A$ by a scalar, $c$. Then $c \det (A) = \det (A_1)$.

**Example 13.1.20** Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $A_1 = \begin{pmatrix} 2 & 4 \\ 3 & 4 \end{pmatrix}$. Then $\det (A) = -2$, $\det (A_1) = -4$.

**Theorem 13.1.21** Let $A$ be an $n \times n$ matrix and let $A_1$ be a matrix which results from switching two rows of $A$. Then $\det (A) = -\det (A_1)$. Also, if one row of $A$ is a multiple of another row of $A$, then $\det (A) = 0$.

**Example 13.1.22** Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $A_1 = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}$. Then $\det (A) = -2$, $\det (A_1) = 2$. 

**Theorem 13.1.23** Let $A$ be an $n \times n$ matrix and let $A_1$ be a matrix which results from applying row operation 3. That is you replace some row by a multiple of another row added to itself. Then $\det(A) = \det(A_1)$.

**Example 13.1.24** Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and let $A_1 = \begin{pmatrix} 1 & 2 \\ 4 & 6 \end{pmatrix}$. Thus the second row of $A_1$ is one times the first row added to the second row. $\det(A) = -2$ and $\det(A_1) = -2$.

**Theorem 13.1.25** In Theorems 13.1.19 - 13.1.23 you can replace the word, “row” with the word “column”.

There are two other major properties of determinants which do not involve row operations.

**Theorem 13.1.26** Let $A$ and $B$ be two $n \times n$ matrices. Then

\[
\det(AB) = \det(A) \det(B).
\]

Also,

\[
\det(A) = \det(A^T).
\]

**Example 13.1.27** Compare $\det(AB)$ and $\det(A) \det(B)$ for

\[
A = \begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix}, B = \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix}.
\]

First

\[
AB = \begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 11 & 4 \\ -1 & -4 \end{pmatrix}
\]

and so \(\det(AB) = \det\begin{pmatrix} 11 & 4 \\ -1 & -4 \end{pmatrix} = -40\).

Now

\[
\det(A) = \det\begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix} = 8
\]

and

\[
\det(B) = \det\begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix} = -5.
\]

Thus $\det(A) \det(B) = 8 \times (-5) = -40$.

### 13.1.4 Finding Determinants Using Row Operations

Theorems 13.1.23 - 13.1.25 can be used to find determinants using row operations. As pointed out above, the method of Laplace expansion will not be practical for any matrix of large size. Here is an example in which all the row operations are used.

**Example 13.1.28** Find the determinant of the matrix,

\[
A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 1 & 2 & 3 \\ 4 & 5 & 4 & 3 \\ 2 & 2 & -4 & 5 \end{pmatrix}
\]
Replace the second row by \((-5)\) times the first row added to it. Then replace the third row by \((-4)\) times the first row added to it. Finally, replace the fourth row by \((-2)\) times the first row added to it. This yields the matrix,

\[
B = \begin{pmatrix}
1 & 2 & 3 & 4 \\
0 & -9 & -13 & -17 \\
0 & -3 & -8 & -13 \\
0 & -2 & -10 & -3
\end{pmatrix}
\]

and from Theorem [13.1.23], it has the same determinant as \(A\). Now using other row operations, \(\det(B) = \left(\frac{-1}{3}\right) \det(C)\) where

\[
C = \begin{pmatrix}
1 & 2 & 3 & 4 \\
0 & 0 & 11 & 22 \\
0 & -3 & -8 & -13 \\
0 & 6 & 30 & 9
\end{pmatrix}.
\]

The second row was replaced by \((-3)\) times the third row added to the second row. By Theorem [13.1.23] this didn’t change the value of the determinant. Then the last row was multiplied by \((-3)\). By Theorem [13.1.19] the resulting matrix has a determinant which is \((-3)\) times the determinant of the unmultiplied matrix. Therefore, we multiplied by \(-1/3\) to retain the correct value. Now replace the last row with \(2\) times the third added to it. This does not change the value of the determinant by Theorem [13.1.23]. Finally switch the third and second rows. This causes the determinant to be multiplied by \((-1)\). Thus \(\det(C) = -\det(D)\) where

\[
D = \begin{pmatrix}
1 & 2 & 3 & 4 \\
0 & -3 & -8 & -13 \\
0 & 0 & 11 & 22 \\
0 & 0 & 14 & -17
\end{pmatrix}
\]

You could do more row operations or you could note that this can be easily expanded along the first column followed by expanding the \(3 \times 3\) matrix which results along its first column. Thus

\[
\det(D) = 1(-3) \begin{vmatrix} 11 & 22 \\ 14 & -17 \end{vmatrix} = 1485
\]

and so \(\det(C) = -1485\) and \(\det(A) = \det(B) = \left(\frac{-1}{3}\right)(-1485) = 495\).

**Example 13.1.29** Find the determinant of the matrix

\[
\begin{pmatrix}
1 & 2 & 3 & 2 \\
1 & -3 & 2 & 1 \\
2 & 1 & 2 & 5 \\
3 & -4 & 1 & 2
\end{pmatrix}
\]

Replace the second row by \((-1)\) times the first row added to it. Next take \(-2\) times the first row and add to the third and finally take \(-3\) times the first row and add to the last row. This yields

\[
\begin{pmatrix}
1 & 2 & 3 & 2 \\
0 & -5 & -1 & -1 \\
0 & -3 & -4 & 1 \\
0 & -10 & -8 & -4
\end{pmatrix}.
\]

By Theorem [13.1.23] this matrix has the same determinant as the original matrix. Remember you can work with the columns also. Take \(-5\) times the last column and add
to the second column. This yields
\[
\begin{pmatrix}
1 & -8 & 3 & 2 \\
0 & 0 & -1 & -1 \\
0 & -8 & -4 & 1 \\
0 & 10 & -8 & -4 \\
\end{pmatrix}
\]
By Theorem 13.1.25 this matrix has the same determinant as the original matrix. Now take \((-1)\) times the third row and add to the top row. This gives.
\[
\begin{pmatrix}
1 & 0 & 7 & 1 \\
0 & 0 & -1 & -1 \\
0 & -8 & -4 & 1 \\
0 & 10 & -8 & -4 \\
\end{pmatrix}
\]
which by Theorem 13.1.23 has the same determinant as the original matrix. Lets expand it now along the first column. This yields the following for the determinant of the original matrix.
\[
\det
\begin{pmatrix}
0 & -1 & -1 \\
-8 & -4 & 1 \\
10 & -8 & -4 \\
\end{pmatrix}
\]
which equals
\[
8 \det
\begin{pmatrix}
-1 & -1 \\
-8 & -4 \\
\end{pmatrix}
+ 10 \det
\begin{pmatrix}
-1 & -1 \\
-4 & 1 \\
\end{pmatrix}
= -82
\]
We suggest you do not try to be fancy in using row operations. That is, stick mostly to the one which replaces a row or column with a multiple of another row or column added to it. Also note there is no way to check your answer other than working the problem more than one way. To be sure you have gotten it right you must do this.

13.2 Applications

13.2.1 A Formula For The Inverse

The definition of the determinant in terms of Laplace expansion along a row or column also provides a way to give a formula for the inverse of a matrix. Recall the definition of the inverse of a matrix in Definition 12.7.2 on Page 238. Also recall the definition of the cofactor matrix given in Definition 13.1.12 on Page 256. This cofactor matrix was just the matrix which results from replacing the \(ij\)th entry of the matrix with the \(ij\)th cofactor.

The following theorem says that to find the inverse, take the transpose of the cofactor matrix and divide by the determinant. The transpose of the cofactor matrix is called the adjugate or sometimes the classical adjoint of the matrix \(A\). In other words, \(A^{-1}\) is equal to one divided by the determinant of \(A\) times the adjugate matrix of \(A\). This is what the following theorem says with more precision.

**Theorem 13.2.1** \(A^{-1}\) exists if and only if \(\det(A) \neq 0\). If \(\det(A) \neq 0\), then \(A^{-1} = (a^{-1}_{ij})\) where

\[
a^{-1}_{ij} = \det(A)^{-1} \text{cof}(A)_{ji},
\]

for \(\text{cof}(A)_{ij}\) the \(ij\)th cofactor of \(A\).

**Example 13.2.2** Find the inverse of the matrix,

\[
A = \begin{pmatrix}
1 & 2 & 3 \\
3 & 0 & 1 \\
1 & 2 & 1 \\
\end{pmatrix}
\]
13.2. APPLICATIONS

First find the determinant of this matrix. Using Theorems 13.1.23 - 13.1.25 on Page 258, the determinant of this matrix equals the determinant of the matrix,

\[
\begin{pmatrix}
1 & 2 & 3 \\
0 & -6 & -8 \\
0 & 0 & -2
\end{pmatrix}
\]

which equals 12. The cofactor matrix of \( A \) is

\[
\begin{pmatrix}
-2 & -2 & 6 \\
4 & -2 & 0 \\
2 & 8 & -6
\end{pmatrix}
\]

Each entry of \( A \) was replaced by its cofactor. Therefore, from the above theorem, the inverse of \( A \) should equal

\[
\frac{1}{12} \begin{pmatrix} -2 & -2 & 6 \\ 4 & -2 & 0 \\ 2 & 8 & -6 \end{pmatrix}^T = \begin{pmatrix} -\frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{6} & \frac{2}{3} \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{pmatrix}
\]

Does it work? You should check to see if it does. When the matrices are multiplied

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

and so it is correct.

**Example 13.2.3** Find the inverse of the matrix,

\[
A = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{6} & \frac{1}{3} & -\frac{5}{6} \\ -\frac{5}{6} & \frac{2}{3} & -\frac{1}{2} \end{pmatrix}
\]

First find its determinant. This determinant is \( \frac{1}{5} \). The inverse is therefore equal to

\[
\begin{pmatrix} \frac{1}{3} & -\frac{1}{2} & -\frac{1}{6} & -\frac{1}{2} & -\frac{1}{6} & \frac{1}{2} \\ \frac{2}{3} & -\frac{1}{2} & -\frac{5}{6} & -\frac{1}{2} & -\frac{5}{6} & \frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{5}{6} & -\frac{1}{2} & -\frac{5}{6} & \frac{1}{2} \\ \frac{1}{3} & -\frac{1}{2} & -\frac{1}{6} & -\frac{1}{2} & -\frac{1}{6} & \frac{1}{2} \end{pmatrix}^T
\]
Expanding all the $2 \times 2$ determinants this yields
\[
6 \begin{pmatrix}
\frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\
\frac{1}{3} & \frac{1}{6} & -\frac{1}{3} \\
-\frac{1}{6} & \frac{1}{8} & \frac{1}{6}
\end{pmatrix}^T = \begin{pmatrix}
1 & 2 & -1 \\
2 & 1 & 1 \\
1 & -2 & 1
\end{pmatrix}
\]
Always check your work.

\[
\begin{pmatrix}
1 & 2 & -1 \\
2 & 1 & 1 \\
1 & -2 & 1
\end{pmatrix} \begin{pmatrix}
\frac{1}{2} & 0 & \frac{1}{2} \\
-\frac{1}{6} & \frac{1}{3} & -\frac{1}{2} \\
-\frac{5}{6} & \frac{2}{3} & -\frac{1}{2}
\end{pmatrix} = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]
and so we got it right. If the result of multiplying these matrices had been something other than the identity matrix, you would know there was an error. When this happens, you need to search for the mistake if you are interested in getting the right answer. A common mistake is to forget to take the transpose of the cofactor matrix.

**Proof of Theorem 13.2.1** From the definition of the determinant in terms of expansion along a column, and letting $(a_{ir}) = A$, if $\det(A) \neq 0$,
\[
\sum_{i=1}^{n} a_{ir} \cof(A)_{ir} \det(A)^{-1} = \det(A) \det(A)^{-1} = 1.
\]
Now consider
\[
\sum_{i=1}^{n} a_{ir} \cof(A)_{rk} \det(A)^{-1}
\]
when $k \neq r$. Replace the $k^{th}$ column with the $r^{th}$ column to obtain a matrix, $B_k$ whose determinant equals zero by Theorem 13.1.21. However, expanding this matrix, $B_k$ along the $k^{th}$ column yields
\[
0 = \det(B_k) \det(A)^{-1} = \sum_{i=1}^{n} a_{ir} \cof(A)_{rk} \det(A)^{-1}
\]
Summarizing,
\[
\sum_{i=1}^{n} a_{ir} \cof(A)_{ik} \det(A)^{-1} = \delta_{rk} \equiv \begin{cases} 1 & \text{if } r = k \\ 0 & \text{if } r \neq k \end{cases}
\]
Now
\[
\sum_{i=1}^{n} a_{ir} \cof(A)_{ik} = \sum_{i=1}^{n} a_{ir} \cof(A)^T_{ki}
\]
which is the $kr^{th}$ entry of $\cof(A)^T A$. Therefore,
\[
\frac{\cof(A)^T}{\det(A)} A = I. \tag{13.2}
\]
Using the other formula in Definition 13.1.13, and similar reasoning,
\[
\sum_{j=1}^{n} a_{rj} \cof(A)_{kj} \det(A)^{-1} = \delta_{rk}
\]
Now
\[ \sum_{j=1}^{n} a_{rj} \cof (A)_{kj} = \sum_{j=1}^{n} a_{rj} \cof (A)^{T}_{jk} \]
which is the \( r_{k}^{th} \) entry of \( A \cof (A)^{T} \). Therefore,
\[
A \frac{\cof (A)^{T}}{\det (A)} = I,
\tag{13.3}
\]
and it follows from [13.2] and [13.3] that \( A^{-1} = (a_{ij}^{-1}) \), where
\[ a_{ij}^{-1} = \cof (A)_{ji} \det (A)^{-1}. \]
In other words,
\[ A^{-1} = \frac{\cof (A)^{T}}{\det (A)}. \]

Now suppose \( A^{-1} \) exists. Then by Theorem [13.1.26],
\[ 1 = \det (I) = \det (AA^{-1}) = \det (A) \det (A^{-1}) \]
so \( \det (A) \neq 0 \). This proves the theorem.

This way of finding inverses is especially useful in the case where it is desired to find the inverse of a matrix whose entries are functions.

**Example 13.2.4** Suppose
\[
A(t) = \begin{pmatrix}
e^{t} & 0 & 0 \\
0 & \cos t & \sin t \\
0 & -\sin t & \cos t
\end{pmatrix}
\]
Show that \( A(t)^{-1} \) exists and then find it.

First note \( \det (A(t)) = e^{t} \neq 0 \) so \( A(t)^{-1} \) exists. The cofactor matrix is
\[
C(t) = \begin{pmatrix}
1 & 0 & 0 \\
0 & e^{t} \cos t & e^{t} \sin t \\
0 & -e^{t} \sin t & e^{t} \cos t
\end{pmatrix}
\]
and so the inverse is
\[
\frac{1}{e^{t}} \begin{pmatrix}
1 & 0 & 0 \\
0 & e^{t} \cos t & e^{t} \sin t \\
0 & -e^{t} \sin t & e^{t} \cos t
\end{pmatrix}^{T} = \begin{pmatrix}
e^{-t} & 0 & 0 \\
0 & \cos t & -\sin t \\
0 & \sin t & \cos t
\end{pmatrix}.
\]

### 13.2.2 Cramer’s Rule

This formula for the inverse also implies a famous procedure known as Cramer’s rule. Cramer’s rule gives a formula for the solutions, \( x \), to a system of equations, \( Ax = y \) in the special case that \( A \) is a square matrix. Note this rule does not apply if you have a system of equations in which there is a different number of equations than variables.

In case you are solving a system of equations, \( Ax = y \) for \( x \), it follows that if \( A^{-1} \) exists,
\[
x = (A^{-1}A)x = A^{-1}(Ax) = A^{-1}y
\]
thus solving the system. Now in the case that $A^{-1}$ exists, there is a formula for $A^{-1}$ given above. Using this formula,

$$x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \frac{1}{\det(A)} \sum_{j=1}^{n} \text{cof}(A)_{ji} y_j,$$

By the formula for the expansion of a determinant along a column,

$$x_i = \frac{1}{\det(A)} \det \begin{pmatrix}
\ast & \cdots & y_1 & \cdots & \ast \\
\vdots & & \vdots & & \vdots \\
\ast & \cdots & y_n & \cdots & \ast
\end{pmatrix},$$

where here the $i^{th}$ column of $A$ is replaced with the column vector, $(y_1, \cdots, y_n)^T$, and the determinant of this modified matrix is taken and divided by $\det(A)$. This formula is known as Cramer’s rule.

**Procedure 13.2.5** Suppose $A$ is an $n \times n$ matrix and it is desired to solve the system

$$Ax = y, y = (y_1, \cdots, y_n)^T$$

for $x = (x_1, \cdots, x_n)^T$. Then Cramer’s rule says

$$x_i = \frac{\det A_i}{\det A},$$

where $A_i$ is obtained from $A$ by replacing the $i^{th}$ column of $A$ with the column $(y_1, \cdots, y_n)^T$.

**Example 13.2.6** Find $x, y$ if

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

From Cramer’s rule,

$$x = \frac{1}{2}$$

Now to find $y,

$$y = \frac{-1}{7}$$

and

$$z = \frac{11}{14}$$
You see the pattern. For large systems Cramer’s rule is less than useful if you want to find an answer. This is because to use it you must evaluate determinants. However, you have no practical way to evaluate determinants for large matrices other than row operations and if you are using row operations, you might just as well use them to solve the system to begin with. It will be a lot less trouble. Nevertheless, there are situations in which Cramer’s rule is useful.

**Example 13.2.7** Solve for \( z \) if

\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & e^t \cos t & e^t \sin t \\
0 & -e^t \sin t & e^t \cos t
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
= 
\begin{pmatrix}
1 \\
t \\
t^2
\end{pmatrix}
\]

You could do it by row operations but it might be easier in this case to use Cramer’s rule because the matrix of coefficients does not consist of numbers but of functions. Thus

\[
z = \frac{1}{\det}
\begin{vmatrix}
1 & 0 & 1 \\
0 & e^t \cos t & t \\
0 & -e^t \sin t & t^2
\end{vmatrix} = t \left((\cos t) t + \sin t \right) e^{-t}.
\]

You end up doing this sort of thing sometimes in ordinary differential equations in the method of variation of parameters.

### 13.3 Exercises

1. Find the determinants of the following matrices.
   - (a) \[
   \begin{pmatrix}
   1 & 2 & 3 \\
   3 & 2 & 2 \\
   0 & 9 & 8
   \end{pmatrix}
   \] (The answer is 31.)
   - (b) \[
   \begin{pmatrix}
   4 & 3 & 2 \\
   1 & 7 & 8 \\
   3 & -9 & 3
   \end{pmatrix}
   \] (The answer is 375.)
   - (c) \[
   \begin{pmatrix}
   1 & 2 & 3 & 2 \\
   1 & 3 & 2 & 3 \\
   4 & 1 & 5 & 0 \\
   1 & 2 & 1 & 2
   \end{pmatrix}
   \] (The answer is -2.)

2. Find the following determinant by expanding along the first row and second column.
   \[
   \begin{vmatrix}
   1 & 2 & 1 \\
   2 & 1 & 3 \\
   2 & 1 & 1
   \end{vmatrix}
   \]

3. Find the following determinant by expanding along the first column and third row.
   \[
   \begin{vmatrix}
   1 & 2 & 1 \\
   1 & 0 & 1 \\
   2 & 1 & 1
   \end{vmatrix}
   \]
4. Find the following determinant by expanding along the second row and first column.
\[
\begin{vmatrix}
1 & 2 & 1 \\
2 & 1 & 3 \\
2 & 1 & 1
\end{vmatrix}
\]

5. Compute the determinant by cofactor expansion. Pick the easiest row or column to use.
\[
\begin{vmatrix}
1 & 0 & 0 & 1 \\
2 & 1 & 1 & 0 \\
0 & 0 & 0 & 2 \\
2 & 1 & 3 & 1
\end{vmatrix}
\]

6. Find the determinant using row operations.
\[
\begin{vmatrix}
1 & 2 & 1 \\
2 & 3 & 2 \\
-4 & 1 & 2
\end{vmatrix}
\]

7. Find the determinant using row operations.
\[
\begin{vmatrix}
2 & 1 & 3 \\
2 & 4 & 2 \\
1 & 4 & -5
\end{vmatrix}
\]

8. Find the determinant using row operations.
\[
\begin{vmatrix}
1 & 2 & 1 & 2 \\
3 & 1 & -2 & 3 \\
-1 & 0 & 3 & 1 \\
2 & 3 & 2 & -2
\end{vmatrix}
\]

9. Find the determinant using row operations.
\[
\begin{vmatrix}
1 & 4 & 1 & 2 \\
3 & 2 & -2 & 3 \\
-1 & 0 & 3 & 3 \\
2 & 1 & 2 & -2
\end{vmatrix}
\]


11. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.
\[
\begin{pmatrix}
a & b \\
c & d
\end{pmatrix} \rightarrow \begin{pmatrix}
a & c \\
b & d
\end{pmatrix}
\]

12. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.
\[
\begin{pmatrix}
a & b \\
c & d
\end{pmatrix} \rightarrow \begin{pmatrix}
c & d \\
a & b
\end{pmatrix}
\]
13. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

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

14. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

\[
\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & b \\ 2c & 2d \end{pmatrix}
\]

15. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

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

16. Let \( A \) be an \( r \times r \) matrix and suppose there are \( r - 1 \) rows (columns) such that all rows (columns) are linear combinations of these \( r - 1 \) rows (columns). Show \( \det (A) = 0 \).

17. Show \( \det (aA) = a^n \det (A) \) where here \( A \) is an \( n \times n \) matrix and \( a \) is a scalar.

18. Prove by doing computations that \( \det (AB) = \det (A) \det (B) \) if \( A \) and \( B \) are \( 2 \times 2 \) matrices.

19. Illustrate with an example of \( 2 \times 2 \) matrices that the determinant of a product equals the product of the determinants.

20. Is it true that \( \det (A + B) = \det (A) + \det (B) \)? If this is so, explain why it is so and if it is not so, give a counter example.

21. An \( n \times n \) matrix is called nilpotent if for some positive integer, \( k \) it follows \( A^k = 0 \). If \( A \) is a nilpotent matrix and \( k \) is the smallest possible integer such that \( A^k = 0 \), what are the possible values of \( \det (A) \)?

22. A matrix is said to be orthogonal if \( A^T A = I \). Thus the inverse of an orthogonal matrix is just its transpose. What are the possible values of \( \det (A) \) if \( A \) is an orthogonal matrix?

23. Fill in the missing entries to make the matrix orthogonal as in Problem 22

\[
\begin{pmatrix}
\frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{\sqrt{12}}{6} \\
\frac{1}{\sqrt{2}} & - & - \\
- & \frac{\sqrt{2}}{3} & -
\end{pmatrix}
\]

24. Let \( A \) and \( B \) be two \( n \times n \) matrices. \( A \sim B \) (\( A \) is similar to \( B \)) means there exists an invertible matrix, \( S \) such that \( A = S^{-1}BS \). Show that if \( A \sim B \), then \( B \sim A \). Show also that \( A \sim A \) and that if \( A \sim B \) and \( B \sim C \), then \( A \sim C \).

25. In the context of Problem 24 show that if \( A \sim B \), then \( \det (A) = \det (B) \).
26. Two \( n \times n \) matrices, \( A \) and \( B \), are similar if \( B = S^{-1}AS \) for some invertible \( n \times n \) matrix, \( S \). Show that if two matrices are similar, they have the same characteristic polynomials. The characteristic polynomial of an \( n \times n \) matrix, \( M \) is the polynomial, \( \det (\lambda I - M) \).

27. Tell whether the statement is true or false.

(a) If \( A \) is a \( 3 \times 3 \) matrix with a zero determinant, then one column must be a multiple of some other column.

(b) If any two columns of a square matrix are equal, then the determinant of the matrix equals zero.

(c) For \( A \) and \( B \) two \( n \times n \) matrices, \( \det (A + B) = \det (A) + \det (B) \).

(d) For \( A \) an \( n \times n \) matrix, \( \det (3A) = 3 \det (A) \)

(e) If \( A^{-1} \) exists then \( \det (A^{-1}) = \det (A)^{-1} \).

(f) If \( B \) is obtained by multiplying a single row of \( A \) by 4 then \( \det (B) = 4 \det (A) \).

(g) For \( A \) an \( n \times n \) matrix, \( \det (-A) = (-1)^n \det (A) \).

(h) If \( A \) is a real \( n \times n \) matrix, then \( \det (A^T A) \geq 0 \).

(i) Cramer’s rule is useful for finding solutions to systems of linear equations in which there is an infinite set of solutions.

(j) If \( A^k = 0 \) for some positive integer, \( k \), then \( \det (A) = 0 \).

(k) If \( A \mathbf{x} = \mathbf{0} \) for some \( \mathbf{x} \neq \mathbf{0} \), then \( \det (A) = 0 \).

28. Use Cramer’s rule to find the solution to

\[
\begin{align*}
x + 2y &= 1 \\
2x - y &= 2
\end{align*}
\]

29. Use Cramer’s rule to find the solution to

\[
\begin{align*}
x + 2y + z &= 1 \\
2x - y - z &= 2 \\
x + z &= 1
\end{align*}
\]

30. Here is a matrix,

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

Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

31. Here is a matrix,

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

Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.
32. Here is a matrix,
\[
\begin{pmatrix}
1 & 3 & 3 \\
2 & 4 & 1 \\
0 & 1 & 1
\end{pmatrix}
\]
Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

33. Here is a matrix,
\[
\begin{pmatrix}
1 & 2 & 3 \\
0 & 2 & 1 \\
2 & 6 & 7
\end{pmatrix}
\]
Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

34. Here is a matrix,
\[
\begin{pmatrix}
1 & 0 & 3 \\
1 & 0 & 1 \\
3 & 1 & 0
\end{pmatrix}
\]
Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

35. Use the formula for the inverse in terms of the cofactor matrix to find if possible the inverses of the matrices
\[
\begin{pmatrix}
1 & 1 \\
1 & 2
\end{pmatrix},
\begin{pmatrix}
1 & 2 & 3 \\
0 & 2 & 1 \\
4 & 1 & 1
\end{pmatrix},
\begin{pmatrix}
1 & 2 & 1 \\
2 & 3 & 0 \\
0 & 1 & 2
\end{pmatrix}.
\]
If the inverse does not exist, explain why.

36. Here is a matrix,
\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & \cos t & -\sin t \\
0 & \sin t & \cos t
\end{pmatrix}
\]
Does there exist a value of \( t \) for which this matrix fails to have an inverse? Explain.

37. Here is a matrix,
\[
\begin{pmatrix}
1 & t & t^2 \\
0 & 1 & 2t \\
t & 0 & 2
\end{pmatrix}
\]
Does there exist a value of \( t \) for which this matrix fails to have an inverse? Explain.

38. Here is a matrix,
\[
\begin{pmatrix}
et & \cosh t & \sinh t \\
et & \sinh t & \cosh t \\
et & \cosh t & \sinh t
\end{pmatrix}
\]
Does there exist a value of \( t \) for which this matrix fails to have an inverse? Explain.

39. Show that if \( \det (A) \neq 0 \) for \( A \) an \( n \times n \) matrix, it follows that if \( Ax = 0 \), then \( x = 0 \).
40. Suppose $A, B$ are $n \times n$ matrices and that $AB = I$. Show that then $BA = I$.

**Hint:** You might do something like this: First explain why det $(A), \det(B)$ are both nonzero. Then $(AB)A = A$ and then show $BA(AB - I) = 0$. From this use what is given to conclude $A(BA - I) = 0$. Then use Problem 39.

41. Use the formula for the inverse in terms of the cofactor matrix to find the inverse of the matrix,

$$A = \begin{pmatrix} e^t & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & e^t \cos t - e^t \sin t & e^t \cos t + e^t \sin t \end{pmatrix}.$$ 

42. Find the inverse if it exists of the matrix,

$$\begin{pmatrix} e^t & \cos t & \sin t \\ e^t & -\sin t & \cos t \\ e^t & -\cos t & -\sin t \end{pmatrix}.$$ 

43. Here is a matrix,

$$\begin{pmatrix} e^t & e^{-t} \cos t & e^{-t} \sin t \\ e^t & -e^{-t} \cos t - e^{-t} \sin t & -e^{-t} \sin t + e^{-t} \cos t \\ e^t & 2e^{-t} \sin t & -2e^{-t} \cos t \end{pmatrix}.$$ 

Does there exist a value of $t$ for which this matrix fails to have an inverse? Explain.

44. Suppose $A$ is an upper triangular matrix. Show that $A^{-1}$ exists if and only if all elements of the main diagonal are non zero. Is it true that $A^{-1}$ will also be upper triangular? Explain. Is everything the same for lower triangular matrices?

45. If $A, B,$ and $C$ are each $n \times n$ matrices and $ABC$ is invertible, why are each of $A, B,$ and $C$ invertible.

### 13.4 Exercises With Answers

1. Find the following determinant by expanding along the second column.

$$\begin{vmatrix} 1 & 3 & 1 \\ 2 & 1 & 5 \\ 2 & 1 & 1 \end{vmatrix}$$

This is

$$3(-1)^{2+1} \begin{vmatrix} 2 & 5 \\ 2 & 1 \end{vmatrix} + 1(-1)^{3+1} \begin{vmatrix} 1 & 1 \\ 2 & 1 \end{vmatrix} + 1(-1)^{3+2} \begin{vmatrix} 1 & 1 \\ 2 & 5 \end{vmatrix} = 20.$$

2. Compute the determinant by cofactor expansion. Pick the easiest row or column to use.

$$\begin{vmatrix} 2 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 3 \\ 2 & 3 & 1 \end{vmatrix}$$

You ought to use the third row. This yields

$$3 \begin{vmatrix} 2 & 0 & 0 \\ 2 & 1 & 1 \\ 2 & 3 & 3 \end{vmatrix} = (3)(2) \begin{vmatrix} 1 & 1 \\ 3 & 3 \end{vmatrix} = 0.$$
3. Find the determinant using row and column operations.

\[
\begin{vmatrix}
5 & 4 & 3 & 2 \\
3 & 2 & 4 & 3 \\
-1 & 2 & 3 & 3 \\
2 & 1 & 2 & -2 \\
\end{vmatrix}
\]

Replace the first row by 5 times the third added to it and then replace the second by 3 times the third added to it and then the last by 2 times the third added to it. This yields

\[
\begin{vmatrix}
0 & 14 & 18 & 17 \\
0 & 8 & 13 & 12 \\
-1 & 2 & 3 & 3 \\
0 & 5 & 8 & 4 \\
\end{vmatrix}
\]

Now lets replace the third column by \(-1\) times the last column added to it.

\[
\begin{vmatrix}
0 & 14 & 1 & 17 \\
0 & 8 & 1 & 12 \\
-1 & 2 & 0 & 3 \\
0 & 5 & 4 & 4 \\
\end{vmatrix}
\]

Now replace the top row by \(-1\) times the second added to it and the bottom row by \(-4\) times the second added to it. This yields

\[
\begin{vmatrix}
0 & 6 & 0 & 5 \\
0 & 8 & 1 & 12 \\
-1 & 2 & 0 & 3 \\
0 & -27 & 0 & -44 \\
\end{vmatrix}
\]

This looks pretty good because it has a lot of zeros. Expand along the first column and next along the second,

\[
(-1) \begin{vmatrix} 6 & 0 & 5 \\ 8 & 1 & 12 \\ -27 & 0 & -44 \end{vmatrix} = (-1)(1) \begin{vmatrix} 6 & 5 \\ -27 & -44 \end{vmatrix} = 129.
\]

Alternatively, you could continue doing row and column operations. Switch the third and first row in 13.4 to obtain

\[
\begin{vmatrix}
-1 & 2 & 0 & 3 \\
0 & 8 & 1 & 12 \\
0 & 6 & 0 & 5 \\
0 & -27 & 0 & -44 \\
\end{vmatrix}
\]

Next take \(9/2\) times the third row and add to the bottom.

\[
\begin{vmatrix}
-1 & 2 & 0 & 3 \\
0 & 8 & 1 & 12 \\
0 & 6 & 0 & 5 \\
0 & 0 & -44 + (9/2) 5 \\
\end{vmatrix}
\]

Finally, take \(-6/8\) times the second row and add to the third.

\[
\begin{vmatrix}
-1 & 2 & 0 & 3 \\
0 & 8 & 1 & 12 \\
0 & 0 & -6/8 & 5 + (-6/8) (12) \\
0 & 0 & 0 & -44 + (9/2) 5 \\
\end{vmatrix}
\]
Therefore, since the matrix is now upper triangular, the determinant is
\[ -((-1)(8)(-6/8)(-44 + (9/2)5)) = 129. \]

4. An operation is done to get from the first matrix to the second. Identify what
was done and tell how it will affect the value of the determinant.
\[
\begin{pmatrix} a & b \\ c & d \end{pmatrix} \rightarrow \begin{pmatrix} a & c \\ b & d \end{pmatrix}
\]
This involved taking the transpose so the determinant of the new matrix is the
same as the determinant of the first matrix.

5. Show that for \( A \) a \( 2 \times 2 \) matrix \( \det (aA) = a^2 \det (A) \) where \( a \) is a scalar.
\[ a^2 \det (A) = a \det (A_1) \] where the first row of \( A \) is replaced by \( a \) times it to get
\( A_1 \). Then \( a \det (A_1) = A_2 \) where \( A_2 \) is obtained from \( A \) by multiplying both rows
by \( a \). In other words, \( A_2 = aA \). Thus the conclusion is established.

6. Use Cramer’s rule to find \( y \) in
\[
\begin{align*}
2x + 2y + z &= 3 \\
2x - y - z &= 2 \\
x + 2z &= 1
\end{align*}
\]
From Cramer’s rule,
\[
y = \frac{\begin{vmatrix} 2 & 3 & 1 \\ 2 & 2 & -1 \\ 1 & 1 & 2 \end{vmatrix}}{\begin{vmatrix} 2 & 2 & 1 \\ 2 & -1 & -1 \\ 1 & 0 & 2 \end{vmatrix}} = \frac{5}{13}
\]

7. Here is a matrix,
\[
\begin{pmatrix} e^t & e^{-t} e^t & e^{-t} \cos t \\
e^{-t} e^t & e^{-t} \cos t - e^{-t} \sin t & -e^{-t} \sin t + e^{-t} \cos t \\
e^{2t} & 2e^{-t} \sin t & -2e^{-t} \cos t \end{pmatrix}
\]
Does there exist a value of \( t \) for which this matrix fails to have an inverse? Explain.
\[
det \begin{pmatrix} e^t & e^{-t} e^t & e^{-t} \cos t \\ e^{-t} e^t & e^{-t} \cos t - e^{-t} \sin t & -e^{-t} \sin t + e^{-t} \cos t \\ e^{2t} & 2e^{-t} \sin t & -2e^{-t} \cos t \end{pmatrix} = 5e^t e^{2(-t)} \cos^2 t + 5e^t e^{2(-t)} \sin^2 t = 5e^{-t} \]
which is never equal to zero for any
value of \( t \) and so there is no value of \( t \) for which the matrix has no inverse.

8. Use the formula for the inverse in terms of the cofactor matrix to find if possible
the inverse of the matrix
\[
\begin{pmatrix} 1 & 2 & 3 \\ 0 & 6 & 1 \\ 4 & 1 & 1 \end{pmatrix}
\]
First you need to take the determinant
\[
det \begin{pmatrix} 1 & 2 & 3 \\ 0 & 6 & 1 \\ 4 & 1 & 1 \end{pmatrix} = -59 \]
and so the matrix has an inverse. Now you need to find the cofactor matrix.

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

\[
= \begin{pmatrix}
5 & 4 & -24 \\
1 & -11 & 7 \\
-16 & -1 & 6
\end{pmatrix}.
\]

Thus the inverse is

\[
\frac{1}{-59} \begin{pmatrix}
-1 & 4 & -24 \\
1 & -11 & 7 \\
-16 & -1 & 6
\end{pmatrix}^T
\]

\[
= \frac{1}{-59} \begin{pmatrix}
-5 & 1 & -16 \\
4 & -11 & -1 \\
-24 & 7 & 6
\end{pmatrix}.
\]

If you check this, it does work.
A.1 The Function $\text{sgn}_n$

It is easiest to give a different definition of the determinant which is clearly well defined and then prove the earlier one in terms of Laplace expansion. Let $(i_1, \cdots, i_n)$ be an ordered list of numbers from $\{1, \cdots, n\}$. This means the order is important so $(1, 2, 3)$ and $(2, 1, 3)$ are different. There will be some repetition between this section and the earlier section on determinants. The main purpose is to give all the missing proofs. Two books which give a good introduction to determinants are Apostol \[2\] and Rudin \[14\]. A recent book which also has a good introduction is Baker \[3\].

The following Lemma will be essential in the definition of the determinant.

**Lemma A.1.1** There exists a unique function, $\text{sgn}_n$ which maps each list of numbers from $\{1, \cdots, n\}$ to one of the three numbers, 0, 1, or $-1$ which also has the following properties.

\begin{align*}
\text{sgn}_n (1, \cdots, n) &= 1 \\
\text{sgn}_n (i_1, \cdots, p, \cdots, q, \cdots, i_n) &= -\text{sgn}_n (i_1, \cdots, q, \cdots, p, \cdots, i_n) \\
\text{sgn}_n (i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) &\equiv (-1)^{n-\theta} \text{sgn}_{n-1} (i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n) \tag{1.3}
\end{align*}

In words, the second property states that if two of the numbers are switched, the value of the function is multiplied by $-1$. Also, in the case where $n > 1$ and $\{i_1, \cdots, i_n\} = \{1, \cdots, n\}$ so that every number from $\{1, \cdots, n\}$ appears in the ordered list, $(i_1, \cdots, i_n)$,

\begin{align*}
\text{sgn}_n (i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) &\equiv (-1)^{n-\theta} \text{sgn}_{n-1} (i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n) \tag{1.3}
\end{align*}

where $n = i_{\theta}$ in the ordered list, $(i_1, \cdots, i_n)$. 275
Proof: To begin with, it is necessary to show the existence of such a function. This is clearly true if \( n = 1 \). Define \( \text{sgn}_n (1) \equiv 1 \) and observe that it works. No switching is possible. In the case where \( n = 2 \), it is also clearly true. Let \( \text{sgn}_2 (1, 2) = 1 \) and \( \text{sgn}_2 (2, 1) = -1 \) while \( \text{sgn}_2 (2, 2) = \text{sgn}_2 (1, 1) = 0 \) and verify it works. Assuming such a function exists for \( n \), \( \text{sgn}_{n+1} \) will be defined in terms of \( \text{sgn}_n \). If there are any repeated numbers in \( (i_1, \cdots, i_{n+1}) \), \( \text{sgn}_{n+1} (i_1, \cdots, i_{n+1}) \equiv 0 \). If there are no repeats, then \( n + 1 \) appears somewhere in the ordered list. Let \( \theta \) be the position of the number \( n + 1 \) in the list. Thus, the list is of the form \( (i_1, \cdots, i_{\theta-1}, n + 1, i_{\theta+1}, \cdots, i_{n+1}) \). From [1.3] it must be that
\[
\begin{align*}
\text{sgn}_{n+1} (i_1, \cdots, i_{\theta-1}, n + 1, i_{\theta+1}, \cdots, i_{n+1}) & = \\
(-1)^{n+1-\theta} \text{sgn}_n (i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_{n+1})
\end{align*}
\]
It is necessary to verify this satisfies [1.1] and [1.2] with \( n \) replaced with \( n + 1 \). The first of these is obviously true because
\[
\text{sgn}_{n+1} (1, \cdots, n, n + 1) \equiv (-1)^{n+1-(n+1)} \text{sgn}_n (1, \cdots, n) = 1.
\]
If there are repeated numbers in \( (i_1, \cdots, i_{n+1}) \), then it is obvious [1.2] holds because both sides would equal zero from the above definition. It remains to verify [1.2] in the case where there are no numbers repeated in \( (i_1, \cdots, i_{n+1}) \). Consider
\[
\text{sgn}_{n+1} (i_1, \cdots, r^p, \cdots, s^q, \cdots, i_{n+1}),
\]
where the \( r \) above the \( p \) indicates the number, \( p \) is in the \( r \)th position and the \( s \) above the \( q \) indicates that the number, \( q \) is in the \( s \)th position. Suppose first that \( r < \theta < s \). Then
\[
\begin{align*}
\text{sgn}_{n+1} (i_1, \cdots, r^p, \cdots, n + 1, \cdots, s^q, \cdots, i_{n+1}) & = \\
(-1)^{n+1-\theta} \text{sgn}_n (i_1, \cdots, r^p, \cdots, s^q, \cdots, i_{n+1})
\end{align*}
\]
while
\[
\begin{align*}
\text{sgn}_{n+1} (i_1, \cdots, q^p, \cdots, n + 1, \cdots, s^q, \cdots, i_{n+1}) & = \\
(-1)^{n+1-\theta} \text{sgn}_n (i_1, \cdots, q^p, \cdots, s^q, \cdots, i_{n+1})
\end{align*}
\]
and so, by induction, a switch of \( p \) and \( q \) introduces a minus sign in the result. Similarly, if \( \theta > s \) or \( \theta < r \) it also follows that [1.2] holds. The interesting case is when \( \theta = r \) or \( \theta = s \). Consider the case where \( \theta = r \) and note the other case is entirely similar.
\[
\begin{align*}
\text{sgn}_{n+1} (i_1, \cdots, r^p, \cdots, s^q, \cdots, i_{n+1}) & = \\
(-1)^{n+1-r} \text{sgn}_n (i_1, \cdots, s^q, \cdots, i_{n+1})
\end{align*}
\] (1.4)
while
\[
\begin{align*}
\text{sgn}_{n+1} (i_1, \cdots, q^p, \cdots, s^q, \cdots, i_{n+1}) & = \\
(-1)^{n+1-s} \text{sgn}_n (i_1, \cdots, r^p, \cdots, i_{n+1})
\end{align*}
\] (1.5)
By making \( s - 1 - r \) switches, move the \( q \) which is in the \( s - 1 \)th position in [1.4] to the \( r \)th position in [1.5]. By induction, each of these switches introduces a factor of \(-1\) and so
\[
\text{sgn}_n (i_1, \cdots, q^p, \cdots, i_{n+1}) = (-1)^{s-1-r} \text{sgn}_n (i_1, \cdots, q^p, \cdots, i_{n+1}).
\]
Therefore,
\[
\text{sgn}_{n+1} (i_1, \cdots, n+1, \cdots, i_n) = (-1)^{n+1-r} \text{sgn}_{n} (i_1, \cdots, r, \cdots, i_n+1)
\]
\[
= (-1)^{n+1-r} (-1)^{s-1-r} \text{sgn}_{n} (i_1, \cdots, r, \cdots, i_n+1)
\]
\[
= (-1)^{n+r} \text{sgn}_{n} (i_1, \cdots, r, \cdots, i_n+1) = (-1)^{2n-1} (-1)^{n+1-s} \text{sgn}_{n} (i_1, \cdots, r, \cdots, i_n+1)
\]
\[
= -\text{sgn}_{n+1} (i_1, \cdots, r, \cdots, n+1, \cdots, i_n+1).
\]

This proves the existence of the desired function.

To see this function is unique, note that you can obtain any ordered list of distinct numbers from a sequence of switches. If there exist two functions, \( f \) and \( g \) both satisfying \( 1.1 \) and \( 1.2 \) you could start with \( f (1, \cdots, n) = g (1, \cdots, n) \) and applying the same sequence of switches, eventually arrive at \( f (i_1, \cdots, i_n) = g (i_1, \cdots, i_n) \). If any numbers are repeated, then \( 1.2 \) gives both functions are equal to zero for that ordered list. This proves the lemma.

### A.2 The Determinant

#### A.2.1 The Definition

In what follows \( \text{sgn} \) will often be used rather than \( \text{sgn}_n \) because the context supplies the appropriate \( n \).

**Definition A.2.1** Let \( f \) be a real valued function which has the set of ordered lists of numbers from \( \{1, \cdots, n\} \) as its domain. Define
\[
\sum_{(k_1, \cdots, k_n)} f (k_1 \cdots k_n)
\]

to be the sum of all the \( f (k_1 \cdots k_n) \) for all possible choices of ordered lists \( (k_1, \cdots, k_n) \) of numbers of \( \{1, \cdots, n\} \). For example,
\[
\sum_{(k_1, k_2)} f (k_1, k_2) = f (1, 2) + f (2, 1) + f (1, 1) + f (2, 2).
\]

**Definition A.2.2** Let \( (a_{ij}) \) = \( A \) denote an \( n \times n \) matrix. The determinant of \( A \), denoted by \( \det (A) \) is defined by
\[
\det (A) \equiv \sum_{(k_1, \cdots, k_n)} \text{sgn} (k_1, \cdots, k_n) a_{1k_1} \cdots a_{nk_n}
\]

where the sum is taken over all ordered lists of numbers from \( \{1, \cdots, n\} \). Note it suffices to take the sum over only those ordered lists in which there are no repeats because if there are, \( \text{sgn} (k_1, \cdots, k_n) = 0 \) and so that term contributes 0 to the sum.

Let \( A \) be an \( n \times n \) matrix, \( A = (a_{ij}) \) and let \( (r_1, \cdots, r_n) \) denote an ordered list of \( n \) numbers from \( \{1, \cdots, n\} \). Let \( A (r_1, \cdots, r_n) \) denote the matrix whose \( k^{\text{th}} \) row is the \( r_k \) row of the matrix, \( A \). Thus
\[
\det (A (r_1, \cdots, r_n)) = \sum_{(k_1, \cdots, k_n)} \text{sgn} (k_1, \cdots, k_n) a_{r_1k_1} \cdots a_{r_nk_n}
\]

(1.6)

and
\[
A (1, \cdots, n) = A.
\]
A.2.2 Permuting Rows Or Columns

Proposition A.2.3 Let
\[(r_1, \cdots, r_n)\]
be an ordered list of numbers from \(\{1, \cdots, n\}\). Then
\[
\text{sgn} (r_1, \cdots, r_n) \det (A)
\]
\[= \sum_{(k_1, \cdots, k_n)} \text{sgn} (k_1, \cdots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n} \quad (1.7)
\]
\[= \det (A (r_1, \cdots, r_n)). \quad (1.8)
\]

Proof: Let \((1, \cdots, n) = (1, \cdots, r, \cdots, s, \cdots, n)\) so \(r < s\).
\[
\det (A (1, \cdots, r, \cdots, s, \cdots, n)) = \sum_{(k_1, \cdots, k_n)} \text{sgn} (k_1, \cdots, k_r, \cdots, k_s, \cdots, k_n) a_{1 k_1} \cdots a_{r k_r} \cdots a_{s k_s} \cdots a_{n k_n},
\]
and renaming the variables, calling \(k_s, k_r\) and \(k_r, k_s\), this equals
\[
= \sum_{(k_1, \cdots, k_n)} \text{sgn} (k_1, \cdots, k_r, \cdots, k_s, \cdots, k_n) a_{1 k_1} \cdots a_{r k_r} \cdots a_{s k_s} \cdots a_{n k_n}
\]
\[= \sum_{(k_1, \cdots, k_n)} -\text{sgn} \left( k_1, \cdots, k_r, k_s, \cdots, k_n \right) a_{1 k_1} \cdots a_{s k_s} \cdots a_{r k_r} \cdots a_{n k_n}
\]
\[= - \det (A (1, \cdots, s, \cdots, r, \cdots, n)). \quad (1.9)
\]
Consequently,
\[
\det (A (1, \cdots, s, \cdots, r, \cdots, n)) = - \det (A (1, \cdots, r, \cdots, s, \cdots, n)) = - \det (A)
\]
Now letting \(A (1, \cdots, s, \cdots, r, \cdots, n)\) play the role of \(A\), and continuing in this way, switching pairs of numbers,
\[
\det (A (r_1, \cdots, r_n)) = (-1)^p \det (A)
\]
where it took \(p\) switches to obtain \((r_1, \cdots, r_n)\) from \((1, \cdots, n)\). By Lemma [A.1.1], this implies
\[
\det (A (r_1, \cdots, r_n)) = (-1)^p \det (A) = \text{sgn} (r_1, \cdots, r_n) \det (A)
\]
and proves the proposition in the case when there are no repeated numbers in the ordered list, \((r_1, \cdots, r_n)\). However, if there is a repeat, say the \(r^{th}\) row equals the \(s^{th}\) row, then the reasoning of (1.9) - (1.10) shows that \(A (r_1, \cdots, r_n) = 0\) and also \(\text{sgn} (r_1, \cdots, r_n) = 0\) so the formula holds in this case also.

Observation A.2.4 There are \(n!\) ordered lists of distinct numbers from \(\{1, \cdots, n\}\).

To see this, consider \(n\) slots placed in order. There are \(n\) choices for the first slot. For each of these choices, there are \(n - 1\) choices for the second. Thus there are \(n \cdot (n - 1)\) ways to fill the first two slots. Then for each of these ways there are \(n - 2\) choices left for the third slot. Continuing this way, there are \(n!\) ordered lists of distinct numbers from \(\{1, \cdots, n\}\) as stated in the observation.
A.2.3 A Symmetric Definition

With the above, it is possible to give a more symmetric description of the determinant from which it will follow that det (A) = det (AT).

Corollary A.2.5 The following formula for det (A) is valid.

\[
\det (A) = \frac{1}{n!} \sum_{(r_1, \ldots, r_n) = (k_1, \ldots, k_n)} \text{sgn} (r_1, \ldots, r_n) \text{sgn} (k_1, \ldots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n}.
\]

(1.11)

And also det (AT) = det (A) where AT is the transpose of A. (Recall that for AT = (a_{ji}^T), a_{ij}^T = a_{ji}.)

**Proof:** From Proposition A.2.3 if the r_i are distinct,

\[
\det (A) = \sum_{(k_1, \ldots, k_n)} \text{sgn} (r_1, \ldots, r_n) \text{sgn} (k_1, \ldots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n}.
\]

Summing over all ordered lists, (r_1, \ldots, r_n) where the r_i are distinct, (If the r_i are not distinct, sgn (r_1, \ldots, r_n) = 0 and so there is no contribution to the sum.)

\[
n! \det (A) = \sum_{(r_1, \ldots, r_n), (k_1, \ldots, k_n)} \text{sgn} (r_1, \ldots, r_n) \text{sgn} (k_1, \ldots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n}.
\]

This proves the corollary since the formula gives the same number for A as it does for AT.

A.2.4 The Alternating Property Of The Determinant

Corollary A.2.6 If two rows or two columns in an n × n matrix, A, are switched, the determinant of the resulting matrix equals (−1) times the determinant of the original matrix. If A is an n × n matrix in which two rows are equal or two columns are equal then det (A) = 0. Suppose the i^{th} row of A equals (xa_1 + yb_1, \ldots, xa_n + yb_n). Then

\[
\det (A) = x \det (A_1) + y \det (A_2)
\]

where the i^{th} row of A_1 is (a_1, \ldots, a_n) and the i^{th} row of A_2 is (b_1, \ldots, b_n), all other rows of A_1 and A_2 coinciding with those of A. In other words, det is a linear function of each row A. The same is true with the word “row” replaced with the word “column”.

**Proof:** By Proposition A.2.3 when two rows are switched, the determinant of the resulting matrix is (−1) times the determinant of the original matrix. By Corollary A.2.5 the same holds for columns because the columns of the matrix equal the rows of the transposed matrix. Thus if A_1 is the matrix obtained from A by switching two columns,

\[
\det (A) = \det (A^T) = - \det (A_1^T) = - \det (A_1).
\]

If A has two equal columns or two equal rows, then switching them results in the same matrix. Therefore, det (A) = − det (A) and so det (A) = 0.

It remains to verify the last assertion.

\[
\det (A) \equiv \sum_{(k_1, \ldots, k_n)} \text{sgn} (k_1, \ldots, k_n) a_{1k_1} \cdots (x a_{k_i} + y b_{k_i}) \cdots a_{nk_n}
\]
\[ x \sum_{(k_1, \ldots, k_n)} \text{sgn}(k_1, \ldots, k_n) a_{1k_1} \cdots a_{nk_n} + y \sum_{(k_1, \ldots, k_n)} \text{sgn}(k_1, \ldots, k_n) a_{1k_1} \cdots b_{k_i} \cdots a_{nk_n} \equiv x \det(A_1) + y \det(A_2). \]

The same is true of columns because \( \det(A^T) = \det(A) \) and the rows of \( A^T \) are the columns of \( A \).

### A.2.5 Linear Combinations And Determinants

**Definition A.2.7** A vector, \( w \), is a linear combination of the vectors \( \{v_1, \cdots, v_r\} \) if there exists scalars, \( c_1, \cdots, c_r \) such that \( w = \sum_{k=1}^{r} c_k v_k. \) This is the same as saying \( w \in \text{span}\{v_1, \cdots, v_r\} \).

The following corollary is also of great use.

**Corollary A.2.8** Suppose \( A \) is an \( n \times n \) matrix and some column (row) is a linear combination of \( r \) other columns (rows). Then \( \det(A) = 0. \)

**Proof:** Let \( A = (a_1, \cdots, a_n) \) be the columns of \( A \) and suppose the condition that one column is a linear combination of \( r \) of the others is satisfied. Then by using Corollary [A.2.6] you may rearrange the columns to have the \( n^{th} \) column a linear combination of the first \( r \) columns. Thus \( a_n = \sum_{k=1}^{r} c_k a_k \) and so

\[ \det(A) = \det(\begin{array}{c} a_1 \cdots a_r \cdots a_{n-1} \sum_{k=1}^{r} c_k a_k \end{array}). \]

By Corollary [A.2.6]

\[ \det(A) = \sum_{k=1}^{r} c_k \det(\begin{array}{c} a_1 \cdots a_r \cdots a_{n-1} a_k \end{array}) = 0. \]

The case for rows follows from the fact that \( \det(A) = \det(A^T) \). This proves the corollary.

### A.2.6 The Determinant Of A Product

Recall the following definition of matrix multiplication.

**Definition A.2.9** If \( A \) and \( B \) are \( n \times n \) matrices, \( A = (a_{ij}) \) and \( B = (b_{ij}) \), \( AB = (c_{ij}) \) where

\[ c_{ij} \equiv \sum_{k=1}^{n} a_{ik} b_{kj}. \]

One of the most important rules about determinants is that the determinant of a product equals the product of the determinants.

**Theorem A.2.10** Let \( A \) and \( B \) be \( n \times n \) matrices. Then

\[ \det(AB) = \det(A) \det(B). \]
Proof: Let \( c_{ij} \) be the \( ij \)th entry of \( AB \). Then by Proposition [A.2.3]

\[
\det(AB) = \sum_{(k_1, \ldots, k_n)} \operatorname{sgn}(k_1, \ldots, k_n) c_{1k_1} \cdots c_{nk_n}
\]

\[
= \sum_{(k_1, \ldots, k_n)} \operatorname{sgn}(k_1, \ldots, k_n) \left( \sum_{r_1} a_{1r_1} b_{r_1k_1} \right) \cdots \left( \sum_{r_n} a_{nr_n} b_{r_nk_n} \right)
\]

\[
= \sum_{(r_1 \ldots r_n)} \sum_{(k_1, \ldots, k_n)} \operatorname{sgn}(k_1, \ldots, k_n) b_{r_1k_1} \cdots b_{r_nk_n} (a_{1r_1} \cdots a_{nr_n})
\]

This proves the theorem.

**A.2.7 Cofactor Expansions**

**Lemma A.2.11** Suppose a matrix is of the form

\[
M = \begin{pmatrix} A & * \\ 0 & a \end{pmatrix}
\]

or

\[
M = \begin{pmatrix} A & 0 \\ * & a \end{pmatrix}
\]

where \( a \) is a number and \( A \) is an \((n-1) \times (n-1)\) matrix and * denotes either a column or a row having length \( n-1 \) and the 0 denotes either a column or a row of length \( n-1 \) consisting entirely of zeros. Then

\[
\det(M) = a \det(A).
\]

**Proof:** Denote \( M \) by \((m_{ij})\). Thus in the first case, \( m_{nn} = a \) and \( m_{ni} = 0 \) if \( i \neq n \) while in the second case, \( m_{nr} = a \) and \( m_{ir} = 0 \) if \( i \neq n \). From the definition of the determinant,

\[
\det(M) \equiv \sum_{(k_1, \ldots, k_n)} \operatorname{sgn}(k_1, \ldots, k_n) m_{1k_1} \cdots m_{nk_n}
\]

Letting \( \theta \) denote the position of \( n \) in the ordered list, \((k_1, \ldots, k_n)\) then using the earlier conventions used to prove Lemma [A.1.1], \( \det(M) \) equals

\[
\sum_{(k_1, \ldots, k_n)} (-1)^{n-\theta} \operatorname{sgn}_{n-1}(k_1, \ldots, k_{\theta-1}, k_{\theta+1}, \ldots, k_n) m_{1k_1} \cdots m_{nk_n}
\]

Now suppose [A.1.13]. Then if \( k_n \neq n \), the term involving \( m_{nk_n} \) in the above expression equals zero. Therefore, the only terms which survive are those for which \( \theta = n \) or in other words, those for which \( k_n = n \). Therefore, the above expression reduces to

\[
a \sum_{(k_1, \ldots, k_{n-1})} \operatorname{sgn}_{n-1}(k_1, \ldots, k_{n-1}) m_{1k_1} \cdots m_{(n-1)k_{n-1}} = a \det(A).
\]

To get the assertion in the situation of [1.12] use Corollary [A.2.5] and [1.13] to write

\[
\det(M) = \det(M^T) = \det \left( \begin{pmatrix} A^T & 0 \\ * & a \end{pmatrix} \right) = a \det(A^T) = a \det(A).
\]
This proves the lemma.

In terms of the theory of determinants, arguably the most important idea is that of Laplace expansion along a row or a column. This will follow from the above definition of a determinant.

**Definition A.2.12** Let \( A = (a_{ij}) \) be an \( n \times n \) matrix. Then a new matrix called the cofactor matrix, \( \text{cof} (A) \) is defined by \( \text{cof} (A) = (c_{ij}) \) where to obtain \( c_{ij} \) delete the \( i^{th} \) row and the \( j^{th} \) column of \( A \), take the determinant of the \((n - 1) \times (n - 1)\) matrix which results, (This is called the \( ij^{th} \) minor of \( A \) ) and then multiply this number by \((-1)^{i+j}\). To make the formulas easier to remember, \( \text{cof} (A)_{ij} \) will denote the \( ij^{th} \) entry of the cofactor matrix.

The following is the main result. Earlier this was given as a definition and the outrageous totally unjustified assertion was made that the same number would be obtained by expanding the determinant along any row or column. The following theorem proves this assertion.

**Theorem A.2.13** Let \( A \) be an \( n \times n \) matrix where \( n \geq 2 \). Then

\[
\det (A) = \sum_{j=1}^{n} a_{ij} \text{cof} (A)_{ij} = \sum_{i=1}^{n} a_{ij} \text{cof} (A)_{ij}.
\]

The first formula consists of expanding the determinant along the \( i^{th} \) row and the second expands the determinant along the \( j^{th} \) column.

**Proof:** Let \((a_i, \ldots, a_n)\) be the \( i^{th} \) row of \( A \). Let \( B_j \) be the matrix obtained from \( A \) by leaving every row the same except the \( i^{th} \) row which in \( B_j \) equals \((0, \cdots, 0, a_{ij}, 0, \cdots, 0)\). Then by Corollary [A.2.6]

\[
\det (A) = \sum_{j=1}^{n} \det (B_j)
\]

Denote by \( A^{ij} \) the \((n - 1) \times (n - 1)\) matrix obtained by deleting the \( i^{th} \) row and the \( j^{th} \) column of \( A \). Thus \( \text{cof} (A)_{ij} \equiv (-1)^{i+j} \det (A^{ij}) \). At this point, recall that from Proposition [A.2.3] when two rows or two columns in a matrix, \( M \), are switched, this results in multiplying the determinant of the old matrix by \(-1\) to get the determinant of the new matrix. Therefore, by Lemma [A.2.11]

\[
\det (B_j) = (-1)^{n-j} (-1)^{n-i} \det \left( \begin{array}{cc}
A^{ij} & * \\
0 & a_{ij}
\end{array} \right) = (-1)^{i+j} \det \left( \begin{array}{cc}
A^{ij} & * \\
0 & a_{ij}
\end{array} \right) = a_{ij} \text{cof} (A)_{ij}.
\]

Therefore,

\[
\det (A) = \sum_{j=1}^{n} a_{ij} \text{cof} (A)_{ij}
\]

which is the formula for expanding \( \det (A) \) along the \( i^{th} \) row. Also,

\[
\det (A) = \det (A^T) = \sum_{j=1}^{n} a_{ij} \text{cof} (A^T)_{ij} = \sum_{j=1}^{n} a_{ji} \text{cof} (A)_{ji}
\]

which is the formula for expanding \( \det (A) \) along the \( i^{th} \) column. This proves the theorem. Note that this gives an easy way to write a formula for the inverse of an \( n \times n \) matrix. Recall the definition of the inverse of a matrix in Definition [12.7.2] on Page [238].
A.2.8 Formula For The Inverse

Theorem A.2.14 $A^{-1}$ exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = (a_{ij}^{-1})$ where

$$a_{ij}^{-1} = \det(A)^{-1} \text{cof}(A)_{ji}$$

for cof $(A)_{ij}$ the $ij^{th}$ cofactor of $A$.

**Proof:** By Theorem A.2.13 and letting $(a_{ir}) = A$, if $\det(A) \neq 0$,

$$\sum_{i=1}^{n} a_{ir} \text{cof}(A)_{ir} \det(A)^{-1} = \det(A) \det(A)^{-1} = 1.$$

Now consider

$$\sum_{i=1}^{n} a_{ir} \text{cof}(A)_{ik} \det(A)^{-1}$$

when $k \neq r$. Replace the $k^{th}$ column with the $r^{th}$ column to obtain a matrix, $B_k$ whose determinant equals zero by Corollary A.2.6. However, expanding this matrix along the $k^{th}$ column yields

$$0 = \det(B_k) \det(A)^{-1} = \sum_{i=1}^{n} a_{ir} \text{cof}(A)_{ik} \det(A)^{-1}.$$

Summarizing,

$$\sum_{i=1}^{n} a_{ir} \text{cof}(A)_{ik} \det(A)^{-1} = \delta_{rk}.$$

Using the other formula in Theorem A.2.13 and similar reasoning,

$$\sum_{j=1}^{n} a_{rj} \text{cof}(A)_{kj} \det(A)^{-1} = \delta_{rk}$$

This proves that if $\det(A) \neq 0$, then $A^{-1}$ exists with $A^{-1} = (a_{ij}^{-1})$, where

$$a_{ij}^{-1} = \text{cof}(A)_{ji} \det(A)^{-1}.$$

Now suppose $A^{-1}$ exists. Then by Theorem A.2.10,

$$1 = \det(I) = \det(AA^{-1}) = \det(A) \det(A^{-1})$$

so $\det(A) \neq 0$. This proves the theorem.

The next corollary points out that if an $n \times n$ matrix, $A$ has a right or a left inverse, then it has an inverse.

**Corollary A.2.15** Let $A$ be an $n \times n$ matrix and suppose there exists an $n \times n$ matrix, $B$ such that $BA = I$. Then $A^{-1}$ exists and $A^{-1} = B$. Also, if there exists $C$ an $n \times n$ matrix such that $AC = I$, then $A^{-1}$ exists and $A^{-1} = C$.

**Proof:** Since $BA = I$, Theorem A.2.10 implies

$$\det B \det A = 1$$

and so $\det A \neq 0$. Therefore from Theorem A.2.14, $A^{-1}$ exists. Therefore,

$$A^{-1} = (BA) A^{-1} = B (AA^{-1}) = BI = B.$$
The case where \( CA = I \) is handled similarly.

The conclusion of this corollary is that left inverses, right inverses and inverses are all the same in the context of \( n \times n \) matrices.

Theorem \( \text{A.2.14} \) says that to find the inverse, take the transpose of the cofactor matrix and divide by the determinant. The transpose of the cofactor matrix is called the adjugate or sometimes the classical adjoint of the matrix \( A \). It is an abomination to call it the adjoint although you do sometimes see it referred to in this way. In words, \( A^{-1} \) is equal to one over the determinant of \( A \) times the adjugate matrix of \( A \).

### A.2.9 Cramer’s Rule

In case you are solving a system of equations, \( Ax = y \) for \( x \), it follows that if \( A^{-1} \) exists,

\[
x = (A^{-1}A)x = A^{-1}(Ax) = A^{-1}y
\]

thus solving the system. Now in the case that \( A^{-1} \) exists, there is a formula for \( A^{-1} \) given above. Using this formula,

\[
x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \sum_{j=1}^{n} \frac{1}{\det(A)} \text{cof}(A)_{ji} y_j.
\]

By the formula for the expansion of a determinant along a column,

\[
x_i = \frac{1}{\det(A)} \det\begin{pmatrix}
* & \cdots & y_1 & \cdots & * \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
* & \cdots & y_n & \cdots & *
\end{pmatrix},
\]

where here the \( i^{th} \) column of \( A \) is replaced with the column vector, \((y_1, \ldots, y_n)^T\), and the determinant of this modified matrix is taken and divided by \( \det(A) \). This formula is known as Cramer’s rule.

### A.2.10 Upper Triangular Matrices

**Definition A.2.16** A matrix \( M \), is upper triangular if \( M_{ij} = 0 \) whenever \( i > j \). Thus such a matrix equals zero below the main diagonal, the entries of the form \( M_{ii} \) as shown.

\[
\begin{pmatrix}
* & * & \cdots & * \\
0 & * & \ddots & \vdots \\
\vdots & \ddots & \ddots & * \\
0 & \ldots & 0 & *
\end{pmatrix}
\]

A lower triangular matrix is defined similarly as a matrix for which all entries above the main diagonal are equal to zero.

With this definition, here is a simple corollary of Theorem \( \text{A.2.13} \).

**Corollary A.2.17** Let \( M \) be an upper (lower) triangular matrix. Then \( \det(M) \) is obtained by taking the product of the entries on the main diagonal.
A.2.11 The Determinant Rank

Definition A.2.18 A submatrix of a matrix $A$ is the rectangular array of numbers obtained by deleting some rows and columns of $A$. Let $A$ be an $m \times n$ matrix. The **determinant rank** of the matrix equals $r$ where $r$ is the largest number such that some $r \times r$ submatrix of $A$ has a non zero determinant.

Theorem A.2.19 If $A$, an $m \times n$ matrix has determinant rank, $r$, then there exist $r$ rows (columns) of the matrix such that every other row (column) is a linear combination of these $r$ rows (columns).

**Proof:** Suppose the determinant rank of $A = (a_{ij})$ equals $r$. Thus some $r \times r$ submatrix has non zero determinant and there is no larger square submatrix which has non zero determinant. Suppose such a submatrix is determined by the $r$ columns whose indices are $j_1 < \cdots < j_r$

and the $r$ rows whose indices are $i_1 < \cdots < i_r$

I want to show that every row is a linear combination of these rows. Consider the $l^{th}$ row and let $p$ be an index between 1 and $n$. Form the following $(r+1) \times (r+1)$ matrix

\[
\begin{pmatrix}
  a_{i_1j_1} & \cdots & a_{i_1j_r} & a_{i_1p} \\
  \vdots & \ddots & \vdots & \vdots \\
  a_{i_rj_1} & \cdots & a_{i_rj_r} & a_{i_rp} \\
  a_{lj_1} & \cdots & a_{lj_r} & a_{lp}
\end{pmatrix}
\]

Of course you can assume $l \notin \{i_1, \cdots, i_r\}$ because there is nothing to prove if the $l^{th}$ row is one of the chosen ones. The above matrix has determinant 0. This is because if $p \notin \{j_1, \cdots, j_r\}$ then the above would be a submatrix of $A$ which is too large to have non zero determinant. On the other hand, if $p \in \{j_1, \cdots, j_r\}$ then the above matrix has two columns which are equal so its determinant is still 0.

Expand the determinant of the above matrix along the last column. Let $C_k$ denote the cofactor associated with the entry $a_{i_1p}$. This is not dependent on the choice of $p$. Remember, you delete the column and the row the entry is in and take the determinant of what is left and multiply by $-1$ raised to an appropriate power. Let $C$ denote the cofactor associated with $a_{lp}$. This is given to be nonzero, it being the determinant of the matrix

\[
\begin{pmatrix}
  a_{i_1j_1} & \cdots & a_{i_1j_r} \\
  \vdots & \ddots & \vdots \\
  a_{i_rj_1} & \cdots & a_{i_rj_r}
\end{pmatrix}
\]

Thus

\[
0 = a_{lp}C + \sum_{k=1}^{r} C_k a_{i_kp}
\]

which implies

\[
a_{lp} = \sum_{k=1}^{r} -\frac{C_k}{C} a_{i_kp} \equiv \sum_{k=1}^{r} m_k a_{i_kp}
\]

Since this is true for every $p$ and since $m_k$ does not depend on $p$, this has shown the $l^{th}$ row is a linear combination of the $i_1, i_2, \cdots, i_r$ rows. The determinant rank does not change when you replace $A$ with $A^T$. Therefore, the same conclusion holds for the columns. This proves the theorem.
A.2.12 Telling Whether \( A \) Is One To One Or Onto

The following theorem is of fundamental importance and ties together many of the ideas presented above.

**Theorem A.2.20** Let \( A \) be an \( n \times n \) matrix. Then the following are equivalent.

1. \( \det (A) = 0 \).
2. \( A, A^T \) are not one to one.
3. \( A \) is not onto.

**Proof:** Suppose \( \det (A) = 0 \). Then the determinant rank of \( A = r < n \). Therefore, there exist \( r \) columns such that every other column is a linear combination of these columns by Theorem A.2.19. In particular, it follows that for some \( m \), the \( m \)th column is a linear combination of all the others. Thus letting \( A = ( a_1 \cdots a_m \cdots a_n ) \) where the columns are denoted by \( a_i \), there exists scalars, \( \alpha_i \) such that

\[
a_m = \sum_{k \neq m} \alpha_k a_k.
\]

Now consider the column vector, \( x \equiv ( \alpha_1 \cdots -1 \cdots \alpha_n )^T \). Then

\[
Ax = -a_m + \sum_{k \neq m} \alpha_k a_k = 0.
\]

Since also \( A0 = 0 \), it follows \( A \) is not one to one. Similarly, \( A^T \) is not one to one by the same argument applied to \( A^T \). This verifies that 1.) implies 2.).

Now suppose 2.). Then since \( A^T \) is not one to one, it follows there exists \( x \neq 0 \) such that

\[
A^T x = 0.
\]

Taking the transpose of both sides yields

\[
x^T A = 0^T
\]

where the \( 0^T \) is a \( 1 \times n \) matrix or row vector. Now if \( Ay = x \), then

\[
|x|^2 = x^T (Ay) = (x^T A) y = 0^T y = 0
\]

contrary to \( x \neq 0 \). Consequently there can be no \( y \) such that \( Ay = x \) and so \( A \) is not onto. This shows that 2.) implies 3.).

Finally, suppose 3.). If 1.) does not hold, then \( \det (A) \neq 0 \) but then from Theorem A.2.14 \( A^{-1} \) exists and so for every \( y \in \mathbb{F}^n \) there exists a unique \( x \in \mathbb{F}^n \) such that \( Ax = y \). In fact \( x = A^{-1} y \). Thus \( A \) would be onto contrary to 3.). This shows 3.) implies 1.) and proves the theorem.

**Corollary A.2.21** Let \( A \) be an \( n \times n \) matrix. Then the following are equivalent.

1. \( \det(A) \neq 0 \).
2. \( A \) and \( A^T \) are one to one.
3. \( A \) is onto.

**Proof:** This follows immediately from the above theorem.
A.2.13 Schur’s Theorem

Consider the following system of equations for \( x_1, x_2, \cdots, x_n \)
\[
\sum_{j=1}^{n} a_{ij} x_j = 0, \quad i = 1, 2, \cdots, m
\]  
(1.14)

where \( m < n \). Then the following theorem is a fundamental observation.

**Theorem A.2.22** Let the system of equations be as just described in (1.14) where \( m < n \). Then letting
\[
x^T \equiv (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n,
\]
there exists \( x \neq 0 \) such that the components satisfy each of the equations of (1.14).

**Proof:** The above system is of the form
\[
Ax = 0
\]
where \( A \) is an \( m \times n \) matrix with \( m < n \). Therefore, if you form the matrix
\[
\begin{pmatrix}
A \\
0
\end{pmatrix},
\]
an \( n \times n \) matrix having \( n - m \) rows of zeros on the bottom, it follows this matrix has determinant equal to 0. Therefore, from Theorem A.2.19 there exists \( x \neq 0 \) such that \( Ax = 0 \). This proves the theorem.

**Definition A.2.23** A set of vectors in \( \mathbb{R}^n \) \( \{x_1, \cdots, x_k\} \) is called an **orthonormal** set of vectors if
\[
x_i \cdot x_j = \delta_{ij} \equiv \begin{cases} 
1 & \text{if } i = j \\
0 & \text{if } i \neq j
\end{cases}
\]

**Theorem A.2.24** Let \( v_1 \) be a unit vector \((|v_1| = 1)\) in \( \mathbb{R}^n, n > 1 \). Then there exist vectors \( \{v_2, \cdots, v_n\} \) such that
\[
\{v_1, \cdots, v_n\}
\]
is an orthonormal set of vectors.

**Proof:** The equation for \( x \)
\[
v_1 \cdot x = 0
\]
has a nonzero solution \( x \) by Theorem A.2.22. Pick such a solution and divide by its magnitude to get \( v_2 \) a unit vector such that \( v_1 \cdot v_2 = 0 \). Now suppose \( v_1, \cdots, v_k \) have been chosen such that \( \{v_1, \cdots, v_k\} \) is an orthonormal set of vectors. Then consider the equations
\[
v_j \cdot x = 0 \quad j = 1, 2, \cdots, k
\]
This amounts to the situation of Theorem A.2.22 in which there are more variables than equations. Therefore, by this theorem, there exists a nonzero \( x \) solving all these equations. Divide by its magnitude and this gives \( v_{k+1} \). This proves the theorem.

**Definition A.2.25** If \( U \) is an \( n \times n \) matrix whose columns form an orthonormal set of vectors, then \( Q \) is called an **orthogonal matrix**. Note that from the way we multiply matrices,
\[
U^T U = UU^T = I.
\]

Thus \( U^{-1} = U^T \).
Note the product of orthogonal matrices is orthogonal because
\[(U_1U_2)^T(U_1U_2) = U_1^TU_1^TU_1U_2 = I.\]

Two matrices \(A\) and \(B\) are similar if there is some invertible matrix \(S\) such that \(A = S^{-1}BS\). Note that similar matrices have the same characteristic equation because by Theorem [A.2.10] which says the determinant of a product is the product of the determinants,
\[
\det(\lambda I - A) = \det(\lambda I - S^{-1}BS) = \det(S^{-1}(\lambda I - B)S) = \det(S^{-1})\det(\lambda I - B)\det(S) = \det(\lambda I - B)
\]

With this preparation, here is a case of Schur’s theorem.

**Theorem A.2.26** Let \(A\) be a real \(n \times n\) matrix which has all real eigenvalues. Then there exists an orthogonal matrix, \(U\) such that
\[
U^TAU = T,
\]
where \(T\) is an upper triangular matrix having the eigenvalues of \(A\) on the main diagonal listed according to multiplicity as zeros of the characteristic equation.

**Proof:** The theorem is clearly true if \(A\) is a \(1 \times 1\) matrix. Just let \(U = 1\) the \(1 \times 1\) matrix which has 1 down the main diagonal and zeros elsewhere. Suppose it is true for \((n - 1) \times (n - 1)\) matrices and let \(A\) be an \(n \times n\) matrix. Then let \(v_1\) be a unit eigenvector for \(A\). Then there exists \(\lambda_1\) such that
\[
Av_1 = \lambda_1v_1, \quad |v_1| = 1.
\]

By Theorem [A.2.24] there exists \(\{v_1, \cdots, v_n\}\), an orthonormal set in \(\mathbb{R}^n\). Let \(U_0\) be a matrix whose \(i^{th}\) column is \(v_i\). Then from the above, it follows \(U_0\) is orthogonal. Then from the way you multiply matrices \(U_0^T A U_0\) is of the form
\[
\begin{pmatrix}
\lambda_1 & * & \cdots & * \\
0 & & & \\
\vdots & & A_1 & \\
0 & & & 
\end{pmatrix}
\]

where \(A_1\) is an \((n - 1) \times (n - 1)\) matrix. The above matrix is similar to \(A\) so it has the same eigenvalues and indeed the same characteristic equation. Also the eigenvalues of \(A_1\) are all real because each of these eigenvalues is an eigenvalue of the above matrix and is therefore an eigenvalue of \(A\). Now by induction there exists an \((n - 1) \times (n - 1)\) orthogonal matrix \(\tilde{U}_1\) such that
\[
\tilde{U}_1^*A_1\tilde{U}_1 = T_{n-1},
\]
an upper triangular matrix. Consider
\[
U_1 \equiv \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}_1 \end{pmatrix}
\]
This is a orthogonal matrix and
\[
U_1^TU_0^TAU_0U_1 = \left( \begin{array}{cc} 1 & 0 \\ 0 & \tilde{U}_1^* \end{array} \right) \left( \begin{array}{cc} \lambda_1 & * \\ \lambda_1 & A_1 \end{array} \right) \left( \begin{array}{cc} 1 & 0 \\ 0 & \tilde{U}_1 \end{array} \right) = \left( \begin{array}{cc} \lambda_1 & * \\ 0 & T_{n-1} \end{array} \right) \equiv T
\]
where \( T \) is upper triangular. Then let \( U = U_0U_1 \). Since \((U_0U_1)^T = U_1^T U_0^T\), it follows \( A \) is similar to \( T \) and that \( U_0U_1 \) is orthogonal. Hence \( A \) and \( T \) have the same characteristic polynomials and since the eigenvalues of \( T \) are the diagonal entries listed according to algebraic multiplicity, this proves the theorem.

### A.2.14 Symmetric Matrices

Recall a real matrix \( A \) is symmetric if \( A = A^T \).

**Lemma A.2.27** A real symmetric matrix has all real eigenvalues.

**Proof:** Recall the eigenvalues are solutions \( \lambda \) to
\[
\det (\lambda I - A) = 0
\]
and so by Theorem [A.2.20], there exists \( x \) a vector such that
\[
Ax = \lambda x, \quad x \neq 0
\]
Of course if \( A \) is real, it is still possible that the eigenvalue could be complex and if this is the case, then the vector \( x \) will also end up being complex. I wish to show the eigenvalues are all real. Suppose then that \( \lambda \) is an eigenvalue and let \( x \) be the corresponding eigenvector described above. Then letting \( \overline{x} \) denote the complex conjugate of \( x \),
\[
\lambda x^T \overline{x} = (Ax)^T \overline{x} = x^TA^T \overline{x} = x^T \overline{A}x = x^T \overline{Ax} = x^T \overline{x} \overline{\lambda}
\]
and so, cancelling \( x^T \overline{x} \), it follows \( \lambda = \overline{\lambda} \) showing \( \lambda \) is real. This proves the lemma.

**Theorem A.2.28** Let \( A \) be a real symmetric matrix. Then there exists a diagonal matrix \( D \) consisting of the eigenvalues of \( A \) down the main diagonal and an orthogonal matrix \( U \) such that
\[
U^T AU = D.
\]

**Proof:** Since \( A \) has all real eigenvalues, it follows from Theorem [A.2.26] there exists an orthogonal matrix \( U \) such that
\[
U^T AU = T
\]
where \( T \) is upper triangular. Now
\[
T^T = U^T A^T U = U^T AU = T
\]
and so in fact \( T \) is a diagonal matrix having the eigenvalues of \( A \) down the diagonal. This proves the theorem.

**Theorem A.2.29** Let \( A \) be a real symmetric matrix which has all positive eigenvalues \( 0 < \lambda_1 \leq \lambda_2 \cdots \leq \lambda_n \). Then
\[
(Ax \cdot x) \equiv x^T Ax \geq \lambda_1 |x|^2
\]

**Proof:** Let \( U \) be the orthogonal matrix of Theorem [A.2.28]. Then
\[
(Ax \cdot x) = x^T Ax = (x^T U) D (U^T x) = (U^T x)^T D (U^T x) \geq \lambda_1 \sum_i |(U^T x)_i|^2 = \lambda_1 (U^T x \cdot U^T x) = \lambda_1(U^T x)^T U^T x = \lambda_1 x^T I x = \lambda_1 |x|^2.
\]
A.3 Exercises

1. Let \( m < n \) and let \( A \) be an \( m \times n \) matrix. Show that \( A \) is not one to one. Hint: Consider the \( n \times n \) matrix, \( A_1 \) which is of the form

\[
A_1 \equiv \begin{pmatrix} A & 0 \\ \end{pmatrix}
\]

where the 0 denotes an \( (n - m) \times n \) matrix of zeros. Thus \( \det A_1 = 0 \) and so \( A_1 \) is not one to one. Now observe that \( A_1 x \) is the vector,

\[
A_1 x = \begin{pmatrix} A x & 0 \\ \end{pmatrix}
\]

which equals zero if and only if \( A x = 0 \).

2. Show that matrix multiplication is associative. That is, \((AB)C = A(BC)\).

3. Show the inverse of a matrix, if it exists, is unique. Thus if \( AB = BA = I \), then \( B = A^{-1} \).

4. In the proof of Theorem A.2.14 it was claimed that \( \det (I) = 1 \). Here \( I = (\delta_{ij}) \). Prove this assertion. Also prove Corollary A.2.17.

5. Let \( v_1, \cdots, v_n \) be vectors in \( \mathbb{F}^n \) and let \( M(v_1, \cdots, v_n) \) denote the matrix whose \( i^{th} \) column equals \( v_i \). Define

\[
d(v_1, \cdots, v_n) \equiv \det (M(v_1, \cdots, v_n)).
\]

Prove that \( d \) is linear in each variable, (multilinear), that

\[
d(v_1, \cdots, v_i, \cdots, v_j, \cdots, v_n) = -d(v_1, \cdots, v_j, \cdots, v_i, \cdots, v_n), \tag{1.16}
\]

and

\[
d(e_1, \cdots, e_n) = 1 \tag{1.17}
\]

where here \( e_j \) is the vector in \( \mathbb{F}^n \) which has a zero in every position except the \( j^{th} \) position in which it has a one.

6. Suppose \( f : \mathbb{F}^n \times \cdots \times \mathbb{F}^n \rightarrow \mathbb{F} \) satisfies \[1.16\] and \[1.17\] and is linear in each variable. Show that \( f = d \).

7. Show that if you replace a row (column) of an \( n \times n \) matrix \( A \) with itself added to some multiple of another row (column) then the new matrix has the same determinant as the original one.

8. If \( A = (a_{ij}) \), show \( \det (A) = \sum_{(k_1, \cdots, k_n)} \text{sgn} (k_1, \cdots, k_n) a_{k_1} \cdots a_{k_n} \).

9. Use the result of Problem 7 to evaluate by hand the determinant

\[
\det \begin{pmatrix} 1 & 2 & 3 & 2 \\ -6 & 3 & 2 & 3 \\ 5 & 2 & 2 & 3 \\ 3 & 4 & 6 & 4 \end{pmatrix}.
\]

10. Find the inverse if it exists of the matrix,

\[
\begin{pmatrix} e^t & \cos t & \sin t \\ e^t & -\sin t & \cos t \\ e^t & -\cos t & -\sin t \end{pmatrix}.
\]
11. Two $n \times n$ matrices, $A$ and $B$, are similar if $B = S^{-1}AS$ for some invertible $n \times n$ matrix, $S$. Show that if two matrices are similar, they have the same characteristic polynomials.

12. Suppose the characteristic polynomial of an $n \times n$ matrix, $A$, is of the form

$$t^n + a_{n-1}t^{n-1} + \cdots + a_1t + a_0$$

and that $a_0 \neq 0$. Find a formula $A^{-1}$ in terms of powers of the matrix, $A$. Show that $A^{-1}$ exists if and only if $a_0 \neq 0$. 
The Fundamental Theorem Of Algebra

The fundamental theorem of algebra states that every non constant polynomial having coefficients in \( \mathbb{C} \) has a zero in \( \mathbb{C} \). If \( \mathbb{C} \) is replaced by \( \mathbb{R} \), this is not true because of the example, \( x^2 + 1 = 0 \). This theorem is a very remarkable result and notwithstanding its title, all the best proofs of it depend on either analysis or topology. It was studied by many people in the eighteenth century. Gauss gave a proof in 1797 which had a loose end. The first correct proof was given by Argand in 1806. A good discussion of the history of this theorem as well as many different ways to prove it are found in the Wikipedia article on the web. Just google fundamental theorem of algebra. The proof given here follows Rudin [14]. See also Hardy [8] for a similar proof, more discussion and references. The best proof is found in the theory of complex analysis.

Recall Corollary 8.0.6 which says every nonzero complex number has \( k \) \( k^{th} \) roots.

Now here is a definition of what it means for a sequence of complex numbers to converge. You will note it is the same definition given for a sequence of real numbers only here the absolute value refers to the absolute value of a complex number.

**Definition B.0.1** Let \( \{z_n\} \) be a sequence of complex numbers. Then

\[
\lim_{n \to \infty} z_n = z
\]

means: For every \( \varepsilon > 0 \) there exists \( n_\varepsilon \) such that if \( n \geq n_\varepsilon \), then

\[
|z_n - z| < \varepsilon.
\]

Now here is a useful observation.

**Proposition B.0.2** Suppose \( \{a_n + ib_n\} \) is a sequence of complex numbers. Then it converges to \( a + ib \) if and only if \( \lim_{n \to \infty} a_n = a \) and \( \lim_{n \to \infty} b_n = b \).

**Proof:** I will leave the details to you. Recall

\[
|(a_n + ib_n) - (a + ib)| = \sqrt{(a - a_n)^2 + (b - b_n)^2}.
\]

This implies

\[
\max (|a_n - a|, |b_n - b|) \leq |(a_n + ib_n) - (a + ib)| \leq |a_n - a| + |b_n - b|
\]

Now the conclusion of the proposition follows right away. This proves the proposition.

Next is an important existence theorem.
Theorem B.0.3 Denote by $[a, b] + i [c, d]$ the set of complex numbers $x + iy$ where $x \in [a, b]$ and $y \in [c, d]$. Suppose $\{z_n\}$ is a sequence of complex numbers in this set. Then there exists a complex number $z \in [a, b] + i [c, d]$ and a subsequence $\{z_{n_k}\}$ such that

$$\lim_{k \to \infty} z_{n_k} = z.$$ 

Proof: Let $z_n = x_n + iy_n$ and consider first the sequence $\{x_n\} \subseteq [a, b]$. Consider the two intervals $[a, \frac{x_n + b}{2}]$ and $[\frac{x_n + a}{2}, b]$, each having length $(b - a)/2$. Then in one of these intervals, perhaps both, there are $x_n$ for infinitely many values of $n$. Pick one of the two intervals for which this is so. Then divide it in half. One of the two halves has $x_n$ for infinitely many values of $n$. Pick the half for which this is so. Divide it in half and pick the half which has the property that it contains $x_n$ for infinitely many $n$. Continue this way. Thus there is a sequence of intervals $\{I_n\}$ where the length of the $n$th interval is no larger than $(b - a)2^{-n+1}$ and each of these intervals contains $x_n$ for infinitely many values of $n$. Now pick $x_{k_1} \in I_1$. If $x_{k_1}, \ldots, x_{k_1}$ have been chosen, let $x_{k_{n+1}}$ be in $I_{n+1}$ such that $k_{n+1} > k_n$. This can be done because each interval contains $x_n$ for infinitely many $n$. Then $\{x_{n_k}\}_{k=1}^\infty$ is a Cauchy sequence because for $k, l > n$,

$$|x_{n_k} - x_{n_l}| < (b - a)2^{-n+1}.$$ 

By Theorem 4.6.15 and Theorem 4.6.6, this sequence converges to some $x \in [a, b]$. Now consider the sequence $\{y_{n_k}\}_{k=1}^\infty \subseteq [c, d]$. By the same reasoning there is a subsequence $\{y_{n_{k_1}}\}_{l=1}^\infty$ which converges to $y \in [c, d]$. It follows from Theorem 4.6.10 that $\{x_{n_{k_1}}\}_{l=1}^\infty$ also converges to $x$. By Proposition B.0.2

$$\lim_{l \to \infty} x_{n_{k_1}} + iy_{n_{k_1}} = x + iy \in [a, b] + i [c, d].$$

This proves the theorem.

Lemma B.0.4 If $p(z) = \sum_{k=0}^n a_k z^k$ is a polynomial and if $\lim_{n \to \infty} z_n = z$, then

$$\lim_{n \to \infty} |p(z_n)| = |p(z)|.$$ 

Proof: $z_n = z_n - z + z$ and so by the binomial theorem,

$$p(z_n) = a_0 + \sum_{k=1}^n a_k \sum_{j=0}^k \binom{k}{j} z^{k-j} (z_n - z)^j$$

$$= a_0 + \sum_{k=1}^n a_k \left( z^k + \sum_{j=1}^k \binom{k}{j} z^{k-j} (z_n - z)^j \right)$$

$$= p(z) + \sum_{k=1}^n a_k \sum_{j=1}^k \binom{k}{j} z^{k-j} (z_n - z)^j \equiv p(z) + e(z_n - z)$$

where $\lim_{n \to \infty} e(z_n - z) = 0$. It follows from the triangle inequality

$$||p(z_n)| - |p(z)|| \leq |p(z_n) - p(z)| \leq |e(z_n - z)|$$

and so $\lim_{n \to \infty} |p(z_n)| = |p(z)|$ as claimed. This proves the lemma.

Lemma B.0.5 Let $p(z)$ be a polynomial with complex coefficients as above. Then $|p(z)|$ achieves its minimum value on any set of the form $[-R, R] + i [-R, R]$. 

Proof: $|p(z)|$ is bounded below by 0 and so

$$\lambda \equiv \inf \{|p(z)| : z \in [-R, R] + i [-R, R]\}$$

exists. By Proposition 1.11.3 there exists a sequence $\{z_n\}$ of points of $[-R, R] + i [-R, R]$ which satisfy

$$\lambda = \lim_{n \to \infty} |p(z_n)|$$

This is called a minimizing sequence. Then by Theorem B.0.3 there is a subsequence $\{z_{n_k}\}$ which converges to a point $z \in [-R, R] + i [-R, R]$. Then by Lemma B.0.4,

$$\lambda = \lim_{k \to \infty} |p(z_{n_k})| = |p(z)|.$$ 

This proves the lemma.

**Theorem B.0.6 (Fundamental theorem of Algebra)** Let $p(z)$ be a nonconstant polynomial. Then there exists $z \in \mathbb{C}$ such that $p(z) = 0$.

**Proof:** Let

$$p(z) = \sum_{k=0}^{n} a_k z^k$$

where $a_n \neq 0$, $n > 0$. Then

$$|p(z)| \geq |a_n| |z|^n - \sum_{k=0}^{n-1} |a_k| |z|^k$$

and so

$$\lim_{|z| \to \infty} |p(z)| = \infty. \quad (2.1)$$

Now let

$$\lambda \equiv \inf \{|p(z)| : z \in \mathbb{C}\}.$$ 

By (2.1) there exists an $R > 0$ such that if $z \notin [-R, R] + i [-R, R]$, it follows that $|p(z)| > \lambda + 1$. Therefore,

$$\lambda \equiv \inf \{|p(z)| : z \in \mathbb{C}\} = \inf \{|p(z)| : z \in [-R, R] + i [-R, R]\}.$$ 

By Lemma B.0.5, there exists $w \in [-R, R] + i [-R, R]$ such that

$$\lambda = |p(w)|$$

I want to argue $\lambda = 0$. Suppose it is greater than 0. Then consider

$$q(z) \equiv \frac{p(z+w)}{p(w)}.$$ 

It follows $q(z)$ is of the form

$$q(z) = 1 + c_k z^k + \cdots + c_n z^n$$

where $c_k \neq 0$, because $q(0) = 1$. It is also true that

$$|q(z)| = \frac{|p(z+w)|}{|p(w)|} \geq 1$$

I want to argue $|q(z)|$ is bounded below by 0 and so
by the assumption that \(|p(w)|\) is the smallest value of \(|p(z)|\). Now let \(\theta \in \mathbb{C}\) be a complex number with \(|\theta| = 1\) and
\[
\theta c_k w^k = -|w|^k |c_k|.
\]
(If \(w \neq 0\), \(\theta = -\frac{|w|^k |c_k|}{w^k c_k}\)
and if \(w = 0\), \(\theta = 1\) will work.) Next let \(\eta^k = \theta\) and let \(t\) be a small positive number.
\[
q(t\eta w) \equiv 1 - t^k |w|^k |c_k| + \cdots + c_n t^n (\eta w)^n
\]
which is of the form
\[1 - t^k |w|^k |c_k| + t^k (g(t, w))\]
where \(\lim_{t \to 0} g(t, w) = 0\). Letting \(t\) be small enough,
\[|g(t, w)| < |w|^k |c_k|/2\]
and so for such \(t\),
\[|q(t\eta w)| < 1 - t^k |w|^k |c_k| + t^k |w|^k |c_k|/2 < 1,
\]a contradiction to \(|q(z)| \geq 1\). This proves the theorem.
Bibliography

Index

∩, 15
∪, 15

adjugate, 258, 282
amplitude, 125
angle between vectors, 185
annuity
  future value, 24
  ordinary, 24
  present value, 25
Archimedian property, 22
area of a cone, 132
area of a parallelogram, 194
augmented matrix, 208
back substitution, 208
barallelepiped
  volume, 199
basic variables, 215
binomial theorem, 25
box product, 199

Cartesian coordinates, 168
Cauchy Schwarz inequality, 183, 190
Cauchy sequence, 183
classical adjoint, 258
cofactor, 254, 280
cofactor matrix, 254
complete the square, 49
completeness, 26
completeness axiom, 27
complex conjugate, 161
complex numbers, 161
component, 178, 192, 193
component of a force, 187
components of a matrix, 224
composition, 39
compound interest, 93
conformable, 228
consistent, 217
Coordinates, 167
Cramer’s rule, 262, 282
cross product, 194
degree, 120
dense, 23
density of rationals, 22
Descartes’ rule of signs, 84
determinant, 275
  product, 278
  transpose, 277
digraph, 231
Dirichlet function, 38
discriminant, 51
distance, 114
distance formula, 171
domain, 37
dot product, 183
digonal form, 210
effective rate, 94
echelon form, 210
ellipse, 153
entries of a matrix, 224
equivalence class, 79
equivalence relation, 79
Euclidean algorithm, 23

factorial, 30
Fibonacci sequence, 87
field
  splitting, 86
field axioms, 10
field extensions, 81
focus, 153
force, 176
free variables, 215
function, 37
fundamental theorem of algebra, 73, 164, 291
fundamental theorem of arithmetic, 61
future value
  annuity, 95
  single payment, 94
INDEX

Gauss Elimination, 217
Gauss elimination, 209
Gauss Jordan method for inverses, 238
greatest common divisor, 59, 66
greatest lower bound, 26

hyperbola, 156
hypergeometric distribution, 35

ideal, 85
maximal, 85
inconsistent, 214, 217
infinite series, 107
infinite sum, 107
inner product, 183
inscribed angle, 132
intervals, 16
inverse function, 39
inverse image, 39
inverses and determinants, 260, 281
invertible, 236
irreducible, 66
isosceles triangle, 116

joule, 189

Laplace expansion, 253, 280
law of cosines, 115
law of sines, 125
leading entry, 209
least upper bound, 26
limit comparison test, 150
linear combination, 278
linear transformation, 246

main diagonal, 254
mathematical induction, 20
matrix, 223
inverse, 236
left inverse, 282
lower triangular, 254, 282
right inverse, 282
upper triangular, 254, 282
minor, 254, 289
minute, 120
monic, 65
multinomial expansion, 34

Newton, 179
nilpotent, 265

oblique asymptote, 57
one to one, 37, 246
onto, 37, 246

order axioms, 14
orthogonal matrix, 265, 285
orthonormal, 285

parabola, 151
parallelepiped, 199
partial fractions, 74
Pascal’s triangle, 25
Pathagorean theorem, 112
permutations and combinations, 33
perpendicular, 186
phase shift, 125
pivot, 215
pivot column, 210
pivot position, 210
Poisson random variable, 148
polar form complex number, 162
polynomial, 40
degree of a polynomial, 41
position vector, 169, 170, 177
present value
ordinary annuity, 96
payment, 96
prime number, 59
projection of a vector, 188
quadratic formula, 51
radian measure, 120
range, 37
rational function, 41
rational root theorem, 63
real numbers, 9
recurrence relation, 87
recursively defined sequence, 87
relatively prime, 59
resultant, 178
right handed system, 194
ring
commutative, 85
row operations, 255
scalar product, 183
scalars, 168, 223
sequence of partial sums, 93
sequences, 87
skew lines, 205
skew symmetric, 235
slope, 34
span, 278
speed, 179
squeezing theorem, 102
standard position, 177
INDEX

subtend, 116
symmetric, 235
synthetic division, 67

triangle inequality, 18, 174, 184
   complex numbers, 162
trichotomy, 15
trigonometric functions, 115

vectors, 167, 176
velocity, 179

well ordered, 20