Calculus, Applications and Theory

Kenneth Kuttler

June 24, 2004

Contents

Introduction

Calculus consists of the study of limits of various sorts and the systematic exploitation of the completeness axiom. It was developed by physicists and engineers over a period of several hundred years in order to solve problems from the physical sciences. It is the language by which precision and quantitative predictions for many complicated problems are obtained. It is used to find lengths of curves, areas and volumes of regions which are not bounded by straight lines. It is used to predict and account for the motion of satellites. It is essential in order to solve many maximization problems and it is prerequisite material in order to understand models based on differential equations. These and other applications are discussed to some extent in this book.

It is assumed the reader has a good understanding of algebra on the level of college algebra or what used to be called algebra II along with some exposure to geometry and trigonometry although the book does contain an extensive review of these things. I have tried to keep the book a manageable length in order to focus more on the important ideas. I have also tried to give complete proofs of all theorems in one variable calculus and to at least give plausibility arguments for those in multiple dimensions. Physical models are derived in the usual way through the use of differentials leading to differential equations which are introduced early and used throughout the book as the basis for physical models.

I expect the reader to be able to use a calculator whenever it would be helpful to do so. Some introduction to the use of computer algebra systems is also presented and there are exercises which require the use of some form of technology. Having said this, calculus is not about using calculators or any other form of technology. I believe that when the syntax and arcane notation associated with technology are presented too prominently, these things become the topic of study rather than the concepts of calculus. This is a book on calculus and should not be considered an instruction manual for the use of technology.

Pictures are often helpful in seeing what is going on and there are many pictures in this book for this reason. However, calculus is not about drawing pictures and ultimately rests on logic and definitions. Algebra plays a central role in gaining the sort of understanding which generalizes to higher dimensions where pictures are not available. Therefore, I have emphasized the algebraic aspects of this subject far more than is usual, especially linear algebra which is absolutely essential to understand in order to do multivariable calculus. I have also featured the repeated index summation convention and the usual reduction identities which allow one to discover vector identities.

Part I

Preliminaries

The Real Numbers

2.0.1 Outcomes

1. Understand the geometric and algebraic significance of a real number.
2. Understand and solve inequalities and be able to use set notation.
3. Understand the absolute value algebraically and geometrically and be able to solve inequalities involving the absolute value. Understand the triangle inequality.
4. Understand well ordering of the natural numbers and the relation to mathematical induction. Be able to prove theorems using math induction.
5. Solve systems of linear equations using row operations.
6. Understand completeness of the real line and its significance.

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.

2.1 The Number Line And Algebra Of The Real Numbers

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

$$\mathbb{Z} \equiv \{\cdots -1, 0, 1, \cdots\},$$

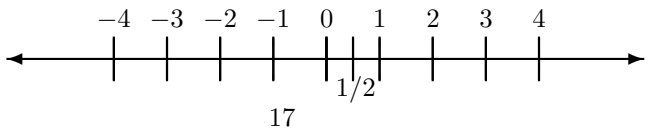
the natural numbers,

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

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.



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 2.1.1 $x+y=y+x$, *(commutative law for addition)*

Axiom 2.1.2 $x+0=x$, *(additive identity).*

Axiom 2.1.3 *For each* $x \in \mathbb{R}$, *there exists* $-x \in \mathbb{R}$ *such that* $x+(-x)=0$, *(existence of additive inverse).*

Axiom 2.1.4 $(x+y)+z=x+(y+z)$, *(associative law for addition).*

Axiom 2.1.5 $xy=yx$, *(commutative law for multiplication).*

Axiom 2.1.6 $(xy)\,z=x\,(yz)$, *(associative law for multiplication).*

Axiom 2.1.7 $1x=x$, *(multiplicative identity).*

Axiom 2.1.8 *For each* $x \neq 0$, *there exists* x^{-1} *such that* $xx^{-1}=1$.*(existence of multiplicative inverse).*

Axiom 2.1.9 $x\,(y+z)=xy+xz$.*(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 $x/y \equiv x\left(y^{-1}\right)$. 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 $x\left(y^{-1}\right)$ 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.[1] 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.

Theorem 2.1.10 *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

$$\begin{aligned} 0 &= (-0x) + 0x = (-0x) + (0x + 0x) \\ &= ((-0x) + 0x) + 0x = 0 + 0x = 0x \end{aligned}$$

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.,

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

[1] 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.

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,

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

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 $xy^2=x\left(y^2\right)$ 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 2.1.11 *Let $x_1,x_2,\cdots,x_m$ 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, $x_1,x_2,\cdots,x_m$ 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 2.1.12 $\sum_{k=1}^{6}(2k+1)=48$.

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,

$$\sum_{j=k}^{m}x_j\equiv x_k+\cdots+x_m.$$

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 2.1.13 *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 $\frac{y\left(x^2+y\right)}{(x-1)(x^2+y)}$. Then since these have the same common denominator, you add them as follows.

$$\begin{aligned} \frac{x}{x^2+y} + \frac{y}{x-1} &= \frac{x\,(x-1)}{(x^2+y)\,(x-1)} + \frac{y\left(x^2+y\right)}{(x-1)\,(x^2+y)} \\ &= \frac{x^2 - x + yx^2 + y^2}{(x^2+y)\,(x-1)}. \end{aligned}$$

2.2 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 $\frac{x}{x^2-1} + \frac{x-1}{x+1}$.
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 $\frac{1}{3} = \frac{1}{x+y} = \frac{1}{x} + \frac{1}{y} = 1 + \frac{1}{2} = \frac{3}{2}$. Then cross multiplying, yields $2 = 9$.

12. Simplify $\frac{x^2y^4z^{-6}}{x^{-2}y^{-1}z}$.

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

    (a) $\frac{x^2y+xy^2+x}{x}$

    (b) $\frac{x^2y+xy^2+x}{xy}$

    (c) $\frac{x^3+2x^2-x-2}{x+1}$

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) $(x - y)\left(x^2 + xy + y^2\right) = x^3 - y^3$

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

16. Find the error in the following.

$$\frac{xy + y}{x} = y + y = 2y.$$

    Now let $x = 2$ and $y = 2$ to obtain

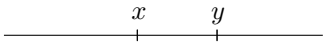
$$3 = 4$$

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

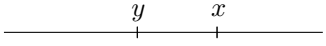
2.3 Order

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



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



$x \leq y$ 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 2.3.2 *The sum of two positive real numbers is positive.*

Axiom 2.3.3 *The product of two positive real numbers is positive.*

Axiom 2.3.4 *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 2.3.5 *If $x < y$ and $y < z$ then $x < z$ (Transitive law).*

Axiom 2.3.6 *If $x < y$ then $x + z < y + z$ (addition to an inequality).*

Axiom 2.3.7 *If $x \leq 0$ and $y \leq 0$, then $xy \geq 0$.*

Axiom 2.3.8 *If $x > 0$ then $x^{-1} > 0$.*

Axiom 2.3.9 *If $x < 0$ then $x^{-1} < 0$.*

Axiom 2.3.10 *If $x < y$ then $xz < yz$ if $z > 0$, (multiplication of an inequality by a positive number).*

Axiom 2.3.11 *If $x < y$ and $z < 0$, then $xz > zy$ (multiplication of an inequality by a negative number).*

Axiom 2.3.12 *Each of the above holds with $>$ and $<$ replaced by $\geq$ and $\leq$ respectively except for 2.3.8 and 2.3.9 in which it is also necessary to stipulate that $x \neq 0$.*

Axiom 2.3.13 *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 2.3.14 *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 2.3.15 *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 nonpositive. 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 2.3.16 *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 $x \in \mathbb{R}$.

To simplify the way such things are written, involves set notation. This is described next.

2.3.1 Set Notation

A set is just a collection of things called elements. 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 $a \leq x < b$. (a,b) consists of the set of real numbers, x such that $a < x < b$ and $(a,b]$ indicates the set of numbers, x such that $a < x \leq b$. $[a,\infty)$ means the set of all numbers, x such that $x \geq a$ 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 ∞ 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 2.3.17 *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 2.3.18 *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 2.3.19 *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 $(-\infty, \infty)$.

2.4 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 $\left[\frac{-1}{3}, 2\right]$.

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

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

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

   Note that $\frac{x+1}{2x-2}$ 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 $[-2, 3]$.

2.5 Exercises

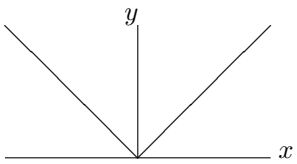
1. Solve $(3x+2)(x-3) \leq 0$.
2. Solve $(3x+2)(x-3) > 0$.
3. Solve $\frac{x+2}{3x-2} < 0$.
4. Solve $\frac{x+1}{x+3} < 1$.
5. Solve $(x-1)(2x+1) \leq 2$.
6. Solve $(x-1)(2x+1) > 2$.
7. Solve $x^2 - 2x \leq 0$.
8. Solve $(x+2)(x-2)^2 \leq 0$.
9. Solve $\frac{3x-4}{x^2+2x+2} \geq 0$.
10. Solve $\frac{3x+9}{x^2+2x+1} \geq 1$.
11. Solve $\frac{x^2+2x+1}{3x+7} < 1$.

2.6 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 2.6.1 $|x| \equiv \begin{cases} x \text{ if } x \geq 0, \\ -x \text{ if } x < 0. \end{cases}$

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.



The following is a fundamental theorem about the absolute value.

Theorem 2.6.2 $|xy| = |x|\,|y|$.

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 2.6.3 *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 2.6.2,

$$\begin{aligned} |x + y|^2 &= \left|(x + y)^2\right| = (x + y)^2 \\ &= x^2 + y^2 + 2xy \leq x^2 + y^2 + 2\,|x|\,|y| \\ &= |x|^2 + |y|^2 + 2\,|x|\,|y| = (|x| + |y|)^2\,. \end{aligned}$$

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

$$\begin{aligned} |x| &= |x - y + y| \\ &\leq |x - y| + |y| \end{aligned}$$

and so

$$|x| - |y| \leq |x - y| \tag{2.1}$$

Now switch the letters to obtain

$$|y| - |x| \leq |y - x| = |x - y|. \tag{2.2}$$

Therefore,

$$||x| - |y|| \leq |x - y|$$

because if $|x| - |y| \geq 0$, then the conclusion follows from (2.1) while if $|x| - |y| \leq 0$, the conclusion follows from (2.2). 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 2.6.4 *Solve the equation* $|x - 1| = 2$

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 2.6.5 *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 2.6.6 *Solve the inequality* $|2x - 1| > 2$.

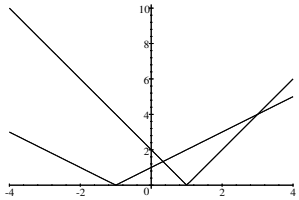
This happens if $2x - 1 > 2$ or if $2x - 1 < -2$. Thus the solution is $x > 3/2$ or $x < -1/2$, $\left(\frac{3}{2}, \infty\right) \cup \left(-\infty, -\frac{1}{2}\right)$.

Example 2.6.7 *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 2.6.8 *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, $\left(\frac{1}{3}, 3\right)$ 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 $(-\infty, \frac{1}{3}] \cup [3, \infty)$.

Example 2.6.9 *Obtain a number, δ, such that if $|x-2| < \delta$, then $\left|x^2 - 4\right| < 1/10$.*

If $|x-2| < 1$, then $||x| - |2|| < 1$ and so $|x| < 3$. Therefore, if $|x-2| < 1$,

$$\begin{aligned} \left|x^2 - 4\right| &= |x+2|\,|x-2| \\ &\leq (|x|+2)\,|x-2| \\ &\leq 5\,|x-2|\,. \end{aligned}$$

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,

$$\begin{aligned} \left|x^2 - 4\right| &= |x+2|\,|x-2| \\ &\leq (|x|+2)\,|x-2| \\ &\leq 7\,|x-2| \end{aligned}$$

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 2.6.10 *Suppose $\varepsilon > 0$ is a given positive number. Obtain a number, $\delta > 0$, such that if $|x-1| < \delta$, then $\left|x^2 - 1\right| < \varepsilon$.*

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

$$\left|x^2 - 1\right| < 3\,|x-1|\,.$$

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

$$\left|x^2 - 1\right| < 3\,|x-1| < 3\frac{\varepsilon}{3} = \varepsilon.$$

2.7 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. 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 $xy \geq yz$.
10. Solve $(x+1)(2x-2)x \geq 0$.
11. Solve $\frac{x+3}{2x+1} > 1$.
12. Solve $\frac{x+2}{3x+1} > 2$.
13. Describe the set of numbers, a such that there is no solution to $|x+1| = 4 - |x+a|$.
14. Suppose $0 < a < b$. Show $a^{-1} > b^{-1}$.
15. Show that if $|x-6| < 1$, then $|x| < 7$.
16. Suppose $|x-8| < 2$. How large can $|x-5|$ be?
17. Obtain a number, $\delta > 0$, such that if $|x-1| < \delta$, then $\left|x^2 - 1\right| < 1/10$.
18. Obtain a number, $\delta > 0$, such that if $|x-4| < \delta$, then $|\sqrt{x} - 2| < 1/10$.
19. 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 δ will depend in some way on ε. You need to tell how.

2.8 Well Ordering Principle And Archimedian Property

Definition 2.8.1 *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 2.8.2 *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 2.8.3 *(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 2.8.4 *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

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

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

$$\begin{aligned}\frac{n(2n+1)}{6} + (n+1) &= \frac{6(n+1) + 2n^2 + n}{6} \\ &= \frac{(n+2)(2n+3)}{6}\end{aligned}$$

Therefore,

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

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

Example 2.8.5 *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

$$\begin{aligned}\frac{1}{2} \cdot \frac{3}{4} \cdots \frac{2n-1}{2n} \cdot \frac{2n+1}{2n+2} &< \frac{1}{\sqrt{2n+1}} \frac{2n+1}{2n+2} \\ &= \frac{\sqrt{2n+1}}{2n+2}.\end{aligned}$$

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.

Lets 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 2.8.6 *The Archimedian property states that whenever $x \in \mathbb{R}$, and $a > 0$, there exists $n \in \mathbb{N}$ such that $na > x$.*

Axiom 2.8.7 $\mathbb{R}$ *has the Archimedian property.*

This is not hard to believe. Just look at the number line. This Archimedian 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 2.8.8 *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

$$S \equiv \{w \in \mathbb{N} : w \geq y\}.$$

The set S is nonempty by the Archimedian 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

$$y - x \leq y - (k-1) = \overbrace{y - k}^{\leq 0} + 1 \leq 1$$

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 2.8.9 *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 2.8.8 there exists $m \in \mathbb{Z}$ such that

$$nx < m < ny$$

and so take $r = m/n$.

Definition 2.8.10 *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 $\frac{79}{22}$. 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 2.8.11 *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,

$$r \equiv b - pa \geq 0.$$

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 2.8.8. 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.

2.9 Exercises

1. The Archimedian property implies the rational numbers are dense in $\mathbb{R}$. Now consider the numbers which are of the form $\frac{k}{2^m}$ 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 $\sum_{k=1}^{n} k^4$ 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 $\sum_{k=1}^{n} ar^k = a\frac{r^{n+1}}{r-1} - a\frac{r}{r-1}$.

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. 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. Now 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? **Hint:** The rate per payment period is .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

$$\begin{array}{c} 1 \\ 1\ 1 \\ 1\ 2\ 1 \\ 1\ 3\ 3\ 1 \\ 1\ 4\ 6\ 4\ 1 \end{array}$$

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. **Hint:** Letting the numbers of the n^{th} row of Pascal's triangle be denoted by $\binom{n}{0}, \binom{n}{1}, \cdots, \binom{n}{n}$ 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 $(n+1)! \equiv (n+1)\,n!$ 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. **Hint:** You might try using the preceding problem.

19. Show that for $p \in (0,1)$, $\sum_{k=0}^{n} \binom{n}{k} k p^k (1-p)^{n-k} = np$.

[2] Blaise Pascal lived in the 1600's and is responsible for the beginnings of the study of probability.

20. Using the binomial theorem prove that for all $n \in \mathbb{N}$, $\left(1+\frac{1}{n}\right)^n \leq \left(1+\frac{1}{n+1}\right)^{n+1}$.
Hint: Show first that $\binom{n}{k} = \frac{n\cdot(n-1)\cdots(n-k+1)}{k!}$. By the binomial theorem,

$$\left(1+\frac{1}{n}\right)^n = \sum_{k=0}^{n}\binom{n}{k}\left(\frac{1}{n}\right)^k = \sum_{k=0}^{n}\frac{\overbrace{n\cdot(n-1)\cdots(n-k+1)}^{k \text{ factors}}}{k!n^k}.$$

Now consider the term $\frac{n\cdot(n-1)\cdots(n-k+1)}{k!n^k}$ 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 $n+1$ whereever 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.

21. Prove by induction that for all $k \geq 4$, $2^k \leq k!$

22. Use the Problems 21 and 20 to verify for all $n \in \mathbb{N}$, $\left(1+\frac{1}{n}\right)^n \leq 3$.

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

24. 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?

25. 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?

2.10 Fundamental Theorem Of Arithmetic*

It is not necessary to read this section in order to do calculus. However, it is good general knowledge and so is included. The following definition describes what is meant by a prime number and also what is meant by the word "divides".

Definition 2.10.1 *The number, a divides the number, b if in Theorem 2.8.11, $r = 0$. 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. The greatest common divisor of two positive integers, m, n is that number, p which has the property that p divides both m and n and also if q divides both m and n, then q divides p. Two integers are relatively prime if their greatest common divisor is one.*

Theorem 2.10.2 *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 $p = x_0 m + y_0 n$. Either p divides m or it does not. If p does not divide m, then by Theorem 2.8.11,

$$m = pq + r$$

where $0 < r < p$. Thus $m = (x_0 m + y_0 n)\, q + r$ and so, solving for r,

$$r = m\,(1 - x_0) + (-y_0 q)\, n \in S.$$

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,

$$p = mx_0 + ny_0 = x_0 qx + y_0 qy = q\,(x_0 x + y_0 y)$$

showing $q|p$. Therefore, $p = (m, n)$.

Theorem 2.10.3 *If p is a prime and $p|ab$ then either $p|a$ or $p|b$.*

Proof: Suppose p does not divide a. Then since the only factors of p are 1 and p it 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 $p|ab$, $ab = pz$ for some integer z. Therefore,

$$b = abx + ybp = pzx + ybp = p\,(xz + yb)$$

and this shows p divides b.

Theorem 2.10.4 *(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 the 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 2.10.3, $p_1|q_j$ for some j. Since these are prime numbers this requires $p_1 = q_1$. 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.

The next theorem is a very nice high school theorem which characterizes all possible rational roots for polynomials having integer coefficients.

Theorem 2.10.5 *(rational root theorem) Let*

$$a_n x^n+\cdots+a_1x+a_0=0$$

where each a_i is an integer and $a_n\neq 0$. Then if the equation has any rational solutions, these are of the form

$$\pm\frac{\textit{factor of } a_0}{\textit{factor of } a_n}.$$

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,

$$a_n p^n+a_{n-1}p^{n-1}q+\cdots+a_1pq^{n-1}+a_0q^n=0.$$

Hence

$$a_np^n=-\left(a_{n-1}p^{n-1}q+\cdots+a_0q^n\right)$$

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 2.10.3 $q|a_n$ because it does not divide p^n due to the fact that p^n and q have no prime factors in common.

Similarly,

$$a_0q^n=-\left(a_np^n+\cdots+a_1pq^{n-1}\right)$$

and so $p|a_0q^n$ but $(p,q^n)=1$. By Theorem 2.10.3 again, $p|a_0$ and this proves the theorem.

Example 2.10.6 *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 2.10.5, the only possible rational roots to this equation are ± 2 and ± 1 and none of these work. Therefore, $\sqrt{2}$ must be irrational.

2.11 Exercises

1. Using Theorem 2.10.5, show $\sqrt[7]{6}$, $\sqrt[3]{7}$, $\sqrt[5]{5}$ are all irrational numbers. This means they are not rational.

2. 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.

3. Euclid[3] 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.

[3]He lived about 300 B.C.

4. 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 $[a, b] = ab/(a, b)$. **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 **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.

5. Show that if $\{a, b, c\}$ are three positive integers, they have a greatest common divisor which may be written as $ax + by + cz$ for some integers x, y, z.

6. Let $a_n = 2^{2^n} + 1$ 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[4] showed that when $n = 5$, the number is not prime[5]. When numbers of this form are prime, they are called Fermat[6] 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, [6]. **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

$$\begin{aligned} pk_2 &= a_m = \left(2^{2^n}\right)^{2^r} + 1 = (pk_1 - 1)^{2^r} + 1 \\ &= p\,(\text{integer}) + 2. \end{aligned}$$

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

2.12 Systems Of Equations

Sometimes it is necessary to solve systems of equations. For example the problem could be to find x and y such that

$$x + y = 7 \text{ and } 2x - y = 8. \tag{2.3}$$

The set of ordered pairs, (x, y) which solve both equations is called the solution set. For example, you can see that $(5, 2) = (x, y)$ is a solution to the above system. To solve this, note that the solution set does not change if any equation is replaced by a non zero multiple of itself. It also does not change if one equation is replaced by itself added to a multiple of the other equation. For example, x and y solve the above system if and only if x and y solve the system

$$x + y = 7, \overbrace{2x - y + (-2)\,(x + y) = 8 + (-2)\,(7)}^{-3y=-6}. \tag{2.4}$$

[4]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.

[5]The number in this case is $4,294,967,297$.

[6]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.

The second equation was replaced by -2 times the first equation added to the second. Thus the solution is $y = 2$, from $-3y = -6$ and now, knowing $y = 2$, it follows from the other equation that $x + 2 = 7$ and so $x = 5$.

Why exactly does the replacement of one equation with a multiple of another added to it not change the solution set? The two equations of (2.3) are of the form

$$E_1 = f_1, E_2 = f_2 \tag{2.5}$$

where E_1 and E_2 are expressions involving the variables. The claim is that if a is a number, then (2.5) has the same solution set as

$$E_1 = f_1,\ E_2 + aE_1 = f_2 + af_1. \tag{2.6}$$

Why is this?

If (x, y) solves (2.5) then it solves the first equation in (2.6). Also, it satisfies $aE_1 = af_1$ and so, since it also solves $E_2 = f_2$ it must solve the second equation in (2.6). If (x, y) solves (2.6) then it solves the first equation of (2.5). Also $aE_1 = af_1$ and it is given that the second equation of (2.6) is verified. Therefore, $E_2 = f_2$ and it follows (x, y) is a solution of the second equation in (2.5). This shows the solutions to (2.5) and (2.6) are exactly the same which means they have the same solution set. Of course the same reasoning applies with no change if there are many more variables than two and many more equations than two. It is still the case that when one equation is replaced with a multiple of another one added to itself, the solution set of the whole system does not change.

The other thing which does not change the solution set of a system of equations consists of listing the equations in a different order. Here is another example.

Example 2.12.1 *Find the solutions to the system,*

$$\begin{aligned} x + 3y + 6z &= 25 \\ 2x + 7y + 14z &= 58 \\ 2y + 5z &= 19 \end{aligned} \tag{2.7}$$

To solve this system replace the second equation by (-2) times the first equation added to the second. This yields. the system

$$\begin{aligned} x + 3y + 6z &= 25 \\ y + 2z &= 8 \\ 2y + 5z &= 19 \end{aligned} \tag{2.8}$$

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

$$\begin{aligned} x + 3y + 6z &= 25 \\ y + 2z &= 8 \\ z &= 3 \end{aligned} \tag{2.9}$$

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 not really much different from what you have always done in solving a single equation. For example, suppose you wanted to solve $2x + 5 = 3x - 6$. You did the same thing to both sides of the equation thus preserving the solution set until you obtained

an equation which was simple enough to give the answer. In this case, you would add $-2x$ to both sides and then add 6 to both sides. This yields $x = 11$.

In (2.9) 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

$$\begin{array}{c} x + 3y = 19 \\ y = 6 \\ z = 3 \end{array}$$

Now add (-3) times the second to the top. This yields

$$\begin{array}{c} x = 1 \\ y = 6 \\ z = 3 \end{array},$$

a system which has the same solution set as the original system.

It is foolish to write the variables every time you do these operations. It is easier to write the system (2.7) as the following "augmented matrix"

$$\begin{pmatrix} 1 & 3 & 6 & 25 \\ 2 & 7 & 14 & 58 \\ 0 & 2 & 5 & 19 \end{pmatrix}.$$

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 (2.7) 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 (2.8). Next take (-2) times the second row and add to the third,

$$\begin{pmatrix} 1 & 3 & 6 & 25 \\ 0 & 1 & 2 & 8 \\ 0 & 0 & 1 & 3 \end{pmatrix}$$

which is the same as (2.9). You get the idea I hope. Write the system as an augmented matrix and follow the procedure of either switching rows, multiplying a row by a non zero number, or replacing a row by a multiple of another row added to it. Each of these operations leaves the solution set unchanged. These operations are called row operations.

Example 2.12.2 *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.

$$\begin{pmatrix} 2 & 4 & -3 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 21 \end{pmatrix}.$$

Putting in the variables, the last two rows say $z = 1$ and $z = 21$. This is impossible so the last system of equations determined by the above augmented matrix has no solution. However, it has the same solution set as the first system of equations. This shows there is no solution to the three given equations. When this happens, the system is called inconsistent.

This should not be surprising that something like this can take place. It can even happen for one equation in one variable. Consider for example, $x = x+1$. There is clearly no solution to this.

Example 2.12.3 *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

$$\begin{pmatrix} 3 & -1 & -5 & 9 \\ 0 & 1 & -10 & 0 \\ 0 & 1 & -10 & 0 \end{pmatrix}$$

Next take -1 times the middle row and add to the bottom.

$$\begin{pmatrix} 3 & -1 & -5 & 9 \\ 0 & 1 & -10 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

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 says $y = 10z$ and $x = 3 + 5z$. Apparently z can equal any number. 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 and this is a good description of the solutions. This is what it is all about, finding the solutions to the system.

The phenomenon of an infinite solution set occurs in equations having only one variable also. For example, consider the equation $x = x$. It doesn't matter what x equals.

Definition 2.12.4 *A system of linear equations is a list of equations,*

$$\sum_{j=1}^{n} a_{ij}x_j = f_j,\ i = 1, 2, 3, \cdots, m$$

where a_{ij} are numbers, f_j is a number, and it is desired to find $(x_1, \cdots, x_n)$ 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. It turns out these are the only three cases which can occur for linear systems. 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. 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. These things are all the same.

Example 2.12.5 *Give the complete solution to the system of equations, $-41x + 15y = 168$, $109x - 40y = -447$, $-3x + y = 12$, and $2x + z = -1$.*

The augmented matrix is

$$\begin{pmatrix} -41 & 15 & 0 & 168 \\ 109 & -40 & 0 & -447 \\ -3 & 1 & 0 & 12 \\ 2 & 0 & 1 & -1 \end{pmatrix}.$$

To solve this multiply the top row by 109, the second row by 41, add the top row to the second row, and multiply the top row by 1/109. This yields

$$\begin{pmatrix} -41 & 15 & 0 & 168 \\ 0 & -5 & 0 & -15 \\ -3 & 1 & 0 & 12 \\ 2 & 0 & 1 & -1 \end{pmatrix}.$$

Now take 2 times the third row and replace the fourth row by this added to 3 times the fourth row.

$$\begin{pmatrix} -41 & 15 & 0 & 168 \\ 0 & -5 & 0 & -15 \\ -3 & 1 & 0 & 12 \\ 0 & 2 & 3 & 21 \end{pmatrix}.$$

Take (-41) times the third row and replace the first row by this added to 3 times the first row. Then switch the third and the first rows.

$$\begin{pmatrix} 123 & -41 & 0 & -492 \\ 0 & -5 & 0 & -15 \\ 0 & 4 & 0 & 12 \\ 0 & 2 & 3 & 21 \end{pmatrix}.$$

Take $-1/2$ times the third row and add to the bottom row. Then take 5 times the third row and add to four times the second. Finally take 41 times the third row and add to 4 times the top row. This yields

$$\begin{pmatrix} 492 & 0 & 0 & -1476 \\ 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 12 \\ 0 & 0 & 3 & 15 \end{pmatrix}$$

It follows $x = \frac{-1476}{492} = -3, y = 3$ and $z = 5$.

You should practice solving systems of equations. Here are some exercises.

2.13 Exercises

1. Give the complete solution to the system of equations, $3x - y + 4z = 6$, $y + 8z = 0$, and $-2x + y = -4$.

2. Give the complete solution to the system of equations, $2x + z = 511$, $x + 6z = 27$, and $y = 1$.

3. Consider the system $-5x + 2y - z = 0$ 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?

4. Give the complete solution to the system of equations, $7x + 14y + 15z = 22$, $2x + 4y + 3z = 5$, and $3x + 6y + 10z = 13$.

5. Give the complete solution to the system of equations, $-5x - 10y + 5z = 0$, $2x + 4y - 4z = -2$, and $-4x - 8y + 13z = 8$.

6. Give the complete solution to the system of equations, $9x - 2y + 4z = -17$, $13x - 3y + 6z = -25$, and $-2x - z = 3$.

7. Give the complete solution to the system of equations, $9x - 18y + 4z = -83$, $-32x + 63y - 14z = 292$, and $-18x + 40y - 9z = 179$.

8. Give the complete solution to the system of equations, $65x + 84y + 16z = 546$, $81x + 105y + 20z = 682$, and $84x + 110y + 21z = 713$.

9. Give the complete solution to the system of equations, $3x - y + 4z = -9$, $y + 8z = 0$, and $-2x + y = 6$.

10. Give the complete solution to the system of equations, $8x + 2y + 3z = -3, 8x + 3y + 3z = -1$, and $4x + y + 3z = -9$.

11. Give the complete solution to the system of equations, $-7x - 14y - 10z = -17$, $2x + 4y + 2z = 4$, and $2x + 4y - 7z = -6$.

12. Give the complete solution to the system of equations, $-8x + 2y + 5z = 18, -8x + 3y + 5z = 13$, and $-4x + y + 5z = 19$.

13. Give the complete solution to the system of equations, $2x + 2y - 5z = 27, 2x + 3y - 5z = 31$, and $x + y - 5z = 21$.

14. Give the complete solution to the system of equations, $3x - y - 2z = 3$, $y - 4z = 0$, and $-2x + y = -2$.

15. Give the complete solution to the system of equations, $3x - y - 2z = 6$, $y - 4z = 0$, and $-2x + y = -4$.

16. Four times the weight of Gaston is 150 pounds more than the weight of Ichabod. Four times the weight of Ichabod is 660 pounds less than seventeen times the weight of Gaston. 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 girls.

17. Give the complete solution to the system of equations, $-19x+8y = -108$, $-71x+30y = -404$, $-2x + y = -12$, $4x + z = 14$.

18. Give the complete solution to the system of equations, $-9x+15y = 66$, $-11x+18y = 79$,$-x + y = 4$, and $z = 3$.

2.14 Completeness of $\mathbb{R}$

By Theorem 2.8.9, 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 2.14.1 *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, l.u.b.(S) or often* $\sup(S)$. *If S is a nonempty set bounded below, define the greatest lower bound, g.l.b.(S) or* $\inf(S)$ *similarly. Thus g is the g.l.b.(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 2.14.2 *(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 2.14.3 *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 $\inf(S)$.

2.15 Review 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. 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

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

9. Using the preceding problem, show the number of ways of selecting a set of k things from a set of n things is $\binom{n}{k}$.

10. Prove the binomial theorem from Problem 9. **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.

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

12. Prove by the binomial theorem and Problem 9 that the number of subsets of a given finite set containing n elements is 2^n.

13. 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.

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

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

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

17. Derive a formula for the multinomial expansion, $\left(\sum_{k=1}^{p} a_k\right)^n$ which is analogous to the binomial expansion. **Hint:** See Problem 10.

18. Suppose $a > 0$ and that x is a real number which satisfies the quadratic equation,

$$ax^2 + bx + c = 0.$$

Find a formula for x in terms of a and b and square roots of expressions involving these numbers. **Hint:** First divide by a to get

$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0.$$

Then add and subtract the quantity $b^2/4a^2$. Verify that

$$x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} = \left(x + \frac{b}{2a}\right)^2.$$

Now solve the result for x. The process by which this was accomplished in adding in the term $b^2/4a^2$ is referred to as completing the square. You should obtain the quadratic formula[7],

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

[7]The ancient Babylonians knew how to solve these quadratic equations sometime before 1700 B.C. It seems they used pretty much the same process outlined in this exercise.

The expression $b^2 - 4ac$ is called the discriminant. When it is positive there are two different real roots. When it is zero, there is exactly one real root and when it equals a negative number there are no real roots.

19. Suppose $f(x) = 3x^2 + 7x - 17$. Find the value of x at which $f(x)$ is smallest by completing the square. Also determine $f(\mathbb{R})$ and sketch the graph of f. **Hint:**

$$\begin{aligned} f(x) &= 3\left(x^2 + \frac{7}{3}x - \frac{17}{3}\right) = 3\left(x^2 + \frac{7}{3}x + \frac{49}{36} - \frac{49}{36} - \frac{17}{3}\right) \\ &= 3\left(\left(x + \frac{7}{6}\right)^2 - \frac{49}{36} - \frac{17}{3}\right). \end{aligned}$$

20. 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. Can you conjecture and prove a result about $y = ax^2 + bx + c$ in terms of the sign of a based on these last two problems?

21. 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 $n + 1$?

Basic Geometry And Trigonometry

3.0.1 Outcomes

1. Understand basic geometry and similar triangles. Be able to give some proof of the Pythagoras theorem from more self evident principles.

2. Understand the distance formula and its relation to the Pythagoras theorem and the law of cosines.

3. Understand the length of a circular arc and its radian measure.

4. Understand the trigonometric functions and their properties, including the important identities for the sum and difference of two angles.

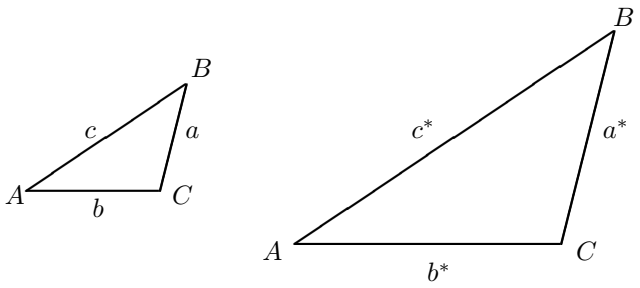
5. Understand basic area formulas including areas of triangles, parallelograms and circular sectors.

6. Define the parabola, hyperbola, and ellipse. Be able to work with these ideas.

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. [10]

3.1 Similar Triangles And Pythagorean Theorem

Definition 3.1.1 *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.*



The fundamental axiom for similar triangles is the following.

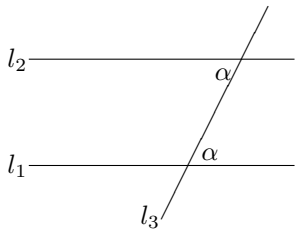
Axiom 3.1.2 *If two triangles are similar then the ratios of corresponding parts are the same.*

For example in the above picture, this says that

$$\frac{a}{b} = \frac{a^*}{b^*}$$

Definition 3.1.3 *Two lines in the plane are said to be parallel if no matter how far they are extended, they never intersect.*

Definition 3.1.4 *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 α.*



As suggested by the above picture, the following axiom will be used.

Axiom 3.1.5 *If l_1 and l_2 are parallel lines intersected by l_3, then alternate interior angles are equal.*

Definition 3.1.6 *An angle is a right angle if when either side is extended, the new angle formed by the extension equals the original angle.*

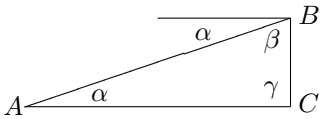
Axiom 3.1.7 *Suppose l_1 and l_2 both intersect a third line, l_3 in a right angle. Then l_1 and l_2 are parallel.*

Definition 3.1.8 *A right triangle is one in which one of the angles is a right angle.*

Axiom 3.1.9 *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 3.1.10 *Let α, β, and γ be the angles of a right triangle with γ the right angle. Then if the angles, α and β 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 3.1.9. 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 3.1.7 this line is parallel to the line joining A and B and by Axiom 3.1.5 the angle between the line joining A and B and this new line is α as shown in the picture. Therefore, the angle formed by placing α and β together is a right angle as claimed.

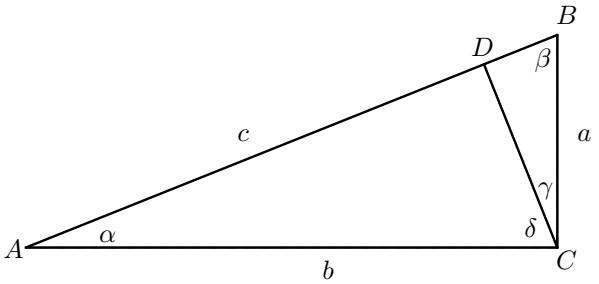
Definition 3.1.11 *When an angle α is placed next to an angle β as shown above, then the resulting angle is denoted by $\alpha + \beta$. A right angle is said to have 90° or to be a 90° angle.*

With this definition, Theorem 3.1.10 says the sum of the two non 90° angles in a right triangle is 90°.

In a right triangle the long side is called the hypotenuse. The similar triangles axiom can be used to prove the Pythagorean theorem.

Theorem 3.1.12 *(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 $\overline{DB}$ the length of the line from D to B.



Then from Theorem 3.1.10, $\delta + \gamma = 90°$ and $\beta + \gamma = 90°$. Therefore, $\delta = \beta$. Also from this same theorem, $\alpha + \delta = 90°$ and so $\alpha = \gamma$. Therefore, the three triangles shown in the picture are all similar. By Axiom 3.1.2,

$$\frac{c}{a} = \frac{a}{\overline{DB}}, \text{ and } \frac{c}{b} = \frac{b}{c - \overline{DB}}.$$

Therefore, $c\overline{DB} = a^2$ and

$$c\left(c - \overline{DB}\right) = b^2$$

so

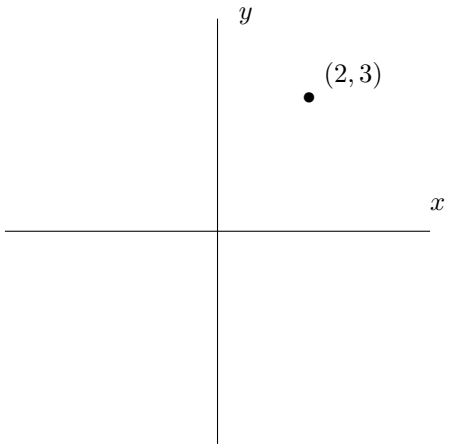
$$\begin{aligned} c^2 &= c\overline{DB} + b^2 \\ &= a^2 + b^2. \end{aligned}$$

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 $\mathbb{R}$.

3.2 Cartesian Coordinates And Straight Lines

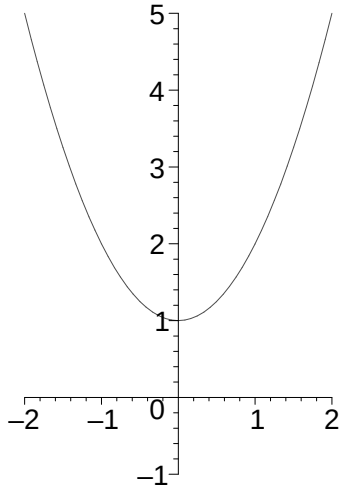
Recall the notion of the Cartesian coordinate system. It involved an x axis, a y axis, two lines which intersect each other at right angles and one identifies a point by specifying a pair of numbers. For example, the number $(2,3)$ involves going 2 units to the right on the x axis and then 3 units directly up on a line perpendicular to the x axis. For example, consider the following picture.



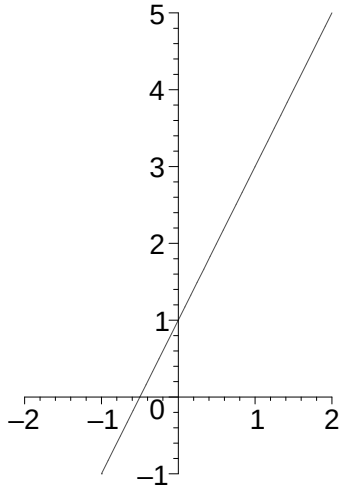
Because of the simple correspondence between points in the plane and the coordinates of a point in the plane, it is often the case that people are a little sloppy in referring to these things. Thus, it is common to see (x,y) referred to as a point in the plane. I will often indulge in this sloppiness.

The reader has likely encountered the notion of graphing relations of the form $y = 2x+3$ or $y = x^2+5$. Recall that you first found lots of ordered pairs which satisfied the relation. For example $(0,3)$,$(1,5)$, and $(-1,1)$ all satisfy the first relation which describes a straight line. Here are some simple examples which you should see that you understand. First here is the graph of $y = x^2+1$.

[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.



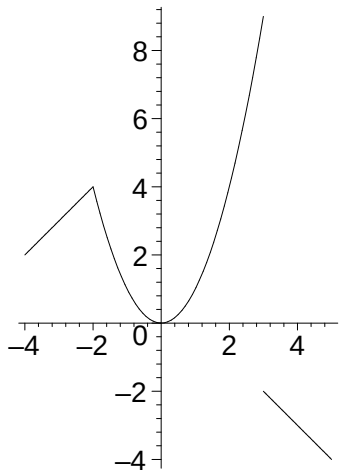
Now here is the graph of the relation $y = 2x + 1$ which is a straight line.



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

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

Then the graph of this relation is sketched below.



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

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

Remember 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.

Example 3.2.1 *Find the relation for a straight line which contains the point* $(1, 2)$ *and has constant slope equal to* 3.

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

3.3 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.

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.

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

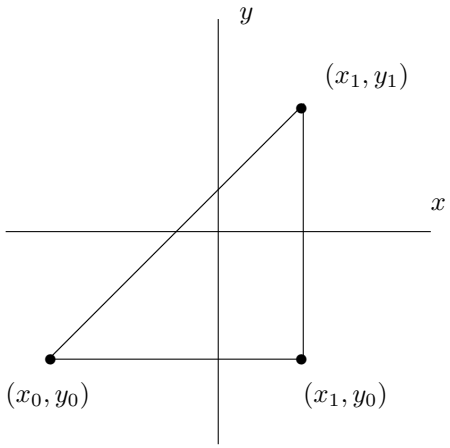
17. Two lines are parallel if they have the same slope. Find the equation of the line through the point $(2,3)$ which is parallel to the line whose equation is $2x+3y=8$.

18. Sketch the graph of the relation defined as

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

3.4 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.



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

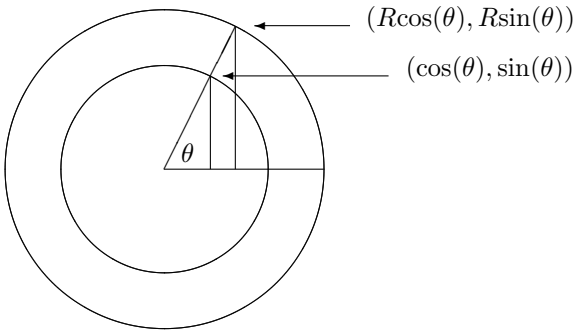
$$\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$$

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 3.1.10 on Page 51 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 3.4.1 *For θ 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 intersection, is given as* $(R\cos\theta, R\sin\theta)$. *In particular, this specifies* $\cos\theta$ *and* $\sin\theta$ *by simply letting* $R = 1$.

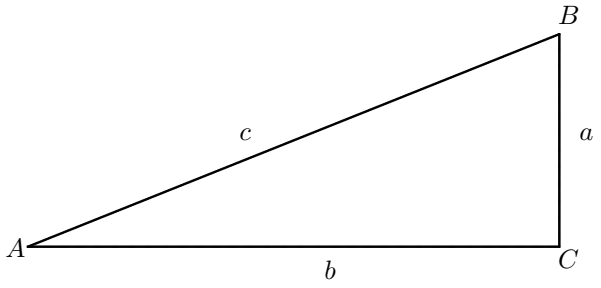
Proposition 3.4.2 *For any angle, θ,* $\cos^2\theta + \sin^2\theta = 1$.

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.

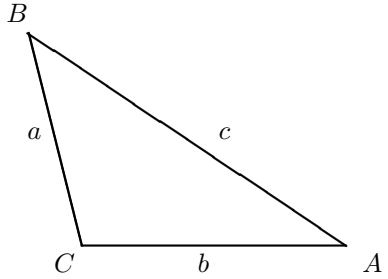
$$\tan\theta \equiv \frac{\sin\theta}{\cos\theta}, \cot\theta \equiv \frac{\cos\theta}{\sin\theta}, \sec\theta \equiv \frac{1}{\cos\theta}, \csc\theta \equiv \frac{1}{\sin\theta}. \tag{3.1}$$

It is also important to understand these functions in terms of a right triangle. Consider the following picture of a right triangle.



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 3.4.3 *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 3.4.2,

$$\begin{aligned} c^2 &= (a\cos C - b)^2 + a^2\sin^2 C \\ &= a^2\cos^2 C - 2ab\cos C + b^2 + a^2\sin^2 C \\ &= a^2 + b^2 - 2ab\cos C \end{aligned}$$

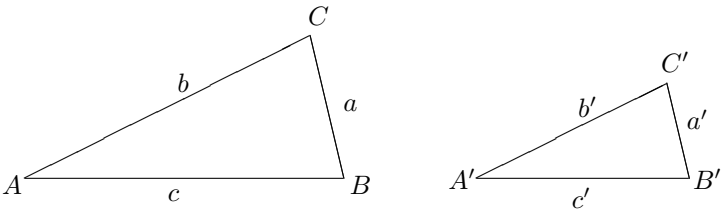
as claimed.

Corollary 3.4.4 *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 3.4.2, $|\cos\theta| \leq 1$ 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 3.4.5 *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,

$$\begin{aligned} a^2 &= b^2 + c^2 - 2bc\cos A \\ &= 2b^2 - 2b^2\cos A \end{aligned}$$

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 isoceles.

3.5 The Circular Arc Subtended By An Angle

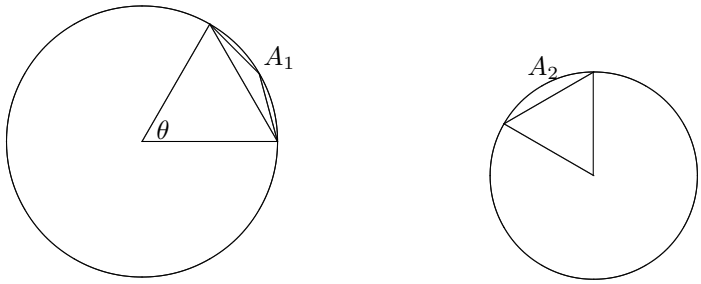
How can angles be measured? This will be done by considering arcs on a circle. To see how this will be done, let θ denote an angle and place the vertex of this angle at the center of

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 2π. In summary, if θ 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.

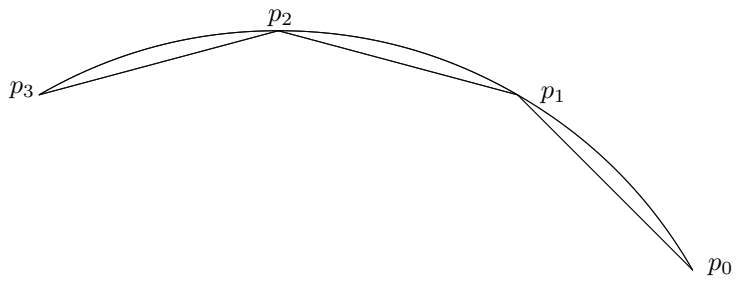


In this picture, there are two circles, a big one having radius, R and a little one having radius r. The angle, θ is situated in two different ways subtending the arcs A_1 and A_2 as shown.

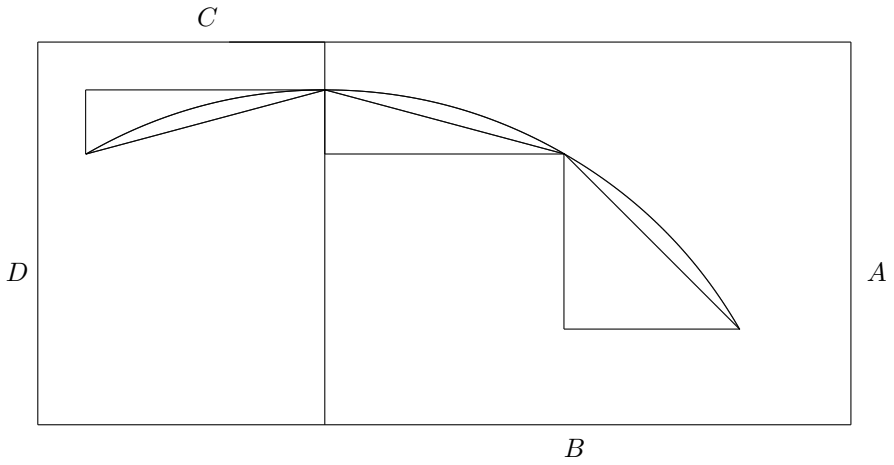
Letting A be an arc of a circle, like those shown in the above picture, A subset of

[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 π as 3. This is not too far off. Later, methods will be given which allow one to calculate π 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 π is transcendental which is the worst sort of irrational number.

$A, \{p_0, \cdots, p_n\}$ 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.

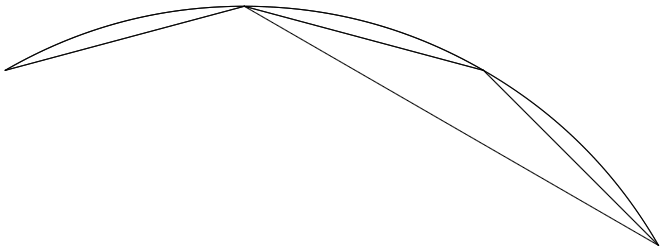


Also, denote by $\mathcal{P}(A)$ the set of all such partitions. For $P = \{p_0, \cdots, p_n\}$, denote by $|p_i - p_{i-1}|$ the distance between p_i and p_{i-1}. Then for $P \in \mathcal{P}(A)$, define $|P| \equiv \sum_{i=1}^{n} |p_i - p_{i-1}|$. Thus $|P|$ consists of the sum of the lengths of the little lines joining successive points of P and appears to be an approximation to the length of the circular arc, A. By Corollary 3.4.4 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.

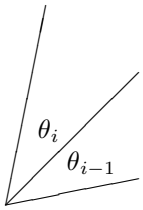


The sum of the lengths of the straight line segments in the part of the picture found in the right rectangle above is less than $A + B$ and the sum of the lengths of the straight line segments in the part of the picture found in the left rectangle above is less than $C + D$ and this would be so for any 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 3.4.4 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, θ_i as follows. The angle θ_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 θ_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 θ_i on the center of the circle and let one side of θ_i coincide with the side of the angle θ_{i-1} which is most counter clockwise, the other side of θ_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 $P_1 \in \mathcal{P}(A_1)$ 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 $Q \in \mathcal{P}(A_2)$ 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 $P_1' \in \mathcal{P}(A_1)$. 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 3.4.5

$$\frac{|P_1'|}{|P_2'|} = \frac{R}{r}$$

and so

$$l(A_2) < |P_2'| + \varepsilon = \frac{r}{R}|P_1'| + \varepsilon \leq \frac{r}{R} l(A_1) + \varepsilon.$$

Since ε 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 ε is arbitrary that $Rl(A_2) \geq rl(A_1)$ and this has proved the following theorem.

Theorem 3.5.1 *Let θ 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 $l(A_1) \leq \frac{R}{r} l(A_2)$. 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 ε 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.

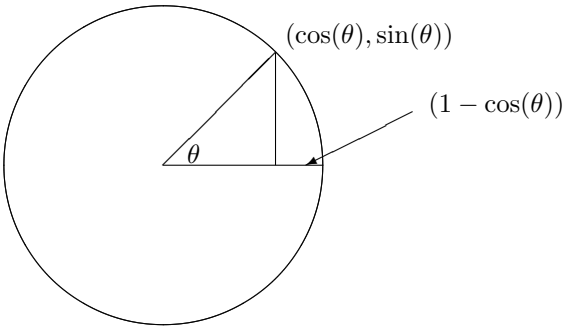
Definition 3.5.2 *Let θ be an angle. The measure of θ is defined to be the length of the circular arc subtended by θ on a circle of radius r divided by r. This is also called the radian measure of the angle.*

You should note the measure of θ is independent of dimension. This is because the units of length cancel when the division takes place.

Proposition 3.5.3 *The above definition is well defined and also, if A is an arc subtended by the angle θ 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 3.5.1. The formula also follows from Theorem 3.5.1 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.

Corollary 3.5.4 *Let $0 <$ radian measure of $\theta < \pi/4$. Then letting A be the arc on the unit circle resulting from situating the angle with one side on the positive x axis and the other side pointing up from the positive x axis,*

$$(1 - \cos\theta) + \sin\theta \geq l(A) \geq \sin\theta \tag{3.2}$$

Proof: Situate the angle, θ such that one side is on the positive x axis and extend the other side till it intersects the unit circle at the point, $(\cos\theta, \sin\theta)$. Then denoting the resulting arc on the circle by A, it follows that for all $P \in \mathcal{P}(A)$ the inequality $(1 - \cos\theta) + \sin\theta \geq |P| \geq \sin\theta$. It follows that $(1 - \cos\theta) + \sin\theta$ is an upper bound for all the $|P|$ where $P \in \mathcal{P}(A)$ and so $(1 - \cos\theta) + \sin\theta$ is at least as large as the sup or least upper bound of the

$|P|$. This proves the top half of the inequality. The bottom half follows because $l(A) \geq L$ where L is the length of the line segment joining $(\cos\theta, \sin\theta)$ and $(1,0)$ due to the definition of $l(A)$. However, $L \geq \sin\theta$ because L is the length of the hypotenuse of a right triangle having $\sin\theta$ as one of the sides.

3.6 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 3.6.1 *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 2.8.11 on Page 32 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, $b = (-p)\,2\pi + (-r_1) = (-p-1)\,2\pi + \left(\overbrace{2\pi - r_1}^{\equiv r}\right)$ 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 3.6.2 *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 3.6.3 *Let $k \in \mathbb{Z}$, then the following formulas hold.*

$$\sin b = -\sin(-b), \cos b = \cos(-b), \tag{3.3}$$

$$\sin(b + 2k\pi) = \sin b, \cos(b + 2k\pi) = \cos b \tag{3.4}$$

$$\cos^2 b + \sin^2 b = 1 \tag{3.5}$$

The other trigonometric functions are defined in the usual way as in (3.1) 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 π 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 3.6.4 *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,

$$x + y = 2\pi p + r_{x+y},\ x = 2\pi p_1 + r_x,\ y = 2\pi p_2 + r_y.$$

From this the result follows because

$$0 = ((x+y)-x)-y = 2\pi\overbrace{((p-p_1)-p_2)}^{\equiv -k} + r_{x+y} - (r_x + r_y).$$

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 3.6.5 *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.

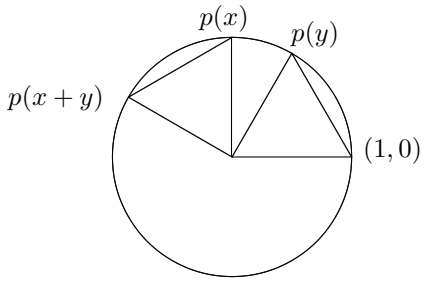
First assume $r_{x+y} \geq r_x$. Then $|r_{x+y} - r_x| = r_{x+y} - r_x = r_y + 2k\pi$. Since both r_{x+y} and r_x are in $[0, 2\pi)$, their difference is also in $[0, 2\pi)$ and so $k = 0$. Therefore, the arc joining $p(x)$ and $p(x+y)$ is of the same length as the arc joining $p(y)$ and $(1,0)$. In 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 2π, 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 3.6.6 *Let $x, y \in \mathbb{R}$. Then*

$$\cos(x+y)\cos y + \sin(x+y)\sin y = \cos x. \tag{3.6}$$

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.



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 3.4.5 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 3.6.3 this implies (3.6). This proves the theorem.

Letting $y=\pi/2$, this shows that

$$\sin(x+\pi/2)=\cos x. \tag{3.7}$$

Now let $u=x+y$ and $v=y$. Then (3.6) implies

$$\cos u\cos v+\sin u\sin v=\cos(u-v) \tag{3.8}$$

Also, from this and (3.3),

$$\begin{aligned}\cos(u+v)&=\cos(u-(-v))\\&=\cos u\cos(-v)+\sin u\sin(-v)\\&=\cos u\cos v-\sin u\sin v\end{aligned} \tag{3.9}$$

Thus, letting $v=\pi/2$,

$$\cos\left(u+\frac{\pi}{2}\right)=-\sin u. \tag{3.10}$$

It follows

$$\begin{aligned}\sin(x+y)&=-\cos\left(x+\frac{\pi}{2}+y\right)\\&=-\left[\cos\left(x+\frac{\pi}{2}\right)\cos y-\sin\left(x+\frac{\pi}{2}\right)\sin y\right]\\&=\sin x\cos y+\sin y\cos x\end{aligned} \tag{3.11}$$

Then using Observation 3.6.3 again, this implies

$$\sin(x-y)=\sin x\cos y-\cos x\sin y. \tag{3.12}$$

In addition to this, Observation 3.6.3 implies

$$\cos 2x=\cos^2 x-\sin^2 x \tag{3.13}$$
$$=2\cos^2 x-1 \tag{3.14}$$
$$=1-2\sin^2 x \tag{3.15}$$

Therefore, making use of the above identities, and Observation 3.6.3,

$$\begin{aligned}\cos(3x)&=\cos 2x\cos x-\sin 2x\sin x\\&=\left(2\cos^2 x-1\right)\cos x-2\cos x\sin^2 x\\&=4\cos^3 x-3\cos x\end{aligned} \tag{3.16}$$

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 do isn't it equal to $-\sqrt{2}/2$? **Hint:** Draw a picture.) Thus $\sin\left(\frac{\pi}{4}\right)=\sqrt{2}/2$ also. (Why?) Here is another one. From (3.16),

$$\begin{aligned}0=\cos\left(\frac{\pi}{2}\right)&=\cos 3\left(\frac{\pi}{6}\right)\\&=4\cos^3\left(\frac{\pi}{6}\right)-3\cos\left(\frac{\pi}{6}\right).\end{aligned}$$

Therefore, $\cos\left(\frac{\pi}{6}\right) = \frac{\sqrt{3}}{2}$ and consequently, $\sin\left(\frac{\pi}{6}\right) = \frac{1}{2}$. Here is a short table including these and a few others. You should make sure you can obtain all these entries. In the table, θ refers to the radian measure of the angle. From now on, angles are considered as real numbers, not as pointy things.

θ	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$
$\cos\theta$	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0
$\sin\theta$	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1

3.7 Exercises

1. Find $\cos\theta$ and $\sin\theta$ for $\theta \in \left\{\frac{2\pi}{3}, \frac{3\pi}{4}, \frac{5\pi}{6}, \pi, \frac{7\pi}{6}, \frac{5\pi}{4}, \frac{4\pi}{3}, \frac{3\pi}{2}, \frac{5\pi}{3}, \frac{7\pi}{4}, \frac{11\pi}{6}, 2\pi\right\}$.
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, $\cos(\pi/12) = \sqrt{\frac{1+(\sqrt{3}/2)}{2}}$. 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. Using Problem 2 graph $y = \cos^2 x$.

16. If $f(x) = A\cos\alpha x + B\sin\alpha x$, show there exists ϕ such that

$$f(x) = \sqrt{A^2 + B^2}\sin(\alpha x + \phi).$$

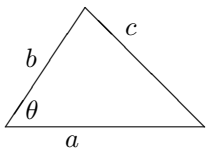
Show there also exists ψ such that $f(x) = \sqrt{A^2 + b^2}\cos(\alpha x + \psi)$. This is a very important result, enough that some of these quantities are given names. $\sqrt{A^2 + B^2}$ is called the amplitude and ϕ or ψ are called phase shifts.

17. Using Problem 16 graph $y = \sin x + \sqrt{3}\cos x$.

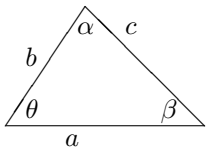
18. Give all solutions to $\sin x + \sqrt{3}\cos x = \sqrt{3}$. **Hint:** Use Problem 17.

19. 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.

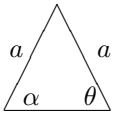
20. In the picture, $a = 5, b = 3$, and $\theta = \frac{2}{3}\pi$. Find c.



21. In the picture, $\theta = \frac{1}{4}\pi$, $\alpha = \frac{2}{3}\pi$ and $c = 3$. Find a.



22. An isoceles triangle is one which has two equal sides. For example the following picture is of an isoceles triangle



the two equal sides having length a. Show the "base angles" θ and α are equal. **Hint:** You might want to use the law of sines.

23. Find a formula for $\tan(\theta+\beta)$ in terms of $\tan\theta$ and $\tan\beta$

24. Find a formula for $\tan(2\theta)$ in terms of $\tan\theta$.

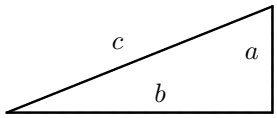
25. Find a formula for $\tan\left(\frac{\theta}{2}\right)$ in terms of $\tan\theta$.

26. 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 $\tan\beta = \frac{1}{5}$, then $4\beta+\theta = \frac{\pi}{4}$. This is the basis for a wonderful formula which has been used to compute π for hundreds of years.

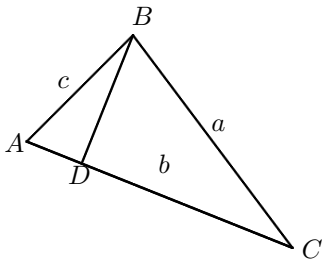
3.8 Some Basic Area Formulas

3.8.1 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.

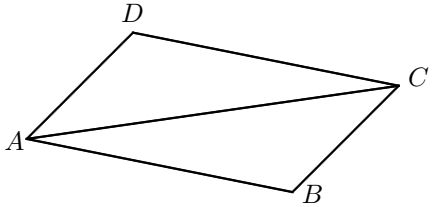


The area of this triangle shown above must equal $ab/2$ because it is half of a rectangle having sides a and b. Now consider a general triangle in which a line perpendicular to the line from A to C has been drawn through B.



The area of this triangle would be the sum of the two right triangles formed. Thus this area would be $\frac{1}{2}\left(\overline{BD}\right)\left(\overline{AD}+\overline{CD}\right) = \frac{1}{2}\left(\overline{BD}\right)b$. In words, the area of the triangle 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 $2\overline{AB}$ (height of parallelogram) $\frac{1}{2} = \overline{AB}$ (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.

3.8.2 The Area Of A Circular Sector

Consider an arc, A, of a circle of radius r which subtends an angle, θ. 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 3.8.1 *Let $1 > \varepsilon > 0$ be given. Then whenever the positive number, α, is small enough,*

$$1 \leq \frac{\alpha}{\sin\alpha} \leq 1+\varepsilon \tag{3.17}$$

and

$$1+\varepsilon \geq \frac{\alpha}{\tan\alpha} \geq 1-\varepsilon \tag{3.18}$$

Proof: This follows from Corollary 3.5.4 on Page 62. In this corollary, $l(A) = \alpha$ and so

$$1-\cos\alpha+\sin\alpha \geq \alpha \geq \sin\alpha.$$

Therefore, dividing by $\sin\alpha$,

$$\frac{1-\cos\alpha}{\sin\alpha}+1 \geq \frac{\alpha}{\sin\alpha} \geq 1. \tag{3.19}$$

Now using the properties of the trig functions,

$$\begin{aligned}\frac{1-\cos\alpha}{\sin\alpha} &= \frac{1-\cos^2\alpha}{\sin\alpha(1+\cos\alpha)}\\ &= \frac{\sin^2\alpha}{\sin\alpha(1+\cos\alpha)} = \frac{\sin\alpha}{1+\cos\alpha}.\end{aligned}$$

From the definition of the sin and cos, whenever α is small enough,

$$\frac{\sin\alpha}{1+\cos\alpha} < \varepsilon$$

and so (3.19) implies that for such α, (3.17) holds. To obtain (3.18), let α be small enough that (3.17) holds and multiply by $\cos\alpha$. Then for such α,

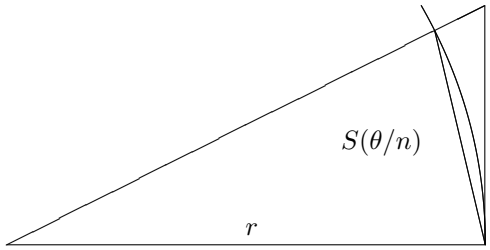
$$\cos\alpha \leq \frac{\alpha}{\tan\alpha} \leq (1+\varepsilon)\cos\alpha$$

Taking α smaller if necessary and noting that for all α small enough, $\cos\alpha$ is very close to 1, yields (3.18). This proves the lemma.

This lemma is very important in another context.

Theorem 3.8.2 *Let $S(\theta)$ denote the sector of a circle of radius r having central angle, θ. Then the area of $S(\theta)$ equals $\frac{r^2}{2}\theta$.*

Proof: Let the angle which A subtends be denoted by θ and divide this sector into n equal sectors each of which has a central angle equal to θ/n. The following is a picture of one of these.



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 θ 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,

$$\frac{r^2}{2}\theta\frac{\sin(\theta/n)}{(\theta/n)}, \ \frac{r^2}{2}\theta\frac{\tan(\theta/n)}{(\theta/n)}.$$

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 ε is an arbitrary small positive number, it follows the area of the sector equals $\frac{r^2}{2}\theta$ as claimed. (Why?)

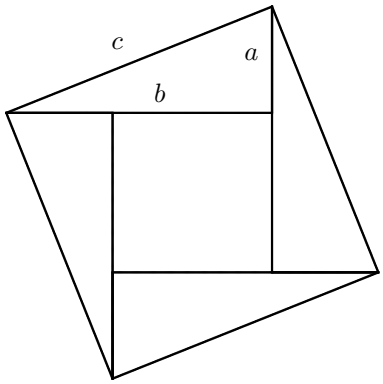
3.9 Exercises

1. Give another argument which verifies the Pythagorean theorem by supplying the details for the following argument[3]. 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

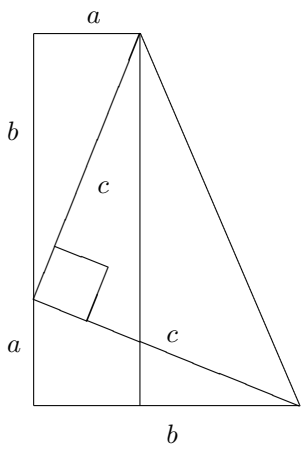
[3]This argument is old and was known to the Indian mathematician Bhaskar who lived 1114-1185 A.D.

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,

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



2. Another very simple and convincing proof of the Pythagorean theorem[4] 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. 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

[4]This argument involving the area of a trapezoid is due to James Garfield who was one of the presidents of the United States.

of the cone and while the paint is still wet, rolling it on the floor yielding a circular sector.

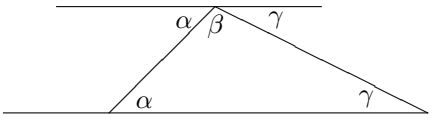
4. An equilateral triangle is one in which all sides are of equal length. Find the area of an equilateral triangle whose sides have length l.

5. 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?

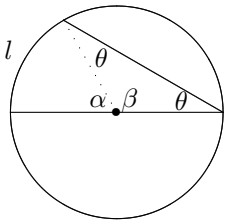
6. Explain why the area of a circle of radius r is πr^2.

7. Explain why through any point in the plane there exists a line parallel to a given line in the plane.

8. Explain why the sum of the radian measures of the angles in any triangle equals π. **Hint:** Consider the following picture and use the result of Problem 7.



9. The following picture is of an "inscribed angle", denoted by θ in a circle of radius a. Drawing a line from the center as shown in the picture, it follows from Problem 22 on Page 67 the two base angles are equal. These are denoted as θ in the picture.



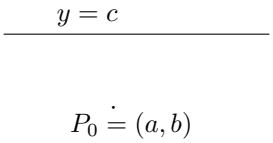
Now the radian measure of α is l/a. Using the result of Problem 8 show the radian measure of θ equals $l/2a$.

10. The inscribed angle in Problem 9 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 9.

3.10 Parabolas, Ellipses, and Hyperbolas

3.10.1 The Parabola

A parabola is a collection of points, P in the plane such that the distance from P to a fixed line is the same as the distance from P to a given point, P_0. From this definition, one can obtain an equation which will describe a parabola. Suppose then that the line is $y = c$ and the point is (a, b) where $b \neq c$ as shown in the picture.



The distance from the point, $P = (x, y)$ to the line is $|c - y|$. Therefore, the description of the parabola requires that

$$\sqrt{(x-a)^2 + (y-b)^2} = |c - y|.$$

Squaring both sides,

$$x^2 - 2xa + a^2 + y^2 - 2yb + b^2 = c^2 - 2cy + y^2$$

and so

$$(x-a)^2 + b^2 - c^2 = (2b - 2c)\, y. \tag{3.20}$$

The simplest case is when $a = 0$ and $b = -c$. Then in this case, it reduces to

$$x^2 = -4cy$$

and the directix is $y = c$ while the focus is $(0, -c)$.

Now consider an arbitrary equation of the form, $y = dx^2 + ex + f$ where $d \neq 0$. By this is meant the set of points (x, y) such that the equation holds. Such a set of points always is a parabola. To see this, complete the square on the right as follows:

$$\begin{aligned} y &= dx^2 + ex + f \\ &= d\left(x^2 + \frac{e}{d}x + \frac{f}{d}\right) \\ &= d\left(x^2 + \frac{e}{d}x + \frac{e^2}{4d^2}\right) - \frac{e^2 d}{4d^2} \\ &= d\left(x - \left(\frac{-e}{2d}\right)\right)^2 - \frac{e^2 d}{4d^2}. \end{aligned}$$

Therefore, letting $a = \frac{-e}{2d}$,

$$\frac{1}{d}y = (x-a)^2 - \frac{e^2}{4d^2}$$

Now you can show that there exists numbers, b and c such that

$$-\frac{e^2}{4d^2} = b^2 - c^2, \ \frac{1}{d} = 2b - 2c. \tag{3.21}$$

Then the above equation reduces to (3.20).

The line, $y = c$ is called the directrix and the point, P_0 in the above is called the focus. Exactly similar results occur if the directrix is of the form $x = c$ and by similar arguments to those above, the set of points in the plane satisfying $ay^2 + by + c = x$ is also a parabola.

Example 3.10.1 *Find the focus and directix of the parabola* $2x^2 - 3x + 1 = 5y$.

First complete the square on the left. Thus $2\left(x^2-\frac{3}{2}x+\frac{9}{16}\right)-\frac{9}{8}+1=5y$. this yields

$$2\left(x-\frac{3}{4}\right)^2=5y+\frac{1}{8}=5\left(y+\frac{1}{40}\right).$$

Dividing both sides by 2,

$$\left(x-\frac{3}{4}\right)^2=\frac{5}{2}\left(y+\frac{1}{40}\right).$$

Now if this were just $x^2=-4cy$, you would know the directix 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

$$u^2=-4\left(\frac{-5}{8}\right)v \tag{3.22}$$

and so the focus for this parabola in the uv plane is $\left(0,\frac{5}{8}\right)$ and its directix is $v=-\frac{5}{2}$. Since $x=u+\frac{3}{4}$ and $y=v-\frac{1}{40}$, if follows the focus of the parabola in the xy plane is $\left(\frac{3}{4},\frac{5}{8}-\frac{1}{40}\right)=\left(\frac{3}{4},\frac{3}{5}\right)$. Similarly, the directix is $y=\frac{-5}{8}-\frac{1}{40}=-\frac{13}{20}$

Example 3.10.2 *Find the focus and directix of the parabola* $x=3y^2+2y+1$.

This is just like the above. You just switch the roles of x and y. Complete the square on the right to get

$$\begin{aligned} x &= 3\left(y^2+\frac{2}{3}y+\frac{1}{9}\right)-\frac{1}{3}+1 \\ &= 3\left(y+\frac{1}{3}\right)^2+\frac{2}{3} \end{aligned}$$

and so

$$-4\left(\frac{-1}{12}\right)\left(x-\frac{2}{3}\right)=\left(y+\frac{1}{3}\right)^2. \tag{3.23}$$

Therefore, the directix is $x=\frac{2}{3}-\frac{1}{12}=\frac{7}{12}$ and the focus is $\left(\frac{1}{12}+\frac{2}{3},0-\frac{1}{3}\right)=\left(\frac{3}{4},-\frac{1}{3}\right)$.

3.10.2 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 $d(P,P_1)+d(P,P_2)=c$, where c is a fixed positive number. These two points are called the foci of the ellipse. Each is called a focus point by itself. Now one can obtain an equation which will describe an ellipse much as was done with the parabola. Let the two given points be (a,b) and $(a,b+h)$. Let a generic point on the ellipse be (x,y). Then according to the description of an ellipse and the distance formula,

$$\sqrt{(x-a)^2+(y-b)^2}+\sqrt{(x-a)^2+(y-b-h)^2}=c. \tag{3.24}$$

Subtracting $\sqrt{(x-a)^2+(y-b)^2}$ from both sides,

$$\sqrt{(x-a)^2+(y-b-h)^2}=c-\sqrt{(x-a)^2+(y-b)^2}\geq 0.$$

Now squaring both sides yields

$$\begin{aligned}(x-a)^2+(y-b-h)^2 &= c^2-2\sqrt{(x-a)^2+(y-b)^2}c \\ &\quad +(x-a)^2+(y-b)^2.\end{aligned}$$

Therefore,

$$\begin{aligned}(y-b)^2-2h(y-b)+h^2 &= c^2-2\sqrt{(x-a)^2+(y-b)^2}c \\ &\quad +(y-b)^2\end{aligned}$$

and so

$$-2h(y-b)+h^2=c^2-2\sqrt{(x-a)^2+(y-b)^2}c$$

Therefore,

$$-2h(y-b)+h^2-c^2=-2\sqrt{(x-a)^2+(y-b)^2}c$$

Now square both sides again to obtain

$$\left(h^2-c^2\right)^2-4h(y-b)\left(h^2-c^2\right)+4h^2(y-b)^2=4c^2\left((x-a)^2+(y-b)^2\right).$$

Simplifying this yields

$$\left(c^2-h^2\right)(y-b)^2+h(y-b)\left(h^2-c^2\right)+c^2(x-a)^2=\frac{1}{4}\left(c^2-h^2\right)^2$$

which simplifies further to

$$(y-b)^2-h(y-b)+\frac{h^2}{4}+\left(\frac{c^2}{c^2-h^2}\right)(x-a)^2=\frac{1+h^2}{4}$$

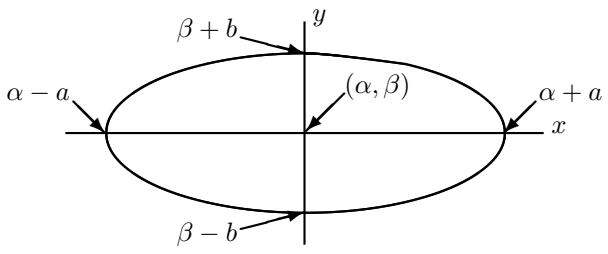
which is equivalent to

$$\frac{\left(y-b-\frac{h}{2}\right)^2}{\left(\frac{1+h^2}{4}\right)}+\frac{(x-a)^2}{\left(\frac{1+h^2}{4}/\frac{c^2}{c^2-h^2}\right)}=1.$$

Thus, redefining the constants, an ellipse has the form

$$\frac{(y-\beta)^2}{b^2}+\frac{(x-\alpha)^2}{a^2}=1. \tag{3.25}$$

Note that if $a=b=r$, this reduces to the equation for a circle of radius r centered at the point (α,β). This last expression is the generic equation for an ellipse. Here is the graph of a typical ellipse.



In this ellipse, $b < a$. If $b > a$, the ellipse would be long in the y direction rather than the x direction. (Why?) Suppose (x_1, y_1) and (x_2, y_2) are two points on the above ellipse in which $b > a$, then $|x_1 - x_2| \leq 2a$ because from the above equation, it follows that $|x_i - \alpha| \leq a$ for $i = 1, 2$ implying that

$$|x_1 - x_2| \leq |x_1 - \alpha| + |\alpha - x_2| \leq a + a = 2a$$

and similarly, $|y_1 - y_2| \leq 2b$. Therefore,

$$|(x_1, y_1) - (x_2, y_2)| \leq \sqrt{4a^2 + 4b^2} \leq 2\sqrt{a^2 + b^2} \leq 2b$$

Thus the greatest distance between two points on the ellipse equals $2b$ and occurs when the two points are $(\alpha, b + \beta)$ and $(\alpha, \beta - b)$. This greatest distance between any two points is called the diameter and this shows the diameter of an ellipse is twice the larger of the two numbers appearing in the denominators on the left in (3.25).

Example 3.10.3 *Find the focus points for the ellipse* $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ *given that* $a^2 \geq b^2$, $a, b > 0$.

Since $a^2 \geq b^2$, the ellipse is at least as long as it is tall and so the focus points will be of the form $(-c, 0)$ and $(c, 0)$. Therefore, the ellipse is of the form

$$\sqrt{(x + c)^2 + y^2} + \sqrt{(x - c)^2 + y^2} = d^2.$$

Thus

$$\begin{aligned} (x + c)^2 + y^2 &= \left(d^2 - \sqrt{(x - c)^2 + y^2}\right)^2 \\ &= d^4 - 2d^2\sqrt{(x - c)^2 + y^2} + (x - c)^2 + y^2 \end{aligned}$$

and so

$$x^2 + 2cx + c^2 + y^2 = d^4 - 2d^2\sqrt{(x - c)^2 + y^2} + x^2 - 2cx + c^2 + y^2$$

which implies

$$4cx = d^4 - 2d^2\sqrt{(x - c)^2 + y^2}$$

and so $d^4 - 4cx = 2d^2\sqrt{x^2 - 2cx + c^2 + y^2}$. Squaring both sides again,

$$d^8 - 8d^4cx + 16c^2x^2 = 4d^4\left(x^2 - 2cx + c^2 + y^2\right).$$

Simplifying this some more,

$$d^8 + \left(16c^2 - 4d^4\right)x^2 - 4d^4y^2 = 4d^4c^2.$$

Therefore,

$$d^4\left(d^4 - 4c^2\right) = 4d^4y^2 + 4\left(d^4 - 4c^2\right)x^2$$

and so

$$1 = \frac{4d^4y^2}{d^4\left(d^4 - 4c^2\right)} + \frac{4\left(d^4 - 4c^2\right)x^2}{d^4\left(d^4 - 4c^2\right)} = \frac{4y^2}{\left(d^4 - 4c^2\right)} + \frac{4x^2}{d^4}$$

Therefore, you need

$$\frac{4}{d^4} = \frac{1}{a^2}, \ \frac{4}{\left(d^4 - 4c^2\right)} = \frac{1}{b^2}.$$

It follows $d^4 = 4a^2$ and $\left(d^4 - 4c^2\right) = 4b^2$. Therefore, $\left(4a^2 - 4c^2\right) = 4b^2$ and solving for c gives $c = \sqrt{a^2 - b^2}$. Therefore, the focus points are $\left(\sqrt{a^2 - b^2}, 0\right)$ and $\left(-\sqrt{a^2 - b^2}, 0\right)$.

Example 3.10.4 *Find the focus points for the ellipse* $\frac{(x-1)^2}{9}+\frac{(y-2)^2}{4}=1$.

From Example 3.10.5 you would see what to do if this were of the form $\frac{x^2}{9}+\frac{y^2}{4}=1$ so simple change the variables. Let $u=x-1$ and $v=y-2$. Then the focus points of the identical ellipse in the uv plane would be $\left(\sqrt{5},0\right)$ and $\left(-\sqrt{5},0\right)$. Now $x=u+1, y=v+2$ so the focus points of the original ellipse in the xy plane are $\left(\sqrt{5}+1,2\right)$ and $\left(-\sqrt{5}+1,2\right)$.

3.10.3 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 $d(P,P_1)-d(P,P_2)=c$, where c is a fixed positive number. These two points are called the foci of the hyperbola. Each is called a focus point by itself. Now one can obtain an equation which will describe a hyperbola. Let the two given points be (a,b) and $(a,b+h)$. Let a generic point on the hyperbola be (x,y). Then according to the description of a hyperbola and the distance formula,

$$\sqrt{(x-a)^2+(y-b)^2}-\sqrt{(x-a)^2+(y-b-h)^2}=c.$$

You can now show that the equation of a hyperbola is of the form

$$\frac{(x-\alpha)^2}{a^2}-\frac{(y-\beta)^2}{b^2}=1 \tag{3.26}$$

or

$$\frac{(y-\beta)^2}{b^2}-\frac{(x-\alpha)^2}{a^2}=1. \tag{3.27}$$

If you like, you can simply take these last two equations as the definition of a hyperbola.

Example 3.10.5 *Find the focus points of the hyperbola* $\frac{x^2}{a^2}-\frac{y^2}{b^2}=1$.

First note that x cannot be equal to zero. Therefore, the hyperoblas open in the direction of the x axis and so the focus points are of the form, $(-c,0)$ and $(c,0)$. Letting (x,y) be the coordinates of a point on the hyperbola, $\sqrt{(x+c)^2+y^2}-\sqrt{(x-c)^2+y^2}=d$. Then $\sqrt{(x+c)^2+y^2}=d+\sqrt{(x-c)^2+y^2}$ and squaring both sides,

$$x^2+2cx+c^2+y^2=d^2+2d\sqrt{(x-c)^2+y^2}+x^2-2cx+c^2+y^2$$

and so

$$4cx-d^2=2d\sqrt{(x-c)^2+y^2}.$$

Now squaring both sides,

$$16c^2x^2-8cd^2x+d^4=4d^2\left(x^2-2cx+c^2+y^2\right)$$

and so

$$16c^2x^2-4d^2x^2-8cd^2x+8d^2cx-4d^2y^2=4d^2c^2-d^4$$

which yields

$$16c^2x^2-4d^2x^2-4d^2y^2=4d^2c^2-d^4$$

It is supposed to reduce to $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$. Therefore, divide both sides by $4d^2c^2 - d^4$. This yields

$$\frac{\overbrace{\left(16c^2 - 4d^2\right)x^2}^{=4/d^2}}{4d^2c^2 - d^4} - \frac{4y^2}{4c^2 - d^2} = 1.$$

Now you need to have

$$\frac{4}{d^2} = \frac{1}{a^2}, \ \frac{4}{4c^2 - d^2} = \frac{1}{b^2}.$$

Therefore, $d^2 = 4a^2$ and $4c^2 - d^2 = 4b^2$ so $4c^2 = 4b^2 + 4a^2$ and so $c = \sqrt{a^2 + b^2}$. Therefore, the focus points are $\left(\sqrt{a^2 + b^2}, 0\right)$ and $\left(-\sqrt{a^2 + b^2}, 0\right)$.

Example 3.10.6 *Find the focus points for the hyperbola* $\frac{(x-1)^2}{4} - \frac{(y-2)^2}{9} = 1$.

From Example 3.10.5 you would see what to do if this were of the form $\frac{x^2}{4} - \frac{y^2}{9} = 1$ so simple 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 $\left(\sqrt{13}, 0\right)$ and $\left(-\sqrt{13}, 0\right)$. Now $x = u+1, y = v+2$ so the focus points of the original hyperbola in the xy plane are $\left(\sqrt{13}+1, 2\right)$ and $\left(-\sqrt{13}+1, 2\right)$.

3.11 Exercises

1. Consider $y = 2x^2 + 3x + 7$. Find the focus and the directrix of this parabola.
2. Find the numbers, b, c which make (3.21) hold.
3. Derive a similar formula to (3.20) in the case that the directrix is of the form $x = c$.
4. Sketch a graph of the ellipse whose equation is $\frac{(x-1)^2}{4} + \frac{(y-2)^2}{9} = 1$.
5. Sketch a graph of the ellipse whose equation is $\frac{(x-1)^2}{9} + \frac{(y-2)^2}{4} = 1$.
6. Sketch a graph of the hyperbola, $\frac{x^2}{4} - \frac{y^2}{9} = 1$.
7. Sketch a graph of the hyperbola, $\frac{y^2}{4} - \frac{x^2}{9} = 1$.
8. Find a formula for the focus points for a hyperbola of the form $\frac{(x-p)^2}{a^2} - \frac{(y-q)^2}{b^2} = 1$.
9. Show the focus points of a hyperbola of the form $\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1$ are $\left(0, \sqrt{a^2 + b^2}\right)$ and $\left(0, -\sqrt{a^2 + b^2}\right)$.
10. Show the focus points of a ellipse of the form $\frac{y^2}{a^2} + \frac{x^2}{b^2} = 1$ for $a \geq b > 0$ are $\left(0, \sqrt{a^2 - b^2}\right)$ and $\left(0, -\sqrt{a^2 - b^2}\right)$.
11. Find a formula for the focus points for a hyperbola of the form $\frac{(y-p)^2}{a^2} - \frac{(x-q)^2}{b^2} = 1$.
12. 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 $a \geq b$.
13. 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 $b \geq a$.

14. Consider the hyperbola, $\frac{y^2}{4} - \frac{x^2}{9} = 1$. Show that $y = \pm\sqrt{b^2 + \frac{b^2x^2}{a^2}}$. The straight lines $y = \frac{bx}{a}$ and $y = -\frac{bx}{a}$ are called the assymptotes of the hyperbola. Show that for large x, $\sqrt{b^2 + \frac{b^2x^2}{a^2}} - \frac{bx}{a}$ is very small.

15. What is the diameter of the ellipse, $\frac{(x-1)^2}{9} + \frac{(y-2)^2}{4} = 1$?

16. Verify that if the focus points are (a, b) and $(a, b+h)$ or (a, b) and $(a+h, b)$, then the hyperbola determined by this pair of points, has an equation given by one of (3.26) or (3.27).

17. Find the focus points for the hyperbola $\frac{(x-1)^2}{4} - \frac{(y-2)^2}{9} = 1$.

18. Show that the set of points which satisfies either (3.26) or (3.27) is unbounded. (If n is any positive number there exist points (x, y) satisfying the equations given such that $|(x, y)| > n$.)

19. 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. **Hint:** First complete the square.

20. 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. **Hint:** First complete the square.

21. 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. **Hint:** First complete the square.

22. 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. **Hint:** First complete the square.

23. 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. Get rid of the square root signs. Why doesn't this equation resemble the ones discussed above?

24. Find the equation of the parabola which has focus $(0, 0)$ and directix $x + y = 1$. This is pretty hard. To do it you need to figure out how to find the distance between a point and the given line.

25. Find a formula for $\sin x \cos y$ in terms of sines and cosines of $x + y$ and $x - y$.

26. 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

4.0.1 Outcomes

1. Understand the geometric significance of a complex number as a point in the plane.

2. Understand the algebra of complex numbers. In particular, understand why the complex numbers satisfy the field axioms.

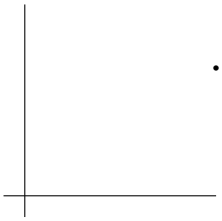
3. Understand the absolute value of a complex number and how to find it as well as its geometric significance.

4. Be able to do computations with complex numbers and understand the conjugate geometrically and algebraically.

5. Understand De'Moivre's theorem and be able to use it to find the roots of a complex number.

This chapter gives a brief treatment of the complex numbers. This will not be needed in Calculus but you will need it when you take differential equations and various other subjects so it is a good idea to consider the subject. These things used to be taught in precalculus classes and people were expected to know them before taking calculus. However, if you are in a hurry to get to calculus, you can skip this short chapter.

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 $(3, 2)$.



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

$$\begin{aligned}(a+ib)(c+id) &= ac+iad+ibc+i^2bd \\ &= (ac-bd)+i(bc+ad).\end{aligned}$$

Every non zero complex number, $a+ib$, with $a^2+b^2 \neq 0$, has a unique multiplicative inverse.

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

You should prove the following theorem.

Theorem 4.0.1 *The complex numbers with multiplication and addition defined as above form a field satisfying all the field axioms listed on Page 18.*

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.

$$\left(\overline{a+ib}\right)(a+ib) = a^2+b^2.$$

Definition 4.0.2 *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, $z=a+ib$,

$$|z| = \left(z\overline{z}\right)^{1/2}.$$

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 4.0.3 *: 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 $(z-w)\left(\overline{z-w}\right) = (1-i3)(1+i3) = 10$ 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

$$\cos\theta = \frac{x}{\sqrt{x^2+y^2}}, \sin\theta = \frac{y}{\sqrt{x^2+y^2}}.$$

The polar form of the complex number is then

$$r\left(\cos\theta + i\sin\theta\right)$$

where θ is this angle just described and $r = \sqrt{x^2+y^2}$.

A fundamental identity is the formula of De Moivre which follows.

Theorem 4.0.4 *Let $r > 0$ be given. Then if n is a positive integer,*

$$\left[r\left(\cos t + i\sin t\right)\right]^n = r^n\left(\cos nt + i\sin nt\right).$$

Proof: It is clear the formula holds if $n = 1$. Suppose it is true for n.

$$\left[r\left(\cos t + i\sin t\right)\right]^{n+1} = \left[r\left(\cos t + i\sin t\right)\right]^n\left[r\left(\cos t + i\sin t\right)\right]$$

which by induction equals

$$= r^{n+1}\left(\cos nt + i\sin nt\right)\left(\cos t + i\sin t\right)$$

$$= r^{n+1}\left(\left(\cos nt\cos t - \sin nt\sin t\right) + i\left(\sin nt\cos t + \cos nt\sin t\right)\right)$$

$$= r^{n+1}\left(\cos\left(n+1\right)t + i\sin\left(n+1\right)t\right)$$

by the formulas for the cosine and sine of the sum of two angles.

Corollary 4.0.5 *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 $z = |z|\left(\cos t + i\sin t\right)$ be the polar form of the complex number. By De Moivre's theorem, a complex number,

$$r\left(\cos\alpha + i\sin\alpha\right),$$

is a k^{th} root of z if and only if

$$r^k\left(\cos k\alpha + i\sin k\alpha\right) = |z|\left(\cos t + i\sin t\right).$$

This requires $r^k = |z|$ and so $r = |z|^{1/k}$ and also both $\cos\left(k\alpha\right) = \cos t$ and $\sin\left(k\alpha\right) = \sin t$. 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 2π, there are exactly k distinct numbers which result from this formula.

Example 4.0.6 *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

$$\cos\left(\frac{\pi}{6}\right) + i\sin\left(\frac{\pi}{6}\right), \cos\left(\frac{5}{6}\pi\right) + i\sin\left(\frac{5}{6}\pi\right),$$

and

$$\cos\left(\frac{3}{2}\pi\right) + i\sin\left(\frac{3}{2}\pi\right).$$

Thus the cube roots of i are $\frac{\sqrt{3}}{2} + i\left(\frac{1}{2}\right)$, $\frac{-\sqrt{3}}{2} + i\left(\frac{1}{2}\right)$, and $-i$.

The ability to find k^{th} roots can also be used to factor some polynomials.

Example 4.0.7 *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 $3, 3\left(\frac{-1}{2} + i\frac{\sqrt{3}}{2}\right)$, 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 $\left(x - 3\left(\frac{-1}{2} + i\frac{\sqrt{3}}{2}\right)\right)\left(x - 3\left(\frac{-1}{2} - i\frac{\sqrt{3}}{2}\right)\right) = x^2 + 3x + 9$ and so

$$x^3 - 27 = (x-3)\left(x^2 + 3x + 9\right)$$

where the quadratic polynomial, $x^2 + 3x + 9$ cannot be factored without using complex numbers.

4.1 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 ω 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 $\cos(5x)$. **Hint:** Use Problem 18 on Page 34 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 θ while the polar form of w involves the angle ϕ, 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 $(x+3)\left(x^2+ax+b\right)$ 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 $\overline{zw}=\overline{z}\,\overline{w}$ and then show by induction that $\overline{z_1\cdots z_m}=\overline{z_1}\cdots\overline{z_m}$. Also verify that $\overline{\sum_{k=1}^m z_k}=\sum_{k=1}^m \overline{z_k}$. 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 $p(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$ 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. 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?

16. 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?

17. 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.

18. 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 $\sum_{\gamma=-0}^{m+n} b_\gamma\cos\gamma\theta+\sum_{\tau=-0}^{n+m} c_\tau\sin\tau\theta$? Explain.

Part II

Functions Of One Variable

Functions

5.0.1 Outcomes

1. Understand the definition of a function and be able to tell whether a given relation defines a function. Define and understand the domain of a function and be able to identify the domain when the function comes as a formula.

2. Understand the definition of a continuous function. Know what it means for a function to be continuous at a point. Be able to prove a given function is continuous or is not continuous.

3. Explain why $\lim_{x\to 0} \frac{\sin(x)}{x} = 0$

4. Be able to explain why the circular functions are continuous.

5. Understand the extreme value theorem and the intermediate value theorem for continuous functions.

6. Explain the concept of limit of a function and be able to find the limit of a function from the definition.

7. Describe the limit of a sequence. Understand the relation between the limit of a sequence and continuity.

8. Understand uniform continuity and the proofs of the theorems about continuous functions.

5.1 General Considerations

By this time, you have seen several examples of functions such as the trig. functions. It is a good idea to formalize this concept before proceding further. The concept of a function is that of something which gives a unique output for a given input.

Definition 5.1.1 *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. 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. 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 5.1.2 *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* $x \in D$.

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 5.1.3 *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, $R \equiv \{2,-2,3,6\}$, *the set of second entries, and let* $f(1)=2, f(2)=-2, f(8)=3,$ *and* $f(7)=6$.

Sometimes functions are not given in terms of a formula. For example, consider the following function defined on the positive real numbers having the following definition.

Example 5.1.4 *For* $x \in \mathbb{R}$ *define*

$$f(x) = \begin{cases} \frac{1}{n} \text{ if } x = \frac{m}{n} \text{ in lowest terms for } m,n \in \mathbb{Z} \\ 0 \text{ if } x \text{ is not rational} \end{cases} \tag{5.1}$$

This is a very interesting function called the Dirichlet function. Note that it is not defined in a simple way from a formula.

Example 5.1.5 *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. The next functions are studied in calculus.

Example 5.1.6 *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.*

Example 5.1.7 *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. Examples 5.1.6 and 5.1.7 are considered later in the book and techniques for finding $x(t)$ and $A(t)$ from the given conditions are presented.

In this chapter the functions are defined on some subset of $\mathbb{R}$ having values in $\mathbb{R}$. Later this will be generalized. 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.

Definition 5.1.8 *Let* f, g *be functions with values in* $\mathbb{R}$. *Let* a, b *be elements 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. Functions are well described by a well known scripture, Matthew 7:20, Wherefore by their fruits ye shall know them. When you have specified what it does to something in its domain, you have told what the function is.

Example 5.1.9 *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 5.1.10 *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.

Definition 5.1.11 *For any function, $f : D(f) \subseteq X \to Y$, define the following set known as the inverse image of y.*

$$f^{-1}(y) \equiv \{x \in D(f) : f(x) = y\}.$$

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, $1-1$. Thus f is one to one, $1-1$, if whenever $f(x) = f(x_1)$, then $x = x_1$. 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 $f\left(f^{-1}(y)\right) = y$ for all $y \in f(D(f))$. Defining id *by* $\mathrm{id}(z) \equiv z$ *this says* $f \circ f^{-1} = \mathrm{id}$ *and* $f^{-1} \circ f = \mathrm{id}$.

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

Definition 5.1.12 *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.

Closely related to the definition of a function is the concept of the graph of a function.

Definition 5.1.13 *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.*

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

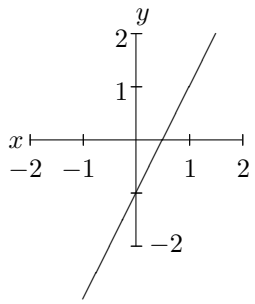
$\mathbb{R}^2$ *denotes the Cartesian product of $\mathbb{R}$ with $\mathbb{R}$.*

The notion of Cartesian product is just an abstraction of the concept of identifying a point in the plane with an ordered pair of numbers.

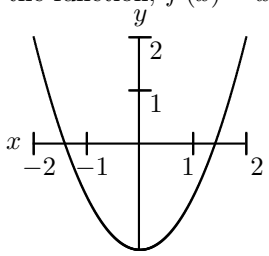
Definition 5.1.14 *Let $f : D(f) \to R(f)$ be a function. The graph of f consists of the set,*

$$\{(x, y) : y = f(x) \text{ for } x \in D(f)\}.$$

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 52. 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, $f(x) = x^2 - 2$



Definition 5.1.15 *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 $\{f_i\}_{i=k}^{\infty}$ or just $\{f_i\}$ for short.*

Example 5.1.16 *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 to us 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 5.1.17 *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 5.1.18 *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 $a_n = \left(n^2 + 1\right)$. Thus $a_1 = 2$, $a_3 = 10$, etc. If

$$n_1 = 1, n_2 = 3, n_3 = 5, \cdots, n_k = 2k - 1,$$

then letting $b_k = a_{n_k}$, it follows

$$b_k = \left((2k-1)^2 + 1\right) = 4k^2 - 4k + 2.$$

5.2 Exercises

1. Let $g(t) \equiv \sqrt{2-t}$ and let $f(t) = \frac{1}{t}$. Find $g \circ f$. Include the domain of $g \circ f$.
2. 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) $f(x) = \sqrt{\frac{x-4}{3x+5}}$
   (e) $f(x) = \sqrt{\frac{x^2-4}{x+1}}$
3. 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}?
4. 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 guess a formula for the k^{th} term of this sequence?
5. 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.
6. 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.
7. 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.
8. 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?
9. 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.

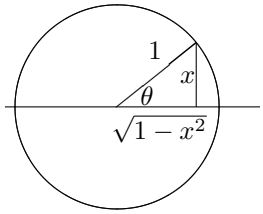
(a) $f+g$
(b) fg
(c) f^3
(d) f/g

10. Draw the graph of the function $f(x) = x^3 + 1$.

11. Draw the graph of the function $f(x) = x^2 + 2x + 2$.

12. Draw the graph of the function $f(x) = \frac{x}{1+x}$.

13. 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 $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, 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, $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Thus $\arcsin\left(\frac{1}{2}\right)$ is the angle whose sin is $\frac{1}{2}$ which is in $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. This angle is $\frac{\pi}{6}$. Suppose you wanted to find $\tan(\arcsin(x))$. How would you do it? Consider the following picture which corresponds to the case where $x > 0$.



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) $\cos(\arcsin(x))$

14. Using Problem 13 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) $\csc\left(\arcsin\left(x^2\right)\right)$
(d) $\cos(2\arcsin(x))$
(e) $\tan(\arcsin(x) + \arcsin(y))$
(f) $\csc(\arcsin(x) - \arcsin(y))$

15. 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) $\sec(\arccos(x))$

16. Using Problem 15 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) $\csc\left(\arccos\left(x^2\right)\right)$

(d) $\cos(\arcsin(x)+\arccos(y))$

(e) $\tan(\arcsin(x)+\arccos(y))$

(f) $\csc(2\arcsin(x)-\arccos(y))$

17. The function, arctan is defined as $\arctan(x) \equiv$ 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) $\sec(\arctan(x))$

18. Using Problem 17 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) $\csc\left(\arccos\left(x^2\right)\right)$

(d) $\cos(2\arctan(x)+\arcsin(y))$

(e) $\tan(\arctan(x)+2\arccos(y))$

(f) $\csc(2\arctan(x)-\arccos(y))$

19. Suppose $a_n = \frac{1}{n}$ and let $n_k = 2^k$. Find b_k where $b_k = a_{n_k}$.

20. If X_i are sets and for some j, $X_j = \emptyset$, the empty set. Verify carefully that $\prod_{i=1}^n X_i = \emptyset$.

21. Suppose $f(x) + f\left(\frac{1}{x}\right) = 7x$ and f is a function defined on $\mathbb{R}\setminus\{0\}$, the nonzero real numbers. Find all values of x where $f(x) = 1$ if there are any. Does there exist any such function?

22. 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?

23. 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, $a_{n+2} \equiv a_n + a_{n+1}$ 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.

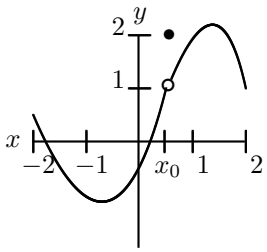
24. In an ordinary annuity, you make constant payments, P at the beginning of each payment period. These accrue interest at the rate of r per payment period. This means at the start of the first payment period, there is the payment $P \equiv A_1$. Then this produces an amount rP in interest so at the beginning of the second payment period, you would have $rP + P + P \equiv A_2$. 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 recurrance 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 $C = \frac{P}{r}, z = (1+r)$, and $s = -\frac{P}{r}$.

25. 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.

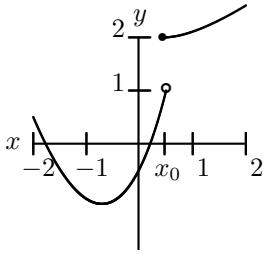
5.3 Continuous Functions

The concept of function is far too general to be useful in calculus. There are various ways to restrict the concept in order to study something interesting and the types of restrictions considered depend very much on what you find interesting. In Calculus, the most fundamental restriction made is to assume the functions are continuous. Continuous functions are those in which a sufficiently small change in x results in a small change in $f(x)$. They rule out things which could never happen physically. For example, it is not possible for a car to jump from one point to another instantly. Making this restriction precise turns out to be surprisingly difficult although many of the most important theorems about continuous functions seem intuitively clear.

Before giving the careful mathematical definitions, here are examples of graphs of functions which are not continuous at the point x_0.



You see, there is a hole in the picture of the graph of this function and instead of filling in the hole with the appropriate value, $f(x_0)$ is too large. This is called a removable discontinuity because the problem can be fixed by redefining the function at the point x_0. Here is another example.



You see from this picture that there is no way to get rid of the jump in the graph of this function by simply redefining the value of the function at x_0. That is why it is called a nonremovable discontinuity or jump discontinuity. Now that pictures have been given of what it is desired to eliminate, it is time to give the precise definition.

The definition which follows, due to Cauchy[1] and Weierstrass[2] is the precise way to

[1] Augustin Louis Cauchy 1789-1857 was the son of a lawyer who was married to an aristocrat. He was born in France just after the fall of the Bastille and his family fled the reign of terror and hid in the countryside till it was over. Cauchy was educated at first by his father who taught him Greek and Latin. Eventually Cauchy learned many languages. He was also a good Catholic.

After the reign of terror, the family returned to Paris and Cauchy studied at the university to be an engineer but became a mathematician although he made fundamental contributions to physics and engineering. Cauchy was one of the most prolific mathematicians who ever lived. He wrote several hundred papers which fill 24 volumes. He also did research on many topics in mechanics and physics including elasticity, optics and astronomy. More than anyone else, Cauchy invented the subject of complex analysis. He is also credited with giving the first rigorous definition of continuity.

He married in 1818 and lived for 12 years with his wife and two daughters in Paris till the revolution of 1830. Cauchy refused to take the oath of allegience to the new ruler and ended up leaving his family and going into exile for 8 years.

Notwithstanding his great achievments he was not known as a popular teacher.

[2] Wilhelm Theodor Weierstrass 1815-1897 brought calculus to essentially the state it is in now. When he was a secondary school teacher, he wrote a paper which was so profound that he was granted a doctor's degree. He made fundamental contributions to partial differential equations, complex analysis, calculus of variations, and many other topics. He also discovered some pathological examples such as space filling curves. Cauchy gave the definition in words and Weierstrass, somewhat later produced the totally rigorous ε δ definition presented here. The need for rigor in the subject of calculus was only realized over a long period of time.

exclude the sort of behavior described above and all statements about continuous functions must ultimately rest on this definition from now on.

Definition 5.3.1 *A function $f : D(f) \subseteq \mathbb{R} \to \mathbb{R}$ is continuous at $x \in D(f)$ if for each $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $y \in D(f)$ and*

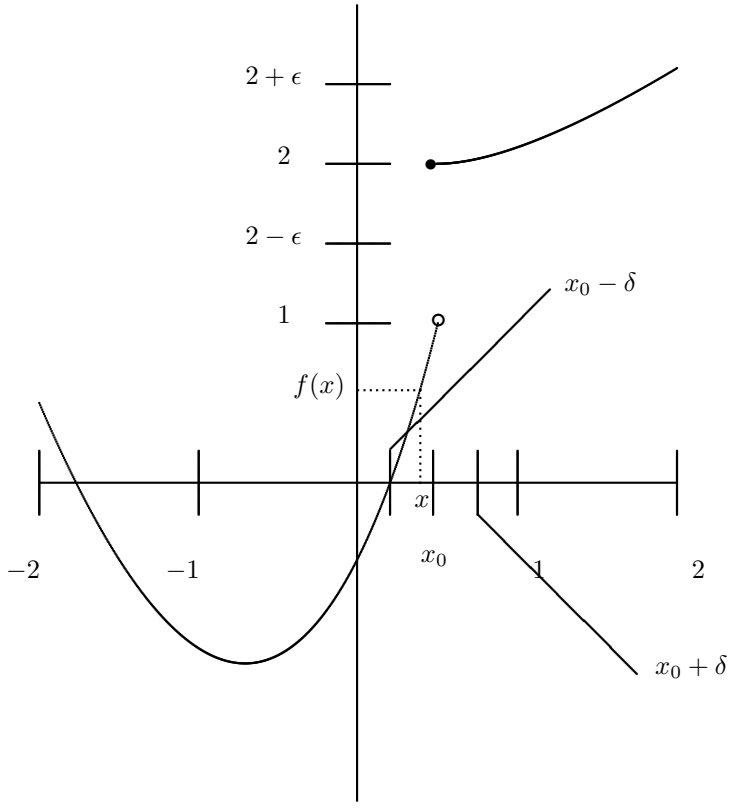
$$|y - x| < \delta$$

it follows that

$$|f(x) - f(y)| < \varepsilon.$$

A function, f is continuous if it is continuous at every point of $D(f)$.

In sloppy English this definition says roughly the following: A function, f is continuous at x when it is possible to make $f(y)$ as close as desired to $f(x)$ provided y is taken close enough to x. In fact this statement in words is pretty much the way Cauchy described it. The completely rigorous definition above is due to Weierstrass. This definition does indeed rule out the sorts of graphs drawn above. Consider the second nonremovable discontinuity. The removable discontinuity case is similar.



For the ε shown you can see from the picture that no matter how small you take δ, there will be points, x, between $x_0 - \delta$ and x_0 where $f(x) < 2 + \varepsilon$. In particular, for these values of x, $|f(x) - f(x_0)| > \varepsilon$. Therefore, the definition of continuity given above excludes the situation in which there is a jump in the function. Similar reasoning shows it excludes the removable discontinuity case as well. There are many ways a function can fail to be

continuous and it is impossible to list them all by drawing pictures. This is why it is so important to use the definition. The other thing to notice is that the concept of continuity as described in the definition is a point property. That is to say it is a property which a function may or may not have at a single point. Here is an example.

Example 5.3.2 *Let*

$$f(x) = \begin{cases} x \text{ if } x \text{ is rational} \\ 0 \text{ if } x \text{ is irrational} \end{cases}.$$

This function is continuous at $x = 0$ and nowhere else.

To verify the assertion about the above function, first show it is not continuous at x if $x \neq 0$. Take such an x and let $\varepsilon = |x|/2$. Now let $\delta > 0$ be completely arbitrary. In the interval, $(x-\delta, x+\delta)$ there are rational numbers, y_1 such that $|y_1| > |x|$ and irrational numbers, y_2. Thus $|f(y_1) - f(y_2)| = |y_1| > |x|$. If f were continuous at x, there would exist $\delta > 0$ such that for every point, $y \in (x-\delta, x+\delta)$, $|f(y) - f(x)| < \varepsilon$. But then, letting y_1 and y_2 be as just described,

$$\begin{aligned} |x| < |y_1| &= |f(y_1) - f(y_2)| \\ &\leq |f(y_1) - f(x)| + |f(x) - f(y_2)| < 2\varepsilon = |x|, \end{aligned}$$

which is a contradiction. Since a contradiction is obtained by assuming that f is continuous at x, it must be concluded that f is not continuous there. To see f is continuous at 0, let $\varepsilon > 0$ be given and let $\delta = \varepsilon$. Then if $|y - 0| < \delta = \varepsilon$, Then

$$\begin{aligned} |f(y) - f(0)| &= 0 \text{ if } y \text{ is irrational} \\ |f(y) - f(0)| &= |y| < \varepsilon \text{ if } y \text{ is rational.} \end{aligned}$$

either way, whenever $|y - 0| < \delta$, it follows $|f(y) - f(0)| < \varepsilon$ and so f is continuous at $x = 0$. How did I know to let $\delta = \varepsilon$? That is a very good question. The choice of δ for a particular ε is usually arrived at by using intuition, the actual ε δ argument reduces to a verification that the intuition was correct. Here is another example.

Example 5.3.3 *Show the function, $f(x) = -5x + 10$ is continuous at $x = -3$.*

To do this, note first that $f(-3) = 25$ and it is desired to verify the conditions for continuity. Consider the following.

$$|-5x + 10 - (25)| = 5|x - (-3)|.$$

This allows one to find a suitable δ. If $\varepsilon > 0$ is given, let $0 < \delta \leq \frac{1}{5}\varepsilon$. Then if $0 < |x - (-3)| < \delta$, it follows from this inequality that

$$|-5x + 10 - (25)| = 5|x - (-3)| < 5\frac{1}{5}\varepsilon = \varepsilon.$$

Sometimes the determination of δ in the verification of continuity can be a little more involved. Here is another example.

Example 5.3.4 *Show the function, $f(x) = \sqrt{2x + 12}$ is continuous at $x = 5$.*

First note $f(5) = \sqrt{22}$. Now consider:

$$\left|\sqrt{2x + 12} - \sqrt{22}\right| = \left|\frac{2x + 12 - 22}{\sqrt{2x + 12} + \sqrt{22}}\right|$$

$$= \frac{2}{\sqrt{2x+12}+\sqrt{22}}\left|x-5\right| \leq \frac{1}{11}\sqrt{22}\left|x-5\right|$$

whenever $|x-5|<1$ because for such x, $\sqrt{2x+12}>0$. Now let $\varepsilon>0$ be given. Choose δ such that $0<\delta\leq \min\left(1,\frac{\varepsilon\sqrt{22}}{2}\right)$. Then if $|x-5|<\delta$, all the inequalities above hold and

$$\left|\sqrt{2x+12}-\sqrt{22}\right| \leq \frac{2}{\sqrt{22}}\left|x-5\right| < \frac{2}{\sqrt{22}}\frac{\varepsilon\sqrt{22}}{2} = \varepsilon.$$

Exercise 5.3.5 *Show $f(x)=-3x^2+7$ is continuous at $x=7$.*

First observe $f(7)=-140$. Now

$$\left|-3x^2+7-(-140)\right| = 3\left|x+7\right|\left|x-7\right| \leq 3\left(|x|+7\right)\left|x-7\right|.$$

If $|x-7|<1$, it follows from the version of the triangle inequality which states $||s|-|t|| \leq |s-t|$ that $|x|<1+7$. Therefore, if $|x-7|<1$, it follows that

$$\left|-3x^2+7-(-140)\right| \leq 3\left((1+7)+7\right)\left|x-7\right|$$

$$= 3\left(1+27\right)\left|x-7\right| = 84\left|x-7\right|.$$

Now let $\varepsilon>0$ be given. Choose δ such that $0<\delta\leq\min\left(1,\frac{\varepsilon}{84}\right)$. Then for $|x-7|<\delta$, it follows

$$\left|-3x^2+7-(-140)\right| \leq 84\left|x-7\right| < 84\left(\frac{\varepsilon}{84}\right) = \varepsilon.$$

These $\varepsilon\ \delta$ proofs will not be emphasized any more than necessary. However, you should try a few of them because until you master this concept, you will not really understand calculus as it has been understood for approximately 150 years. The best you can do without this definition is to gain an understanding of the subject as it was understood by people in the 1700's, before the need for rigor was realized. Don't be discouraged by these historical observations. If you are able to master calculus as understood by Lagrange or Laplace[3], you will have learned some very profound ideas even if they did originate in the eighteenth century.

5.4 Sufficient Conditions For Continuity

The next theorem is a fundamental result which is convenient for avoiding the $\varepsilon\ \delta$ definition of continuity.

Theorem 5.4.1 *The following assertions are valid for f,g functions and a,b numbers.*

1. *The function, $af+bg$ is continuous at x when f, g are continuous at $x\in D(f)\cap D(g)$ and $a,b\in\mathbb{R}$.*
2. *If f and g are each real valued functions continuous at x, then fg is continuous at x. If, in addition to this, $g(x)\neq 0$, then f/g is continuous at x.*
3. *If f is continuous at x, $f(x)\in D(g)\subseteq\mathbb{R}$, and g is continuous at $f(x)$,then $g\circ f$ is continuous at x.*

[3]Lagrange and Laplace were two great physicists of the 1700's. They made fundamental contributions to the calculus of variations and to mechanics and astronomy.

4. The function $f : \mathbb{R} \to \mathbb{R}$, given by $f(x) = |x|$ is continuous.

The proof of this theorem is in the last section of this chapter but its conclusions are not surprising. For example the first claim says that $(af + bg)(y)$ is close to $(af + bg)(x)$ when y is close to x provided the same can be said about f and g. For the third claim, continuity of f indicates that if y is close enough to x then $f(x)$ is close to $f(y)$ and so by continuity of g at $f(x)$, $g(f(y))$ is close to $g(f(x))$. The fourth claim is verified as follows.

$$|x| = |x - y + y| \leq |x - y| + |y|$$

and so

$$|x| - |y| \leq |x - y|.$$

Similarly,

$$|y| - |x| \leq |x - y|.$$

Therefore,

$$||x| - |y|| \leq |x - y|.$$

It follows that if $\varepsilon > 0$ is given, one can take $\delta = \varepsilon$ and obtain that for $|x - y| < \delta = \varepsilon$,

$$||x| - |y|| \leq |x - y| < \delta = \varepsilon$$

which shows continuity of the function, $f(x) = |x|$.

5.5 Continuity Of Circular Functions

The functions $\sin x$ and $\cos x$ are often called the circular functions. This is because for each $x \in \mathbb{R}$, $(\cos x, \sin x)$ is a point on the unit circle.

Theorem 5.5.1 *The functions,* cos *and* sin *are continuous.*

Proof: First it will be shown that cos and sin are continuous at 0. By Corollary 3.5.4 on Page 62 the following inequality is valid for small positive values of θ.

$$1 - \cos\theta + \sin\theta \geq \theta \geq \sin\theta.$$

It follows that for θ small and positive, $|\theta| \geq |\sin\theta| = \sin\theta$. If $\theta < 0$, then $-\theta = |\theta| > 0$ and $-\theta \geq \sin(-\theta)$. But then this means $|\sin\theta| = -\sin\theta = \sin(-\theta) \leq -\theta = |\theta|$ in this case also. Therefore, whenever $|\theta|$ is small enough,

$$|\theta| \geq |\sin\theta|.$$

Now let $\varepsilon > 0$ be given and take $\delta = \varepsilon$. Then if $|\theta| < \delta$, it follows

$$|\sin\theta - 0| = |\sin\theta - \sin 0| = |\sin\theta| \leq |\theta| < \delta = \varepsilon,$$

showing sin is continuous at 0.

Next, note that for $|\theta| < \pi/2$, $\cos\theta \geq 0$ and so for such θ,

$$\sin^2\theta \geq \frac{\sin^2\theta}{1 + \cos\theta} = \frac{1 - \cos^2\theta}{1 + \cos\theta} = 1 - \cos\theta \geq 0. \tag{5.2}$$

From the first part of this argument for sin, given $\varepsilon > 0$ there exists $\delta > 0$ such that if $|\theta| < \delta$, then $|\sin\theta| < \sqrt{\varepsilon}$. It follows from (5.2) that if $|\theta| < \delta$, then $\varepsilon > 1 - \cos\theta \geq 0$. This proves these functions are continuous at 0. Now $y = (y - x) + x$ and so

$$\cos y = \cos(y - x)\cos x - \sin(x - y)\sin x.$$

Therefore,

$$\cos y - \cos x = \cos(y-x)\cos x - \sin(x-y)\sin x - \cos x$$

and so, since $|\cos x|, |\sin x| \leq 1$,

$$\begin{aligned} |\cos y - \cos x| &\leq |\cos x(\cos(y-x)-1)| + |\sin x||\sin(y-x)| \\ &\leq |\cos(y-x)-1| + |\sin(y-x)|. \end{aligned}$$

From the first part of this theorem, if $|y-x|$ is sufficiently small, both of these last two terms are less than $\varepsilon/2$ and this proves cos is continuous at x. The proof that sin is continuous is left for you to verify.

5.6 Exercises

1. Let $f(x) = 2x+7$. Show f is continuous at every point x. **Hint:** You need to let $\varepsilon > 0$ be given. In this case, you should try $\delta \leq \varepsilon/2$. Note that if one δ works in the definition, then so does any smaller δ.

2. Let $f(x) = x^2+1$. Show f is continuous at $x=3$. **Hint:**

$$\begin{aligned} |f(x)-f(3)| &= \left|x^2+1-(9+1)\right| \\ &= |x+3||x-3|. \end{aligned}$$

   Thus if $|x-3|<1$, it follows from the triangle inequality, $|x|<1+3=4$ and so

$$|f(x)-f(3)| < 4|x-3|.$$

   Now try to complete the argument by letting $\delta \leq \min(1, \varepsilon/4)$. The symbol, min means to take the minimum of the two numbers in the parenthesis.

3. Let $f(x) = x^2+1$. Show f is continuous at $x=4$.

4. Let $f(x) = 2x^2+1$. Show f is continuous at $x=1$.

5. Let $f(x) = x^2+2x$. Show f is continuous at $x=2$. Then show it is continuous at every point.

6. Let $f(x) = |2x+3|$. Show f is continuous at every point. **Hint:** Review the two versions of the triangle inequality for absolute values.

7. Let $f(x) = \frac{1}{x^2+1}$. Show f is continuous at every value of x.

8. Show sin is continuous.

9. Let $f(x) = \sqrt{x}$ show f is continuous at every value of x in its domain. **Hint:** You might want to make use of the identity,

$$\sqrt{x} - \sqrt{y} = \frac{x-y}{\sqrt{x}+\sqrt{y}}$$

   at some point in your argument.

10. Using Theorem 5.4.1, show all polynomials are continuous and that a rational function is continuous at every point of its domain. **Hint:** First show the function given as $f(x) = x$ is continuous and then use the Theorem 5.4.1.

11. Let $f(x) = \begin{cases} 1 \text{ if } x \in \mathbb{Q} \\ 0 \text{ if } x \notin \mathbb{Q} \end{cases}$ and consider $g(x) = f(x)\sin x$. Determine where g is continuous and explain your answer.

12. Suppose f is any function whose domain is the integers. Thus $D(f) = \mathbb{Z}$, the set of whole numbers, $\cdots, -3, -2, -1, 0, 1, 2, 3, \cdots$. Then f is continuous. Why? **Hint:** In the definition of continuity, what if you let $\delta = \frac{1}{4}$? Would this δ work for a given $\varepsilon > 0$? This shows that the idea that a continuous function is one for which you can draw the graph without taking the pencil off the paper is a lot of nonsense.

13. Describe the points where $\tan, \cot, \sec$, and $\csc$ are continuous.

14. Give an example of a function, f which is not continuous at some point but $|f|$ is continuous at that point.

15. Find two functions which fail to be continuous but whose product is continuous.

16. Find two functions which fail to be continuous but whose sum is continuous.

17. Find two functions which fail to be continuous but whose quotient is continuous.

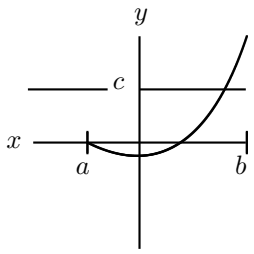
18. Where is the function, $\sin(\tan(x))$ continuous? What is its domain?

19. Where is the function, $\tan(\sin(x))$ continuous? What is its domain?

20. Suppose f is a function defined on $\mathbb{R}$ and f is continuous at 0. Suppose also that $f(x+y) = f(x) + f(y)$. Show that if this is so, then f must be continuous at every value of $x \in \mathbb{R}$. Next show that for every rational number, r, $f(r) = rf(1)$. Finally explain why $f(r) = rf(1)$ for every r a real number. **Hint:** To do this last part, you need to use the density of the rational numbers and continuity of f.

5.7 Properties Of Continuous Functions

Continuous functions have many important properties which are consequences of the completeness axiom. Proofs of these theorems are in the last section at the end of this chapter. The next theorem is called the intermediate value theorem and the following picture illustrates its conclusion. It gives the existence of a certain point.



You see in the picture there is a horizontal line, $y = c$ and a continuous function which starts off less than c at the point a and ends up greater than c at point b. The intermediate value theorem says there is some point between a and b such that the value of the function at this point equals c. You see this taking place in the above picture where the line and the graph of the function cross. The x value at this point is the one whose existence is guaranteed by the theorem. It may seem this is obvious but without completeness the conclusion of the theorem cannot be drawn. Nevertheless, the above picture makes this theorem very easy to believe.

Theorem 5.7.1 *Suppose $f : [a,b] \to \mathbb{R}$ is continuous and suppose $f(a) < c < f(b)$. Then there exists $x \in (a,b)$ such that $f(x) = c$.*

Example 5.7.2 *Does there exist a solution to the equation $\sqrt{x^4+7} - x^3 \sin x = 0$?*

By Theorem 5.4.1 and Problem 9 on Page 103 it follows easily that the function, f, given by $f(x) = \sqrt{x^4+7} - x^3 \sin x$ is continuous. Also, $f(0) = \sqrt{7} > 0$ while

$$f\left(\frac{5\pi}{2}\right) = \sqrt{\left(\frac{5\pi}{2}\right)^4 + 7} - \left(\frac{5\pi}{2}\right)^3 \sin\left(\frac{5\pi}{2}\right)$$

which is approximately equal to $-422.731\,331\,831\,631\,6 < 0$. Therefore, by the intermediate value theorem there must exist $x \in \left(0, \frac{5\pi}{2}\right)$ such that $f(x) = 0$.

This example illustrates the use of this major theorem very well. It says something exists but it does not tell how to find it.

Definition 5.7.3 *A function, f, defined on some interval is strictly increasing if whenever $x < y$, it follows $f(x) < f(y)$. The function is strictly decreasing if whenever $x < y$, it follows $f(x) > f(y)$.*

You should draw a picture of the graph of a strictly increasing or decreasing function from the definition.

Lemma 5.7.4 *Let $\phi : [a,b] \to \mathbb{R}$ be a one to one continuous function. Then ϕ is either strictly increasing or strictly decreasing.*

This lemma is not real easy to prove but it is one of those things which seems obvious. To say a function is one to one is to say that every horizontal line intersects the graph of the function in no more than one point. (This is called the horizontal line test.) Now if your function is continuous (having no jumps) and is one to one, try to imagine how this could happen without it being either strictly increasing or decreasing and you will soon see this is highly believable and in fact, for it to fail would be incredible. The proof of this lemma is in the last section of this chapter in case you are interested.

Corollary 5.7.5 *Let $\phi : (a,b) \to \mathbb{R}$ be a one to one continuous function. Then ϕ is either strictly increasing or strictly decreasing.*

The proof of this corollary is the same as the proof of the lemma. The next corollary follows from the above.

Corollary 5.7.6 *Let $f : (a,b) \to \mathbb{R}$ be one to one and continuous. Then $f(a,b)$ is an open interval, (c,d) and $f^{-1} : (c,d) \to (a,b)$ is continuous. Also, if $f : [a,b] \to \mathbb{R}$ is one to one and continuous, then $f([a,b])$ is a closed interval, $[c,d]$ and $f^{-1} : [c,d] \to [a,b]$ is continuous.*

This corollary is not too surprising either. To view the graph of the inverse function, simply turn things on the side and switch x and y. If the original graph has no jumps in it, neither will the new graph. Of course, the concept of continuity is tied to a rigorous definition, not to the drawing of pictures. This is why there is a proof in the last section of this chapter.

In Russia there is a kind of doll called a matrushka doll. You pick it up and notice it comes apart in the center. Separating the two halves you find an identical doll inside. Then you notice this inside doll also comes apart in the center. Separating the two halves, you

find yet another identical doll inside. This goes on quite a while until the final doll is in one piece. The nested interval lemma is like a matrushka doll except the process never stops. It involves a sequence of intervals, the first containing the second, the second containing the third, the third containing the fourth and so on. The fundamental question is whether there exists a point in all the intervals.

Lemma 5.7.7 *Let $I_k = \left[a^k, b^k\right]$ and suppose that for all $k = 1, 2, \cdots$,*

$$I_k \supseteq I_{k+1}.$$

Then there exists a point, $c \in \mathbb{R}$ which is an element of every I_k.

Proof: Since $I_k \supseteq I_{k+1}$, this implies

$$a^k \leq a^{k+1},\ b^k \geq b^{k+1}. \tag{5.3}$$

Consequently, if $k \leq l$,

$$a^l \leq a^l \leq b^l \leq b^k. \tag{5.4}$$

Now define

$$c \equiv \sup\left\{a^l : l = 1, 2, \cdots\right\}$$

By the first inequality in (5.3), and (5.4)

$$a^k \leq c = \sup\left\{a^l : l = k, k+1, \cdots\right\} \leq b^k \tag{5.5}$$

for each $k = 1, 2 \cdots$. Thus $c \in I_k$ for every k and this proves the lemma. If this went too fast, the reason for the last inequality in (5.5) is that from (5.4), b^k is an upper bound to $\left\{a^l : l = k, k+1, \cdots\right\}$. Therefore, it is at least as large as the least upper bound.

This is really quite a remarkable result and may not seem so obvious. Consider the intervals $I_k \equiv (0, 1/k)$. Then there is no point which lies in all these intervals because no negative number can be in all the intervals and $1/k$ is smaller than a given positive number whenever k is large enough. Thus the only candidate for being in all the intervals is 0 and 0 has been left out of them all. The problem here is that the endpoints of the intervals were not included contrary to the hypotheses of the above lemma in which all the intervals included the endpoints.

With the nested interval lemma, it becomes possible to prove the following lemma which shows a function continuous on a closed interval in $\mathbb{R}$ is bounded.

Lemma 5.7.8 *Let $I = [a, b]$ and let $f : I \to \mathbb{R}$ be continuous. Then f is bounded. That is there exist numbers, m and M such that for all $x \in [a, b]$,*

$$m \leq f(x) \leq M.$$

Proof: Let $I \equiv I_0$ and suppose f is not bounded on I_0. Consider the two sets, $\left[a, \frac{a+b}{2}\right]$ and $\left[\frac{a+b}{2}, b\right]$. Since f is not bounded on I_0, it follows that f must fail to be bounded on at least one of these sets. Let I_1 be one of these on which f is not bounded. Now do to I_1 what was done to I_0 to obtain $I_2 \subseteq I_1$ and for any two points, $x, y \in I_2$

$$|x - y| \leq 2^{-1}\frac{b-a}{2} \leq 2^{-2}(b-a).$$

Continue in this way obtaining sets, I_k such that $I_k \supseteq I_{k+1}$ and for any two points in I_k, x, y, $|x - y| \leq 2^{-k}(b-a)$. By the nested interval lemma,there exists a point, c which is contained in each I_k. Also, by continuity, there exists a $\delta > 0$ such that if $|c - y| < \delta$, then

$$|f(c) - f(y)| < 1. \tag{5.6}$$

Let k be so large that $2^{-k}(b-a) < \delta$. Then for every $y \in I_k$, $|c-y| < \delta$ and so (5.6) holds for all such y. But this implies that for all $y \in I_k$,

$$|f(y)| \leq |f(c)| + 1$$

which shows that f is bounded on I_k contrary to the way I_k was chosen. This contradiction proves the lemma.

Example 5.7.9 *Let $f(x) = 1/x$ for $x \in (0,1)$.*

Clearly, f is not bounded. Does this violate the conclusion of the nested interval lemma? It does not because the end points of the interval involved are not in the interval. The same function defined on $[.000001, 1)$ would have been bounded although in this case the boundedness of the function would not follow from the above lemma because it fails to include the right endpoint.

The next theorem is known as the max min theorem or extreme value theorem.

Theorem 5.7.10 *Let $I = [a,b]$ and let $f : I \to \mathbb{R}$ be continuous. Then f achieves its maximum and its minimum on I. This means there exist, $x_1, x_2 \in I$ such that for all $x \in I$,*

$$f(x_1) \leq f(x) \leq f(x_2).$$

Proof: By completeness of $\mathbb{R}$ and Lemma 5.7.8 $f(I)$ has a least upper bound, M. If for all $x \in I$, $f(x) \neq M$, then by Theorem 5.4.1, the function, $g(x) \equiv (M - f(x))^{-1} = \frac{1}{M-f(x)}$ is continuous on I. Since M is the least upper bound of $f(I)$ there exist points, $x \in I$ such that $(M - f(x))$ is as small as desired. Consequently, g is not bounded above, contrary to Lemma 5.7.8. Therefore, there must exist some $x \in I$ such that $f(x) = M$. This proves f achieves its maximum. The argument for the minimum is similar. Alternatively, you could consider the function $h(x) = M - f(x)$. Then use what was just proved to conclude h achieves its maximum at some point, x_1. Thus $h(x_1) \geq h(x)$ for all $x \in I$ and so $M - f(x_1) \geq M - f(x)$ for all $x \in I$ which implies $f(x_1) \leq f(x)$ for all $x \in I$. This proves the theorem.

5.8 Exercises

1. Give an example of a continuous function defined on $(0,1)$ which does not achieve its maximum on $(0,1)$.

2. Give an example of a continuous function defined on $(0,1)$ which is bounded but which does not achieve either its maximum or its minimum.

3. Give an example of a discontinuous function defined on $[0,1]$ which is bounded but does not achieve either its maximum or its minimum.

4. Give an example of a continuous function defined on $[0,1) \cup (1,2]$ which is positive at 2, negative at 0 but is not equal to zero for any value of x.

5. Let $f(x) = x^5 + ax^4 + bx^3 + cx^2 + dx + e$ where a, b, c, d, and e are numbers. Show there exists x such that $f(x) = 0$.

6. Give an example of a function which is one to one but neither strictly increasing nor strictly decreasing. **Hint:** Look for discontinuous functions satisfying the horizontal line test.

7. Do you believe in $\sqrt[7]{8}$? That is, does there exist a number which multiplied by itself seven times yields 8? Before you jump to any conclusions, the number you get on your calculator is wrong. In fact, your calculator does not even know about $\sqrt[7]{8}$. All it can do is try to approximate it and what it gives you is this approximation. Why does it exist? **Hint:** Use the intermediate value theorem on the function, $f(x) = x^7 - 8$.

8. Let $f(x) = x - \sqrt{2}$ for $x \in \mathbb{Q}$, the rational numbers. Show that even though $f(0) < 0$ and $f(2) > 0$, there is no point in $\mathbb{Q}$ where $f(x) = 0$. Does this contradict the intermediate value theorem? Explain.

9. It has been known since the time of Pythagoras that $\sqrt{2}$ is irrational. If you throw out all the irrational numbers, show that the conclusion of the intermediate value theorem could no longer be obtained. That is, show there exists a function which starts off less than zero and ends up larger than zero and yet there is no number where the function equals zero. **Hint:** Try $f(x) = x^2 - 2$. You supply the details.

10. Let f be a continuous function defined on a closed interval, $I_1 \equiv [a, b]$ such that $f(a) < 0$ and $f(b) > 0$. Consider $\frac{a+b}{2}$. If $f\left(\frac{a+b}{2}\right) \geq 0$, let $I_2 = \left[a, \frac{a+b}{2}\right]$ and if $f\left(\frac{a+b}{2}\right) < 0$, let $I_2 \equiv \left[\frac{a+b}{2}, c\right]$. Thus $I_1 \supset I_2$ and the interval I_2 has exactly the same property that I_1 had in terms of f being negative at the left endpoint and nonnegative at the right endpoint. Continue this way obtaining a sequence of nested closed and bounded intervals, $\{I_k\}$. Show there is exactly one point, x in all these intervals and that $f(x) = 0$. This is called the method of bisection and can be used to find a solution to the equation, $f(x) = 0$.

11. Apply the method of bisection described in Problem 10 to find $\sqrt[7]{8}$. Use a calculator to raise things to the seventh power. It will be much easier than doing it by hand.

12. Use the method of bisection and the nested interval lemma to prove the intermediate value theorem.

13. A circular hula hoop lies partly in the shade and partly in the hot sun. Show there exist two points on the hula hoop which are at opposite sides of the hoop which have the same temperature. **Hint:** Imagine this is a circle and points are located by specifying their angle, θ from a fixed diameter. Then letting $T(\theta)$ be the temperature in the hoop, $T(\theta + 2\pi) = T(\theta)$. You need to have $T(\theta) = T(\theta + \pi)$ for some θ. Assume T is a continuous function of θ.

14. A car starts off on a long trip with a full tank of gas. The driver intends to drive the car till it runs out of gas. Show that at some time the number of miles the car has gone exactly equals the number of gallons of gas in the tank.

15. Suppose f is a continuous function defined on $[0, 1]$ which maps $[0, 1]$ into $[0, 1]$. Show there exists $x \in [0, 1]$ such that $x = f(x)$. **Hint:** Consider $h(x) \equiv x - f(x)$ and the intermediate value theorem.

5.9 Limits Of A Function

A concept closely related to continuity is that of the limit of a function.

Definition 5.9.1 *Let $f : D(f) \subseteq \mathbb{R} \to \mathbb{R}$ be a function where $D(f) \supseteq (x - r, x) \cup (x, x + r)$ for some $r > 0$. Note that f is not necessarily defined at x. Then*

$$\lim_{y \to x} f(y) = L$$

if and only if the following condition holds. For all $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$0 < |y - x| < \delta,$$

then,

$$|L - f(y)| < \varepsilon.$$

If everything is the same as the above, except y is required to be larger than x and f is only required to be defined on $(x, x + r)$, then the notation is

$$\lim_{y \to x+} f(y) = L.$$

If f is only required to be defined on $(x - r, x)$ and y is required to be less than x,with the same conditions above, we write

$$\lim_{y \to x-} f(y) = L.$$

Limits are also taken as a variable "approaches" infinity. Of course nothing is "close" to infinity and so this requires a slightly different definition.

$$\lim_{x \to \infty} f(x) = L$$

if for every $\varepsilon > 0$ there exists l such that whenever $x > l$,

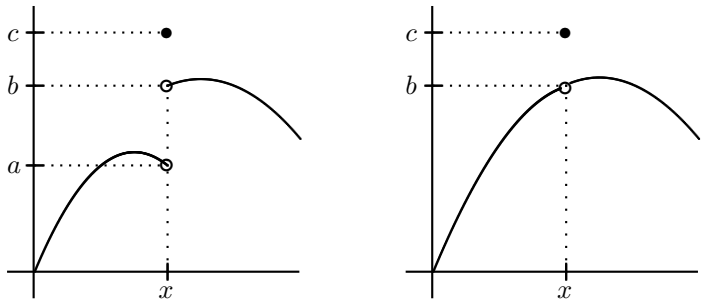
$$|f(x) - L| < \varepsilon \tag{5.7}$$

and

$$\lim_{x \to -\infty} f(x) = L$$

if for every $\varepsilon > 0$ there exists l such that whenever $x < l$, (5.7) holds.

The following pictures illustrate some of these definitions.



In the left picture is shown the graph of a function. Note the value of the function at x equals c while $\lim_{y \to x+} f(y) = b$ and $\lim_{y \to x-} f(y) = a$. In the second picture, $\lim_{y \to x} f(y) = b$. Note that the value of the function at the point x has nothing to do with the limit of the function in any of these cases. **The value of a function at x is irrelevant to the value of the limit at x!** This must always be kept in mind. You do not evaluate interesting limits by computing $f(x)$! In the above picture, $f(x)$ is always wrong! It may be the case that $f(x)$ is right but this is merely a happy coincidence when it occurs and as explained below in Theorem 5.9.6, this is equivalent to f being continuous at x.

Theorem 5.9.2 *If $\lim_{y \to x} f(y) = L$ and $\lim_{y \to x} f(y) = L_1$, then $L = L_1$.*

Proof: Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that if $0 < |y - x| < \delta$, then

$$|f(y) - L| < \varepsilon, \ |f(y) - L_1| < \varepsilon.$$

Therefore, for such y,

$$|L - L_1| \leq |L - f(y)| + |f(y) - L_1| < \varepsilon + \varepsilon = 2\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows $L = L_1$.

The above theorem holds for any of the kinds of limits presented in the above definition.

Another concept is that of a function having either ∞ or $-\infty$ as a limit. In this case, the values of the function do not ever get close to their target because nothing can be close to $\pm\infty$. Roughly speaking, the limit of the function equals ∞ if the values of the function are ultimately larger than any given number. More precisely:

Definition 5.9.3 *If $f(x) \in \mathbb{R}$, then $\lim_{y\to x} f(x) = \infty$ if for every number l, there exists $\delta > 0$ such that whenever $|y - x| < \delta$, then $f(x) > l$. $\lim_{x\to\infty} f(x) = \infty$ if for all k, there exists l such that $f(x) > k$ whenever $x > l$. One sided limits and limits as the variable approaches $-\infty$, are defined similarly.*

It may seem there is a lot to memorize here. In fact, this is not so because all the definitions are intuitive when you understand them.

Theorem 5.9.4 *In this theorem, the symbol, $\lim_{y\to x}$ denotes any of the limits described above. Suppose $\lim_{y\to x} f(y) = L$ and $\lim_{y\to x} g(y) = K$ where K and L are real numbers in $\mathbb{R}$. Then if $a, b \in \mathbb{R}$,*

$$\lim_{y\to x} (af(y) + bg(y)) = aL + bK, \tag{5.8}$$

$$\lim_{y\to x} fg(y) = LK \tag{5.9}$$

and if $K \neq 0$,

$$\lim_{y\to x} \frac{f(y)}{g(y)} = \frac{L}{K}. \tag{5.10}$$

Also, if h is a continuous function defined near L, then

$$\lim_{y\to x} h \circ f(y) = h(L). \tag{5.11}$$

Suppose $\lim_{y\to x} f(y) = L$. If $f(y) \leq a$ all y of interest, then $L \leq a$ and if $f(y) \geq a$ then $L \geq a$.

Proof: The proof of (5.8) is left for you. It is like a corresponding theorem for continuous functions. Next consider (5.9). Let $\varepsilon > 0$ be given. Then by the triangle inequality,

$$\begin{aligned} |fg(y) - LK| &\leq |fg(y) - f(y)K| + |f(y)K - LK| \\ &\leq |f(y)|\,|g(y) - K| + |K|\,|f(y) - L|. \end{aligned} \tag{5.12}$$

There exists δ_1 such that if $0 < |y - x| < \delta_1$, then

$$|f(y) - L| < 1,$$

and so for such y, and the triangle inequality, $|f(y)| < 1 + |L|$. Therefore, for $0 < |y - x| < \delta_1$,

$$|fg(y) - LK| \leq (1 + |K| + |L|)\,[|g(y) - K| + |f(y) - L|]. \tag{5.13}$$

Now let $0 < \delta_2$ be such that for $0 < |x - y| < \delta_2$,

$$|f(y) - L| < \frac{\varepsilon}{2(1 + |K| + |L|)},\ |g(y) - K| < \frac{\varepsilon}{2(1 + |K| + |L|)}.$$

Then letting $0 < \delta \leq \min(\delta_1, \delta_2)$, it follows from (5.13) that

$$|fg(y) - LK| < \varepsilon$$

and this proves (5.9). Limits as $x \to \pm\infty$ and one sided limits are handled similarly.

The proof of (5.10) is left to you. It is just like the theorem about the quotient of continuous functions being continuous provided the function in the denominator is non zero at the point of interest.

Consider (5.11). Since h is continuous near L, it follows that for $\varepsilon > 0$ given, there exists $\eta > 0$ such that if $|y - L| < \eta$, then

$$|h(y) - h(L)| < \varepsilon$$

Now since $\lim_{y \to x} f(y) = L$, there exists $\delta > 0$ such that if $0 < |y - x| < \delta$, then

$$|f(y) - L| < \eta.$$

Therefore, if $0 < |y - x| < \delta$,

$$|h(f(y)) - h(L)| < \varepsilon.$$

The same theorem holds for one sided limits and limits as the variable moves toward $\pm\infty$. The proofs are left to you. They are minor modifications of the above.

It only remains to verify the last assertion. Assume $f(y) \leq a$. It is required to show that $L \leq a$. If this is not true, then $L > a$. Letting ε be small enough that $a < L - \varepsilon$, it follows that ultimately, for y close enough to x, $f(y) \in (L - \varepsilon, L + \varepsilon)$ which requires $f(y) > a$ contrary to assumption.

A very useful theorem for finding limits is called the squeezing theorem.

Theorem 5.9.5 *Suppose* $\lim_{x \to a} f(x) = L = \lim_{x \to a} g(x)$ *and for all* x *near* a,

$$f(x) \leq h(x) \leq g(x).$$

Then

$$\lim_{x \to a} h(x) = L.$$

Proof: If $L \geq h(x)$, then

$$|h(x) - L| \leq |f(x) - L|.$$

If $L < h(x)$, then

$$|h(x) - L| \leq |g(x) - L|.$$

Therefore,

$$|h(x) - L| \leq |f(x) - L| + |g(x) - L|.$$

Now let $\varepsilon > 0$ be given. There exists δ_1 such that if $0 < |x - a| < \delta_1$,

$$|f(x) - L| < \varepsilon/2$$

and there exists δ_2 such that if $0 < |x - a| < \delta_2$, then

$$|g(x) - L| < \varepsilon/2.$$

Letting $0 < \delta \leq \min(\delta_1, \delta_2)$, if $0 < |x - a| < \delta$, then

$$\begin{aligned} |h(x) - L| &\leq |f(x) - L| + |g(x) - L| \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

This proves the theorem.

Theorem 5.9.6 *For $f : I \to \mathbb{R}$, and I is an interval of the form $(a,b), [a,b), (a,b]$, or $[a,b]$, then f is continuous at $x \in I$ if and only if $\lim_{y \to x} f(y) = f(x)$.*

Proof: You fill in the details. Compare the definition of continuous and the definition of the limit just given.

Example 5.9.7 *Find $\lim_{x \to 3} \frac{x^2-9}{x-3}$.*

Note that $\frac{x^2-9}{x-3} = x + 3$ whenever $x \neq 3$. Therefore, if $0 < |x - 3| < \varepsilon$,

$$\left| \frac{x^2 - 9}{x - 3} - 6 \right| = |x + 3 - 6| = |x - 3| < \varepsilon.$$

It follows from the definition that this limit equals 6.

You should be careful to note that in the definition of limit, the variable **never equals the thing it is getting close to**. In this example, x is never equal to 3. This is very significant because, in interesting limits, the function whose limit is being taken will usually not be defined at the point of interest.

Example 5.9.8 *Let*

$$f(x) = \frac{x^2 - 9}{x - 3} \text{ if } x \neq 3.$$

How should f be defined at $x = 3$ so that the resulting function will be continuous there?

In the previous example, the limit of this function equals 6. Therefore, by Theorem 5.9.6 it is necessary to define $f(3) \equiv 6$.

Example 5.9.9 *Find $\lim_{x \to \infty} \frac{x}{1+x}$.*

Write $\frac{x}{1+x} = \frac{1}{1+(1/x)}$. Now it seems clear that $\lim_{x \to \infty} 1 + (1/x) = 1 \neq 0$. Therefore, Theorem 5.9.4 implies

$$\lim_{x \to \infty} \frac{x}{1 + x} = \lim_{x \to \infty} \frac{1}{1 + (1/x)} = \frac{1}{1} = 1.$$

Example 5.9.10 *Show $\lim_{x \to a} \sqrt{x} = \sqrt{a}$ whenever $a \geq 0$. In the case that $a = 0$, take the limit from the right.*

There are two cases. First consider the case when $a > 0$. Let $\varepsilon > 0$ be given. Multiply and divide by $\sqrt{x} + \sqrt{a}$. This yields

$$\left| \sqrt{x} - \sqrt{a} \right| = \left| \frac{x - a}{\sqrt{x} + \sqrt{a}} \right|.$$

Now let $0 < \delta_1 < a/2$. Then if $|x-a| < \delta_1, x > a/2$ and so

$$\left|\sqrt{x}-\sqrt{a}\right| = \left|\frac{x-a}{\sqrt{x}+\sqrt{a}}\right| \leq \frac{|x-a|}{\left(\sqrt{a}/\sqrt{2}\right)+\sqrt{a}}$$
$$\leq \frac{2\sqrt{2}}{\sqrt{a}}|x-a|.$$

Now let $0 < \delta \leq \min\left(\delta_1, \frac{\varepsilon\sqrt{a}}{2\sqrt{2}}\right)$. Then for $0 < |x-a| < \delta$,

$$\left|\sqrt{x}-\sqrt{a}\right| \leq \frac{2\sqrt{2}}{\sqrt{a}}|x-a| < \frac{2\sqrt{2}}{\sqrt{a}}\frac{\varepsilon\sqrt{a}}{2\sqrt{2}} = \varepsilon.$$

Next consider the case where $a = 0$. In this case, let $\varepsilon > 0$ and let $\delta = \varepsilon^2$. Then if $0 < x - 0 < \delta = \varepsilon^2$, it follows that $0 \leq \sqrt{x} < \left(\varepsilon^2\right)^{1/2} = \varepsilon$.

5.10 Exercises

1. Find the following limits if possible
   (a) $\lim_{x\to 0+} \frac{|x|}{x}$
   (b) $\lim_{x\to 0+} \frac{x}{|x|}$
   (c) $\lim_{x\to 0-} \frac{|x|}{x}$
   (d) $\lim_{x\to 4} \frac{x^2-16}{x+4}$
   (e) $\lim_{x\to 3} \frac{x^2-9}{x+3}$
   (f) $\lim_{x\to -2} \frac{x^2-4}{x-2}$
   (g) $\lim_{x\to\infty} \frac{x}{1+x^2}$
   (h) $\lim_{x\to\infty} -2\frac{x}{1+x^2}$
2. Find $\lim_{h\to 0} \frac{\frac{1}{(x+h)^3}-\frac{1}{x^3}}{h}$.
3. Find $\lim_{x\to 4} \frac{\sqrt[4]{x}-\sqrt{2}}{\sqrt{x}-2}$.
4. Find $\lim_{x\to\infty} \frac{\sqrt[5]{3x}+\sqrt[4]{x}+7\sqrt{x}}{\sqrt{3x+1}}$.
5. Find $\lim_{x\to\infty} \frac{(x-3)^{20}(2x+1)^{30}}{(2x^2+7)^{25}}$.
6. Find $\lim_{x\to 2} \frac{x^2-4}{x^3+3x^2-9x-2}$.
7. Find $\lim_{x\to\infty}\left(\sqrt{1-7x+x^2}-\sqrt{1+7x+x^2}\right)$.
8. Prove Theorem 5.9.2 for right, left and limits as $y \to \infty$.
9. Prove from the definition that $\lim_{x\to a}\sqrt[3]{x} = \sqrt[3]{a}$ for all $a \in \mathbb{R}$. **Hint:** You might want to use the formula for the difference of two cubes,

$$a^3 - b^3 = (a-b)\left(a^2+ab+b^2\right).$$

10. Find $\lim_{h\to 0}\frac{(x+h)^2-x^2}{h}$.

11. Prove Theorem 5.9.6 from the definitions of limit and continuity.

12. Find $\lim_{h\to 0}\frac{(x+h)^3-x^3}{h}$

13. Find $\lim_{h\to 0}\frac{\frac{1}{x+h}-\frac{1}{x}}{h}$

14. Find $\lim_{x\to -3}\frac{x^3+27}{x+3}$

15. Find $\lim_{h\to 0}\frac{\sqrt{(3+h)^2}-3}{h}$ if it exists.

16. Find the values of x for which $\lim_{h\to 0}\frac{\sqrt{(x+h)^2}-x}{h}$ exists and find the limit.

17. Find $\lim_{h\to 0}\frac{\sqrt[3]{(x+h)}-\sqrt[3]{x}}{h}$ if it exists. Here $x\neq 0$.

18. Suppose $\lim_{y\to x+} f(y) = L_1 \neq L_2 = \lim_{y\to x-} f(y)$. Show $\lim_{y\to x} f(x)$ does not exist. **Hint:** Roughly, the argument goes as follows: For $|y_1 - x|$ small and $y_1 > x$, $|f(y_1) - L_1|$ is small. Also, for $|y_2 - x|$ small and $y_2 < x$, $|f(y_2) - L_2|$ is small. However, if a limit existed, then $f(y_2)$ and $f(y_1)$ would both need to be close to some number and so both L_1 and L_2 would need to be close to some number. However, this is impossible because they are different.

19. Show $\lim_{x\to 0}\frac{\sin x}{x} = 1$. **Hint:** You might consider Theorem 5.5.1 on Page 102 to write the inequality $|\sin x| + 1 - \cos x \geq |x| \geq |\sin x|$ whenever $|x|$ is small. Then divide both sides by $|\sin x|$ and use some trig. identities to write $\frac{\sin^2 x}{|\sin x|(1+\cos x)} + 1 \geq \frac{|x|}{|\sin x|} \geq 1$ and then use squeezing theorem.

20. Let $f(x,y) = \frac{x^2-y^2}{x^2+y^2}$. Find $\lim_{x\to 0}(\lim_{y\to 0} f(x,y))$ and $\lim_{y\to 0}(\lim_{x\to 0} f(x,y))$. If you did it right you got -1 for one answer and 1 for the other. What does this tell you about interchanging limits?

5.11 The Limit Of A Sequence

A closely related concept is the limit of a sequence. This was defined precisely a little before the definition of the limit by Bolzano[4]. The following is the precise definition of what is meant by the limit of a sequence.

Definition 5.11.1 *A sequence $\{a_n\}_{n=1}^{\infty}$ converges to a,*

$$\lim_{n\to\infty} a_n = a \text{ or } a_n \to a$$

if and only if for every $\varepsilon > 0$ there exists n_ε such that whenever $n \geq n_\varepsilon$,

$$|a_n - a| < \varepsilon.$$

[4]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 concience. His convictions got him in trouble with Emporer 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.

In words the definition says that given any measure of closeness, ε, the terms of the sequence are eventually all this close to a. Note the similarity with the concept of limit. Here, the word "eventually" refers to n being sufficiently large. Earlier, it referred to y being sufficiently close to x on one side or another or else x being sufficiently large in either the positive or negative directions. The above definition is always the definition of what is meant by the limit of a sequence. If the a_n are complex numbers or later on, vectors the definition remains the same. If $a_n = x_n + iy_n$ and $a = x + iy, |a_n - a| = \sqrt{(x_n - x)^2 + (y_n - y)^2}$. Recall the way you measure distance between two complex numbers.

Theorem 5.11.2 *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_ε such that if $n \geq n_\varepsilon$, then $|a_n - a| < \varepsilon$ and $|a_n - a_1| < \varepsilon$. Therefore, for such n,

$$\begin{aligned} |a_1 - a| &\leq |a_1 - a_n| + |a_n - a| \\ &< \varepsilon + \varepsilon < |a_1 - a|/2 + |a_1 - a|/2 = |a_1 - a|, \end{aligned}$$

a contradiction.

Example 5.11.3 *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 5.11.4 *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 5.11.5 *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.

Theorem 5.11.6 *Suppose* $\{a_n\}$ *and* $\{b_n\}$ *are sequences and that*

$$\lim_{n\to\infty} a_n = a \text{ and } \lim_{n\to\infty} b_n = b.$$

Also suppose x *and* y *are real numbers. Then*

$$\lim_{n\to\infty} xa_n + yb_n = xa + yb \tag{5.14}$$

$$\lim_{n\to\infty} a_n b_n = ab \tag{5.15}$$

If $b \neq 0$,

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{a}{b}. \tag{5.16}$$

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

$$\begin{aligned} |a_n b_n - ab| &\leq |a_n b_n - a_n b| + |a_n b - ab| \\ &\leq |a_n| \, |b_n - b| + |b| \, |a_n - a| \\ &\leq (|a| + 1) \, |b_n - b| + |b| \, |a_n - a| \, . \end{aligned}$$

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 $n \geq n_\varepsilon$,

$$\begin{aligned} |a_n b_n - ab| &\leq (|a| + 1) \, |b_n - b| + |b| \, |a_n - a| \\ &< (|a| + 1) \frac{\varepsilon}{2(|a| + 1)} + |b| \frac{\varepsilon}{2(|b| + 1)} \leq \varepsilon. \end{aligned}$$

This proves (5.15). Next consider (5.16).

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,

$$\left| \frac{a_n}{b_n} - \frac{a}{b} \right| = \left| \frac{a_n b - ab_n}{bb_n} \right| \leq \frac{2}{|b|^2} \left[|a_n b - ab| + |ab - ab_n| \right]$$

$$\leq \frac{2}{|b|} |a_n - a| + \frac{2|a|}{|b|^2} |b_n - b| \, .$$

Now choose n_2 so large that if $n \geq n_2$, then

$$|a_n - a| < \frac{\varepsilon |b|}{4}, \text{ and } |b_n - b| < \frac{\varepsilon |b|^2}{4(|a| + 1)}.$$

Letting $n_\varepsilon > \max(n_1, n_2)$, it follows that for $n \geq n_\varepsilon$,

$$\begin{aligned}\left|\frac{a_n}{b_n} - \frac{a}{b}\right| &\leq \frac{2}{|b|}|a_n - a| + \frac{2|a|}{|b|^2}|b_n - b| \\ &< \frac{2}{|b|}\frac{\varepsilon|b|}{4} + \frac{2|a|}{|b|^2}\frac{\varepsilon|b|^2}{4(|a|+1)} < \varepsilon.\end{aligned}$$

Another very useful theorem for finding limits is the squeezing theorem.

Theorem 5.11.7 *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 $n \geq n_1$,

$$|a_n - a| < \varepsilon/2 \text{ and } |b_n - b| < \varepsilon/2.$$

Then for such n,

$$|c_n - a| \leq |a_n - a| + |b_n - a| < \varepsilon.$$

This proves the theorem.

As an example, consider the following.

Example 5.11.8 *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.*

Theorem 5.11.9 $\lim_{n\to\infty} r^n = 0$. *Whenever* $|r| < 1$.

Proof: If $0 < r < 1$ if follows $r^{-1} > 1$. Why? Letting $\alpha = \frac{1}{r} - 1$, it follows

$$r = \frac{1}{1+\alpha}.$$

Therefore, by the binomial theorem,

$$0 < r^n = \frac{1}{(1+\alpha)^n} \leq \frac{1}{1+\alpha n}.$$

Therefore, $\lim_{n\to\infty} r^n = 0$ 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 5.1.18 on Page 94 at this point.

Theorem 5.11.10 *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_ε 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.

5.11.1 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 5.11.11 *$\{a_n\}$ is a Cauchy sequence if for all $\varepsilon > 0$, there exists n_ε 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 5.11.12 *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 5.11.13 *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_ε 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 5.11.14 *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 exists different descriptions of the completeness axiom. If you like you can simply add the three new criteria in the following theorem to the list of things which you mean when you say $\mathbb{R}$ is complete 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 2.14.1 on Page 44 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.

It is the concept of completeness and the notion of limits which sets analysis apart from algebra. You will find that every existence theorem, a theorem which asserts the existence of something, in analysis depends on the assumption that some space is complete.

Theorem 5.11.15 *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 $s_1, \cdots, s_n$ and $b_1, \cdots, b_n$ 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

$$b_{n+1} \equiv \frac{s_n + b_n}{2}$$

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 $n > m$,

$$0 \leq b_m - b_n = |b_m - b_n|$$

$$= \sum_{k=m}^{n-1} b_k - b_{k+1} \leq \sum_{k=m}^{n-1} b_k - s_k \leq \sum_{k=m}^{n-1} 2^{-k} (b_1 - s_1)$$

$$= \frac{2^{-m} - 2^{-n}}{2^{-1}} (b_1 - s_1) \leq 2^{-m+1} (b_1 - s_1)$$

and $\lim_{m\to\infty} 2^{-m} = 0$ by Theorem 5.11.9. 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

$$\begin{aligned} |l - l_1| &\leq |l - s_n| + |s_n - b_n| + |b_n - l_1| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon. \end{aligned}$$

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_ε 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.

Theorem 5.11.16 *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 2.14.3 on Page 45 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 $a = \lim_{n\to\infty} a_n$.

5.11.2 Decimals

You are all 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.[5] 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 5.11.17 *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 5.11.18 *The above definition makes sense.*

[5] 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.

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 5.11.16 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

$$\begin{aligned}\frac{9}{10}\left(\sum_{k=1}^{n}\frac{1}{10^k}\right) &= \sum_{k=1}^{n}\frac{1}{10^k}-\frac{1}{10}\sum_{k=1}^{n}\frac{1}{10^k}=\sum_{k=1}^{n}\frac{1}{10^k}-\sum_{k=2}^{n+1}\frac{1}{10^k}\\ &= \frac{1}{10}-\frac{1}{10^{n+1}}\end{aligned}$$

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.

5.11.3 Continuity And The Limit Of A Sequence

There is a very useful way of thinking of continuity in terms of limits of sequences found in the following theorem. In words, it says a function is continuous if it takes convergent sequences to convergent sequences whenever possible.

Theorem 5.11.19 *A function $f:D(f)\to\mathbb{R}$ is continuous at $x\in D(f)$ if and only if, whenever $x_n\to x$ with $x_n\in D(f)$, it follows $f(x_n)\to f(x)$.*

Proof: Suppose first that f is continuous at x and let $x_n\to x$. Let $\varepsilon>0$ be given. By continuity, there exists $\delta>0$ such that if $|y-x|<\delta$, then $|f(x)-f(y)|<\varepsilon$. However, there exists n_δ such that if $n\geq n_\delta$, then $|x_n-x|<\delta$ and so for all n this large,

$$|f(x)-f(x_n)|<\varepsilon$$

which shows $f(x_n)\to f(x)$.

Now suppose the condition about taking convergent sequences to convergent sequences holds at x. Suppose f fails to be continuous at x. Then there exists $\varepsilon>0$ and $x_n\in D(f)$ such that $|x-x_n|<\frac{1}{n}$, yet

$$|f(x)-f(x_n)|\geq\varepsilon.$$

But this is clearly a contradiction because, although $x_n\to x$, $f(x_n)$ fails to converge to $f(x)$. It follows f must be continuous after all. This proves the theorem.

5.12 Exercises

1. Find $\lim_{n\to\infty}\frac{n}{3n+4}$.
2. Find $\lim_{n\to\infty}\frac{3n^4+7n+1000}{n^4+1}$.
3. Find $\lim_{n\to\infty}\frac{2^n+7(5^n)}{4^n+2(5^n)}$.

4. Find $\lim_{n\to\infty} n\tan\frac{1}{n}$. **Hint:** See Problem 19 on Page 114.

5. Find $\lim_{n\to\infty} n\sin\frac{2}{n}$. **Hint:** See Problem 19 on Page 114.

6. Find $\lim_{n\to\infty}\sqrt{\left(n\sin\frac{9}{n}\right)}$. **Hint:** See Problem 19 on Page 114.

7. Find $\lim_{n\to\infty}\sqrt{(n^2+6n)}-n$. **Hint:** Multiply and divide by $\sqrt{(n^2+6n)}+n$.

8. Find $\lim_{n\to\infty}\sum_{k=1}^{n}\frac{1}{10^k}$.

9. Suppose $\{x_n+iy_n\}$ is a sequence of complex numbers which converges to the complex number $x+iy$. Show this happens if and only if $x_n\to x$ and $y_n\to y$.

10. For $|r|<1$, find $\lim_{n\to\infty}\sum_{k=0}^{n}r^k$. **Hint:** First show $\sum_{k=0}^{n}r^k=\frac{r^{n+1}}{r-1}-\frac{1}{r-1}$. Then recall Theorem 5.11.9.

11. Suppose $x=.3434343\overline{34}$ 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 $100x=34.3434343\overline{34}$ and then subtract to get $99x=34$. From this you conclude that $x=34/99$. Fully justify this procedure. **Hint:** $.3434343\overline{34}=\lim_{n\to\infty}34\sum_{k=1}^{n}\left(\frac{1}{100}\right)^k$ now use Problem 10.

12. Suppose $D(f)=[0,1]\cup\{9\}$ and $f(x)=x$ on $[0,1]$ while $f(9)=5$. Is f continuous at the point, 9? Use whichever definition of continuity you like.

13. Suppose $x_n\to x$ and $x_n\le c$. Show that $x\le c$. Also show that if $x_n\to x$ and $x_n\ge c$, then $x\ge c$. **Hint:** If this is not true, argue that for all n large enough $x_n>c$.

14. Let $a\in[0,1]$. Show $a=.a_1a_2a_3\cdots$ for a unique choice of integers, $a_1,a_2,\cdots$ if it is possible to do this. Otherwise, give an example.

15. Show every rational number between 0 and 1 has a decimal expansion which either repeats or terminates.

16. Consider the number whose decimal expansion is $.0100100010000100000100000001\cdots$. Show this is an irrational number. Now using this, show that between any two integers there exists an irrational number. Next show that between any two numbers there exists an irrational number.

17. Using the binomial theorem prove that for all $n\in\mathbb{N}$, $\left(1+\frac{1}{n}\right)^n\le\left(1+\frac{1}{n+1}\right)^{n+1}$. **Hint:** Show first that $\binom{n}{k}=\frac{n\cdot(n-1)\cdots(n-k+1)}{k!}$. By the binomial theorem,

$$\left(1+\frac{1}{n}\right)^n=\sum_{k=0}^{n}\binom{n}{k}\left(\frac{1}{n}\right)^k=\sum_{k=0}^{n}\frac{\overbrace{n\cdot(n-1)\cdots(n-k+1)}^{k\text{ factors}}}{k!n^k}.$$

Now consider the term $\frac{n\cdot(n-1)\cdots(n-k+1)}{k!n^k}$ and note that a similar term occurs in the binomial expansion for $\left(1+\frac{1}{n+1}\right)^{n+1}$ except you replace n with $n+1$ whereever 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.

18. Prove by induction that for all $k\ge4$, $2^k\le k!$

19. Use the Problems 21 and 17 to verify for all $n \in \mathbb{N}$, $\left(1+\frac{1}{n}\right)^n \leq 3$.

20. Prove $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$ exists and equals a number less than 3.

21. Using Problem 19, prove $n^{n+1} \geq (n+1)^n$ for all integers, $n \geq 3$.

22. Find $\lim_{n\to\infty} n\sin n$ if it exists. If it does not exist, explain why it does not.

23. 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. **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 $[a, a+\varepsilon]$.

24. Let $A_n = \sum_{k=2}^{n} \frac{1}{k(k-1)}$ for $n \geq 2$. Show $\lim_{n\to\infty} A_n$ exists. **Hint:** Show there exists an upper bound to the A_n as follows.

$$\begin{aligned}\sum_{k=2}^{n} \frac{1}{k\,(k-1)} &= \sum_{k=2}^{n}\left(\frac{1}{k-1}-\frac{1}{k}\right)\\ &= \frac{1}{2}-\frac{1}{n-1} \leq \frac{1}{2}.\end{aligned}$$

25. Let $H_n = \sum_{k=1}^{n} \frac{1}{k^2}$ for $n \geq 2$. Show $\lim_{n\to\infty} H_n$ exists. **Hint:** Use the above problem to obtain the existence of an upper bound.

26. Let a be a positive number and let $x_1 = b > 0$ where $b^2 > a$. Explain why there exists such a number, b. Now having defined x_n, define $x_{n+1} \equiv \frac{1}{2}\left(x_n + \frac{a}{x_n}\right)$. Verify that $\{x_n\}$ is a decreasing sequence and that it satisfies $x_n^2 \geq a$ for all n and is therefore, bounded below. Explain why $\lim_{n\to\infty} x_n$ exists. If x is this limit, show that $x^2 = a$. Explain how this shows that every positive real number has a square root. This is an example of a recursively defined sequence. Note this does not give a formula for x_n, just a rule which tells us how to define x_{n+1} if x_n is known.

27. Let $a_1 = 0$ and suppose that $a_{n+1} = \frac{9}{9-a_n}$. Write a_2, a_3, a_4. Now prove that for all n, it follows that $a_n \leq \frac{9}{2} + \frac{3}{2}\sqrt{5}$ (By Problem 7 on Page 108 there is no problem with the existence of various roots of positive numbers.) and so the sequence is bounded above. Next show that the sequence is increasing and so it converges. Find the limit of the sequence. **Hint:** You should prove these things by induction. Finally, to find the limit, let $n \to \infty$ in both sides and argue that the limit, a, must satisfy $a = \frac{9}{9-a}$.

28. If $x \in \mathbb{R}$, show there exists a sequence of rational numbers, $\{x_n\}$ such that $x_n \to x$ and a sequence of irrational numbers, $\{x_n'\}$ such that $x_n' \to x$. Now consider the following function.

$$f(x) = \begin{cases} 1 \text{ if } x \text{ is rational} \\ 0 \text{ if } x \text{ is irrational} \end{cases}.$$

Show using the sequential version of continuity in Theorem 5.11.19 that f is discontinuous at every point.

29. If $x \in \mathbb{R}$, show there exists a sequence of rational numbers, $\{x_n\}$ such that $x_n \to x$ and a sequence of irrational numbers, $\{x_n'\}$ such that $x_n' \to x$. Now consider the following function.

$$f(x) = \begin{cases} x \text{ if } x \text{ is rational} \\ 0 \text{ if } x \text{ is irrational} \end{cases}.$$

Show using the sequential version of continuity in Theorem 5.11.19 that f is continuous at 0 and nowhere else.

30. The nested interval lemma and Theorem 5.11.19 can be used to give an easy proof of the intermediate value theorem. Suppose $f(a) > 0$ and $f(b) < 0$ for f a continuous function defined on $[a,b]$. The intermediate value theorem states that under these conditions, there exists $x \in (a,b)$ such that $f(x) = 0$. Prove this theorem as follows: Let $c = \frac{a+b}{2}$ and consider the intervals $[a,c]$ and $[c,b]$. Show that on one of these intervals, f is nonnegative at one end and nonpositive at the other. Now consider that interval, divide it in half as was done for the original interval and argue that on one of these smaller intervals, the function has different signs at the two endpoints. Continue in this way. Next apply the nested interval lemma to get x in all these intervals and argue there exist sequences, $x_n \to x$ and $y_n \to x$ such that $f(x_n) < 0$ and $f(y_n) > 0$. By continuity, you can assume $f(x_n) \to f(x)$ and $f(y_n) \to f(x)$. Show this requires that $f(x) = 0$.

31. 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.

32. Show the following converge to 0.

    (a) $\frac{n^5}{1.01^n}$

    (b) $\frac{10^n}{n!}$

33. 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.

34. Suppose $r \in (0,1)$. Show that $\lim_{n\to\infty} r^n = 0$. **Hint:** Use the binomial theorem. $r = \frac{1}{1+\delta}$ where $\delta > 0$. Therefore, $r^n = \frac{1}{(1+\delta)^n} < \frac{1}{1+n\delta}$, etc.

35. Prove $\lim_{n\to\infty} \sqrt[n]{n} = 1$. **Hint:** Let $e_n \equiv \sqrt[n]{n} - 1$ so that $(1+e_n)^n = n$. Now observe that $e_n > 0$ and use the binomial theorem to conclude $1 + ne_n + \frac{n(n-1)}{2}e_n^2 \leq n$. This nice approach to establishing this limit using only elementary algebra is in Rudin [23].

36. Find $\lim_{n\to\infty} (x^n + 5)^{1/n}$ for $x \geq 0$. There are two cases here, $x \leq 1$ and $x > 1$. Show that if $x > 1$, the limit is x while if $x \leq 1$ the limit equals 1. **Hint:** Use the argument of Problem 35. This interesting example is in [8].

5.13 Uniform Continuity

There is a theorem about the integral of a continuous function which requires the notion of uniform continuity. This is discussed in this section. Consider the function $f(x) = \frac{1}{x}$ for $x \in (0,1)$. This is a continuous function because, by Theorem 5.4.1, it is continuous at every point of $(0,1)$. However, for a given $\varepsilon > 0$, the δ needed in the ε, δ definition of continuity becomes very small as x gets close to 0. The notion of uniform continuity involves being able to choose a single δ which works on the whole domain of f. Here is the definition.

Definition 5.13.1 *Let $f : D \subseteq \mathbb{R} \to \mathbb{R}$ be a function. Then f is uniformly continuous if for every $\varepsilon > 0$, there exists a δ **depending only on** ε such that if $|x - y| < \delta$ then $|f(x) - f(y)| < \varepsilon$.*

It is an amazing fact that under certain conditions continuity implies uniform continuity.

Definition 5.13.2 *A set, $K \subseteq \mathbb{R}$ is sequentially compact if whenever $\{a_n\} \subseteq K$ is a sequence, there exists a subsequence, $\{a_{n_k}\}$ such that this subsequence converges to a point of K.*

The following theorem is part of the Heine Borel theorem.

Theorem 5.13.3 *Every closed interval, $[a,b]$ is sequentially compact.*

Proof: Let $\{x_n\} \subseteq [a,b] \equiv I_0$. Consider the two intervals $\left[a, \frac{a+b}{2}\right]$ and $\left[\frac{a+b}{2}, b\right]$ each of which has length $(b-a)/2$. At least one of these intervals contains x_n for infinitely many values of n. Call this interval I_1. Now do for I_1 what was done for I_0. Split it in half and let I_2 be the interval which contains x_n for infinitely many values of n. Continue this way obtaining a sequence of nested intervals $I_0 \supseteq I_1 \supseteq I_2 \supseteq I_3 \cdots$ where the length of I_n is $(b-a)/2^n$. Now pick n_1 such that $x_{n_1} \in I_1$, n_2 such that $n_2 > n_1$ and $x_{n_2} \in I_2, n_3$ such that $n_3 > n_2$ and $x_{n_3} \in I_3$, etc. (This can be done because in each case the intervals contained x_n for infinitely many values of n.) By the nested interval lemma there exists a point, c contained in all these intervals. Furthermore,

$$|x_{n_k} - c| < (b-a)\, 2^{-k}$$

and so $\lim_{k\to\infty} x_{n_k} = c \in [a,b]$. This proves the theorem.

Theorem 5.13.4 *Let $f : K \to \mathbb{R}$ be continuous where K is a sequentially compact set in $\mathbb{R}$. Then f is uniformly continuous on K.*

Proof: If this is not true, there exists $\varepsilon > 0$ such that for every $\delta > 0$ there exists a pair of points, x_δ and y_δ such that even though $|x_\delta - y_\delta| < \delta$, $|f(x_\delta) - f(y_\delta)| \geq \varepsilon$. Taking a succession of values for δ equal to $1, 1/2, 1/3, \cdots$, and letting the exceptional pair of points for $\delta = 1/n$ be denoted by x_n and y_n,

$$|x_n - y_n| < \frac{1}{n}, |f(x_n) - f(y_n)| \geq \varepsilon.$$

Now since K is sequentially compact, there exists a subsequence, $\{x_{n_k}\}$ such that $x_{n_k} \to z \in K$. Now $n_k \geq k$ and so

$$|x_{n_k} - y_{n_k}| < \frac{1}{k}.$$

Consequently, $y_{n_k} \to z$ also. (x_{n_k} is like a person walking toward a certain point and y_{n_k} is like a dog on a leash which is constantly getting shorter. Obviously y_{n_k} must also move toward the point also. You should give a precise proof of what is needed here.) By continuity of f and Problem 13 on Page 122,

$$0 = |f(z) - f(z)| = \lim_{k\to\infty} |f(x_{n_k}) - f(y_{n_k})| \geq \varepsilon,$$

an obvious contradiction. Therefore, the theorem must be true.

The following corollary follows from this theorem and Theorem 5.13.3.

Corollary 5.13.5 *Suppose I is a closed interval, $I = [a,b]$ and $f : I \to \mathbb{R}$ is continuous. Then f is uniformly continuous.*

5.14 Exercises

1. A function, $f : D \subseteq \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous or just Lipschitz for short if there exists a constant, K such that
$$|f(x) - f(y)| \leq K|x - y|$$
for all $x, y \in D$. Show every Lipschitz function is uniformly continuous.

2. If $|x_n - y_n| \to 0$ and $x_n \to z$, show that $y_n \to z$ also.

3. Consider $f : (1, \infty) \to \mathbb{R}$ given by $f(x) = \frac{1}{x}$. Show f is uniformly continuous even though the set on which f is defined is not sequentially compact.

4. If f is uniformly continuous, does it follow that $|f|$ is also uniformly continuous? If $|f|$ is uniformly continuous does it follow that f is uniformly continuous? Answer the same questions with "uniformly continuous" replaced with "continuous". Explain why.

5. Suppose K is a sequentially compact set and $f : K \to \mathbb{R}$. Show that f achieves both its maximum and its minimum on K. **Hint:** Let $M \equiv \sup\{f(x) : x \in K\}$. Argue there exists a sequence, $\{x_n\} \subseteq K$ such that $f(x_n) \to M$. Now use sequential compactness to get a subsequence, $\{x_{n_k}\}$ such that $\lim_{k\to\infty} x_{n_k} = x \in K$ and use the continuity of f to verify that $f(x) = M$. Incidentally, this shows f is bounded on K as well. A similar argument works to give the part about achieving the minimum.

5.15 Theorems About Continuous Functions

In this section, proofs of some theorems which have not been proved yet are given.

Theorem 5.15.1 *The following assertions are valid*

1. *The function, $af + bg$ is continuous at x when f, g are continuous at $x \in D(f) \cap D(g)$ and $a, b \in \mathbb{R}$.*

2. *If and f and g are each real valued functions continuous at x, then fg is continuous at x. If, in addition to this, $g(x) \neq 0$, then f/g is continuous at x.*

3. *If f is continuous at x, $f(x) \in D(g) \subseteq \mathbb{R}$, and g is continuous at $f(x)$, then $g \circ f$ is continuous at x.*

4. *The function $f : \mathbb{R} \to \mathbb{R}$, given by $f(x) = |x|$ is continuous.*

Proof: First consider 1.) Let $\varepsilon > 0$ be given. By assumption, there exist $\delta_1 > 0$ such that whenever $|x - y| < \delta_1$, it follows $|f(x) - f(y)| < \frac{\varepsilon}{2(|a|+|b|+1)}$ and there exists $\delta_2 > 0$ such that whenever $|x - y| < \delta_2$, it follows that $|g(x) - g(y)| < \frac{\varepsilon}{2(|a|+|b|+1)}$. Then let $0 < \delta \leq \min(\delta_1, \delta_2)$. If $|x - y| < \delta$, then everything happens at once. Therefore, using the triangle inequality
$$|af(x) + bf(x) - (ag(y) + bg(y))|$$
$$\leq |a||f(x) - f(y)| + |b||g(x) - g(y)|$$
$$< |a|\left(\frac{\varepsilon}{2(|a| + |b| + 1)}\right) + |b|\left(\frac{\varepsilon}{2(|a| + |b| + 1)}\right) < \varepsilon.$$

Now consider 2.) There exists $\delta_1 > 0$ such that if $|y-x| < \delta_1$, then $|f(x) - f(y)| < 1$. Therefore, for such y,

$$|f(y)| < 1 + |f(x)|.$$

It follows that for such y,

$$|fg(x) - fg(y)| \leq |f(x)g(x) - g(x)f(y)| + |g(x)f(y) - f(y)g(y)|$$

$$\begin{aligned} &\leq |g(x)||f(x) - f(y)| + |f(y)||g(x) - g(y)| \\ &\leq (1 + |g(x)| + |f(y)|)[|g(x) - g(y)| + |f(x) - f(y)|]. \end{aligned}$$

Now let $\varepsilon > 0$ be given. There exists δ_2 such that if $|x-y| < \delta_2$, then

$$|g(x) - g(y)| < \frac{\varepsilon}{2(1 + |g(x)| + |f(y)|)},$$

and there exists δ_3 such that if $|x-y| < \delta_3$, then

$$|f(x) - f(y)| < \frac{\varepsilon}{2(1 + |g(x)| + |f(y)|)}$$

Now let $0 < \delta \leq \min(\delta_1, \delta_2, \delta_3)$. Then if $|x-y| < \delta$, all the above hold at once and so

$$|fg(x) - fg(y)| \leq$$

$$\begin{aligned} &(1 + |g(x)| + |f(y)|)[|g(x) - g(y)| + |f(x) - f(y)|] \\ &< (1 + |g(x)| + |f(y)|)\left(\frac{\varepsilon}{2(1 + |g(x)| + |f(y)|)} + \frac{\varepsilon}{2(1 + |g(x)| + |f(y)|)}\right) = \varepsilon. \end{aligned}$$

This proves the first part of 2.) To obtain the second part, let δ_1 be as described above and let $\delta_0 > 0$ be such that for $|x-y| < \delta_0$,

$$|g(x) - g(y)| < |g(x)|/2$$

and so by the triangle inequality,

$$-|g(x)|/2 \leq |g(y)| - |g(x)| \leq |g(x)|/2$$

which implies $|g(y)| \geq |g(x)|/2$, and $|g(y)| < 3|g(x)|/2$.

Then if $|x-y| < \min(\delta_0, \delta_1)$,

$$\begin{aligned} \left|\frac{f(x)}{g(x)} - \frac{f(y)}{g(y)}\right| &= \left|\frac{f(x)g(y) - f(y)g(x)}{g(x)g(y)}\right| \\ &\leq \frac{|f(x)g(y) - f(y)g(x)|}{\left(\frac{|g(x)|^2}{2}\right)} \\ &= \frac{2|f(x)g(y) - f(y)g(x)|}{|g(x)|^2} \end{aligned}$$

$$\begin{aligned}
&\leq \frac{2}{|g(x)|^2}\left[|f(x)\,g(y)-f(y)\,g(y)+f(y)\,g(y)-f(y)\,g(x)|\right]\\
&\leq \frac{2}{|g(x)|^2}\left[|g(y)|\,|f(x)-f(y)|+|f(y)|\,|g(y)-g(x)|\right]\\
&\leq \frac{2}{|g(x)|^2}\left[\frac{3}{2}|g(x)|\,|f(x)-f(y)|+(1+|f(x)|)\,|g(y)-g(x)|\right]\\
&\leq \frac{2}{|g(x)|^2}(1+2|f(x)|+2|g(x)|)\left[|f(x)-f(y)|+|g(y)-g(x)|\right]\\
&\equiv M\left[|f(x)-f(y)|+|g(y)-g(x)|\right]
\end{aligned}$$

where M is defined by

$$M \equiv \frac{2}{|g(x)|^2}(1+2|f(x)|+2|g(x)|)$$

Now let δ_2 be such that if $|x-y|<\delta_2$, then

$$|f(x)-f(y)|<\frac{\varepsilon}{2}M^{-1}$$

and let δ_3 be such that if $|x-y|<\delta_3$, then

$$|g(y)-g(x)|<\frac{\varepsilon}{2}M^{-1}.$$

Then if $0<\delta\leq\min(\delta_0,\delta_1,\delta_2,\delta_3)$, and $|x-y|<\delta$, everything holds and

$$\left|\frac{f(x)}{g(x)}-\frac{f(y)}{g(y)}\right|\leq M\left[|f(x)-f(y)|+|g(y)-g(x)|\right]$$

$$<M\left[\frac{\varepsilon}{2}M^{-1}+\frac{\varepsilon}{2}M^{-1}\right]=\varepsilon.$$

This completes the proof of the second part of 2.)

Note that in these proofs no effort is made to find some sort of "best" δ. The problem is one which has a yes or a no answer. Either is it or it is not continuous.

Now consider 3.). If f is continuous at x, $f(x)\in D(g)\subseteq\mathbb{R}^p$, and g is continuous at $f(x)$,then $g\circ f$ is continuous at x. Let $\varepsilon>0$ be given. Then there exists $\eta>0$ such that if $|y-f(x)|<\eta$ and $y\in D(g)$, it follows that $|g(y)-g(f(x))|<\varepsilon$. From continuity of f at x, there exists $\delta>0$ such that if $|x-z|<\delta$ and $z\in D(f)$, then $|f(z)-f(x)|<\eta$. Then if $|x-z|<\delta$ and $z\in D(g\circ f)\subseteq D(f)$, all the above hold and so

$$|g(f(z))-g(f(x))|<\varepsilon.$$

This proves part 3.)

To verify part 4.), let $\varepsilon>0$ be given and let $\delta=\varepsilon$. Then if $|x-y|<\delta$, the triangle inequality implies

$$\begin{aligned}
|f(x)-f(y)| &= ||x|-|y||\\
&\leq |x-y|<\delta=\varepsilon.
\end{aligned}$$

This proves part 4.) and completes the proof of the theorem.

Next here is a proof of the intermediate value theorem.

Theorem 5.15.2 *Suppose $f : [a,b] \to \mathbb{R}$ is continuous and suppose $f(a) < c < f(b)$. Then there exists $x \in (a,b)$ such that $f(x) = c$.*

Proof: Let $d = \frac{a+b}{2}$ and consider the intervals $[a,d]$ and $[d,b]$. If $f(d) \geq c$, then on $[a,d]$, the function is $\leq c$ at one end point and $\geq c$ at the other. On the other hand, if $f(d) \leq c$, then on $[d,b]$ $f \geq 0$ at one end point and ≤ 0 at the other. Pick the interval on which f has values which are at least as large as c and values no larger than c. Now consider that interval, divide it in half as was done for the original interval and argue that on one of these smaller intervals, the function has values at least as large as c and values no larger than c. Continue in this way. Next apply the nested interval lemma to get x in all these intervals. In the n^{th} interval, let x_n, y_n be elements of this interval such that $f(x_n) \leq c, f(y_n) \geq c$. Now $|x_n - x| \leq (b-a)\,2^{-n}$ and $|y_n - x| \leq (b-a)\,2^{-n}$ and so $x_n \to x$ and $y_n \to x$. Therefore,

$$f(x) - c = \lim_{n\to\infty} (f(x_n) - c) \leq 0$$

while

$$f(x) - c = \lim_{n\to\infty} (f(y_n) - c) \geq 0.$$

Consequently $f(x) = c$ and this proves the theorem. (For the last step, see Problem 13 on Page 122).

Lemma 5.15.3 *Let $\phi : [a,b] \to \mathbb{R}$ be a continuous function and suppose ϕ is $1-1$ on (a,b). Then ϕ is either strictly increasing or strictly decreasing on $[a,b]$.*

Proof: First it is shown that ϕ is either strictly increasing or strictly decreasing on (a,b).

If ϕ is not strictly decreasing on (a,b), then there exists $x_1 < y_1$, $x_1, y_1 \in (a,b)$ such that

$$(\phi(y_1) - \phi(x_1))(y_1 - x_1) > 0.$$

If for some other pair of points, $x_2 < y_2$ with $x_2, y_2 \in (a,b)$, the above inequality does not hold, then since ϕ is $1-1$,

$$(\phi(y_2) - \phi(x_2))(y_2 - x_2) < 0.$$

Let $x_t \equiv tx_1 + (1-t)\,x_2$ and $y_t \equiv ty_1 + (1-t)\,y_2$. Then $x_t < y_t$ for all $t \in [0,1]$ because

$$tx_1 \leq ty_1 \text{ and } (1-t)\,x_2 \leq (1-t)\,y_2$$

with strict inequality holding for at least one of these inequalities since not both t and $(1-t)$ can equal zero. Now define

$$h(t) \equiv (\phi(y_t) - \phi(x_t))(y_t - x_t).$$

Since h is continuous and $h(0) < 0$, while $h(1) > 0$, there exists $t \in (0,1)$ such that $h(t) = 0$. Therefore, both x_t and y_t are points of (a,b) and $\phi(y_t) - \phi(x_t) = 0$ contradicting the assumption that ϕ is one to one. It follows ϕ is either strictly increasing or strictly decreasing on (a,b).

This property of being either strictly increasing or strictly decreasing on (a,b) carries over to $[a,b]$ by the continuity of ϕ. Suppose ϕ is strictly increasing on (a,b), a similar argument holding for ϕ strictly decreasing on (a,b). If $x > a$, then pick $y \in (a,x)$ and from the above, $\phi(y) < \phi(x)$. Now by continuity of ϕ at a,

$$\phi(a) = \lim_{x\to a+} \phi(z) \leq \phi(y) < \phi(x).$$

Therefore, $\phi(a) < \phi(x)$ whenever $x \in (a,b)$. Similarly $\phi(b) > \phi(x)$ for all $x \in (a,b)$. This proves the lemma.

Corollary 5.15.4 *Let $f : (a,b) \to \mathbb{R}$ be one to one and continuous. Then $f(a,b)$ is an open interval, (c,d) and $f^{-1} : (c,d) \to (a,b)$ is continuous.*

Proof: Since f is either strictly increasing or strictly decreasing, it follows that $f(a,b)$ is an open interval, (c,d). Assume f is decreasing. Now let $x \in (a,b)$. Why is f^{-1} is continuous at $f(x)$? Since f is decreasing, if $f(x) < f(y)$, then $y \equiv f^{-1}(f(y)) < x \equiv f^{-1}(f(x))$ and so f^{-1} is also decreasing. Let $\varepsilon > 0$ be given. Let $\varepsilon > \eta > 0$ and $(x-\eta, x+\eta) \subseteq (a,b)$. Then $f(x) \in (f(x+\eta), f(x-\eta))$. Let $\delta = \min(f(x) - f(x+\eta), f(x-\eta) - f(x))$. Then if

$$|f(z) - f(x)| < \delta,$$

it follows

$$z \equiv f^{-1}(f(z)) \in (x-\eta, x+\eta) \subseteq (x-\varepsilon, x+\varepsilon)$$

so

$$\left|f^{-1}(f(z)) - x\right| = \left|f^{-1}(f(z)) - f^{-1}(f(x))\right| < \varepsilon.$$

This proves the theorem in the case where f is strictly decreasing. The case where f is increasing is similar.

Derivatives

6.0.1 Outcomes

1. Understand the relation between velocity and the derivative.
2. Define and understand the definition of the derivative, both graphically and in terms of a limit. Be able to find derivatives using this definition.
3. Define and use the derivative to find local extrema.
4. Understand the proof of the mean value theorem, both the Lagrange and the Cauchy version.
5. Use the derivative as an aid to curve sketching.

6.1 Velocity

Imagine an object which is moving along the real line in the positive direction and that at time $t > 0$, the position of the object is $r(t) = -10 + 30t + t^2$ where distance is measured in kilometers and t in hours. Thus at $t = 0$, the object is at the point -10 kilometers and when $t = 1$, the object is at 21 kilometers. The average velocity during this time is the distance traveled divided by the elapsed time. Thus the average velocity would be $\frac{21-(-10)}{1} = 31$ kilometers per hour. It came out positive because the object moved in the positive direction along the real line, from -10 to 21. Suppose it was desired to find something which deserves to be referred to as the instantaneous velocity when $t = 1/2$? If the object were a car, it is reasonable to suppose that the magnitude of the average velocity of the object over a very small interval of time would be very close to the number that would appear on the speedometer. For example, if considering the average velocity of the object on the interval $[.5, .5 + .0001]$, this average velocity would be pretty close to the thing which deserves to be called the instantaneous velocity at $t = .5$ hours. Thus the velocity at $t = .5$ would be close to

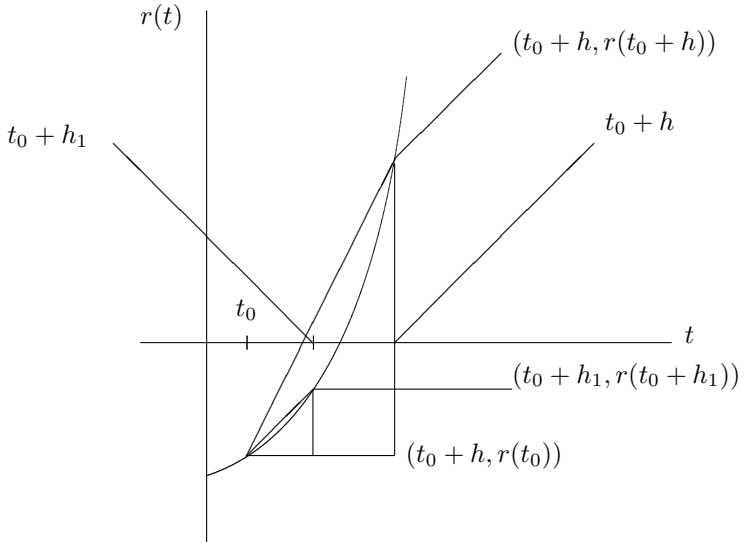
$$\left(r\left(.5 + .0001\right) - r\left(.5\right)\right)/.0001$$

$$= \left(30\left(.5 + .01\right) + \left(.5 + .0001\right)^2 - 30\left(.5\right) - \left(.5\right)^2\right)/.0001 = 31.\,000\,1$$

Of course, you would expect to be even closer using a time interval of length .000001 instead of just .0001. In general, consider a time interval of length h and then define the instantaneous velocity to be the number which all these average velocities get close to as h gets smaller and smaller. Thus in this case form the average velocity on the interval, $[.5, .5 + h]$ to get

$$\left(30\left(.5 + h\right) + \left(.5 + h\right)^2 - \left(30\left(.5\right) + \left(.5\right)^2\right)\right)/h = 30 + 2\left(.5\right) + h.$$

What number does this average get close to as h gets smaller and smaller? Clearly it gets close to 31 and for this reason, the velocity at time .5 is defined as 31. It is positive because the object is moving in the positive direction. If the object were moving in the negative direction, the number would be negative. The notion just described of finding an instantaneous velocity has a geometrical application to finding the slope of a line tangent to a curve.



In the above picture, you see the slope of the line joining the two points $(t_0, r(t_0))$ and $(t_0 + h, r(t_0 + h))$ is given by

$$\frac{r(t_0 + h) - r(t_0)}{h}$$

which equals the average velocity on the time interval, $[t_0, t_0 + h]$. You can also see the effect of making h closer and closer to zero as illustrated by changing h to the smaller h_1 in the picture. The slope of the resulting line segment appears to get closer and closer to what ought to be considered the slope of the line tangent to the curve at the point $(t_0, r(t_0))$.

It is time to make this heuristic material much more precise.

6.2 The Derivative

The derivative of a function of one variable is a function given by the following definition.

Definition 6.2.1 *The derivative of a function, $f'(x)$, is defined as the following limit whenever the limit exists. If the limit does not exist, then neither does $f'(x)$.*

$$\lim_{h \to 0} \frac{f(x + h) - f(x)}{h} \equiv f'(x) \tag{6.1}$$

The function of h on the left is called the difference quotient.

Note that the difference quotient on the left of the equation is a function of h which is not defined at $h = 0$. This is why, in the definition of limit, $|h| > 0$. **It is not necessary to have the function defined at the point in order to consider its limit.** The

distinction between the limit of a function and its value is very important and must be kept in mind. Also it is clear from setting $y = x + h$ that

$$f'(x) = \lim_{y \to x} \frac{f(y) - f(x)}{y - x}. \tag{6.2}$$

Theorem 6.2.2 *If $f'(x)$ exists, then f is continuous at x.*

Proof: Suppose $\varepsilon > 0$ is given and choose $\delta_1 > 0$ such that if $|h| < \delta_1$,

$$\left| \frac{f(x+h) - f(x)}{h} - f'(x) \right| < 1.$$

then for such h, the triangle inequality implies

$$|f(x+h) - f(x)| < |h| + |f'(x)|\,|h|\,.$$

Now letting $\delta < \min\left(\delta_1, \frac{\varepsilon}{1+|f'(x)|}\right)$ it follows if $|h| < \delta$, then

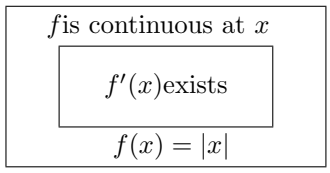
$$|f(x+h) - f(x)| < \varepsilon.$$

Letting $y = h + x$, this shows that if $|y - x| < \delta$,

$$|f(y) - f(x)| < \varepsilon$$

which proves f is continuous at x.

It is very important to remember that just because f is continuous, does not mean f has a derivative. The following picture describes the situation.



As indicated in the above picture the function $f(x) = |x|$ does not have a derivative at $x = 0$. To see this,

$$\lim_{h \to 0+} \frac{f(h) - f(0)}{h} = \lim_{h \to 0+} \frac{h}{h} = 1$$

while

$$\lim_{h \to 0-} \frac{f(h) - f(0)}{h} = \lim_{h \to 0-} \frac{-h}{h} = -1.$$

Thus the two limits, one from the right and one from the left do not agree as they would have to do if the function had a derivative at $x = 0$. See Problem 18 on Page 114. Geometrically, this lack of differentiability is manifested by there being a pointy place in the graph of $y = |x|$ at $x = 0$. In short, the pointy places don't have derivatives.

Example 6.2.3 *Let $f(x) = c$ where c is a constant. Find $f'(x)$.*

Set up the difference quotient,

$$\frac{f(x+h) - f(x)}{h} = \frac{c - c}{h} = 0$$

Therefore,

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} 0 = 0$$

Example 6.2.4 *Let $f(x) = cx$ where c is a constant. Find $f'(x)$.*

Set up the difference quotient,

$$\frac{f(x+h)-f(x)}{h} = \frac{c(x+h)-cx}{h} = \frac{ch}{h} = c.$$

Therefore,

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} c = c.$$

Example 6.2.5 *Let $f(x) = \sqrt{x}$ for $x > 0$. Find $f'(x)$.*

Set up the difference quotient,

$$\begin{aligned}\frac{f(x+h)-f(x)}{h} &= \frac{\sqrt{x+h}-\sqrt{x}}{h} = \frac{x+h-x}{h\left(\sqrt{x+h}+\sqrt{x}\right)}\\ &= \frac{1}{\sqrt{x+h}+\sqrt{x}}\end{aligned}$$

and so

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = \lim_{h\to 0}\frac{1}{\sqrt{x+h}+\sqrt{x}} = \frac{1}{2\sqrt{x}}.$$

There are rules of derivatives which make finding the derivative very easy.

Theorem 6.2.6 *Let $a, b \in \mathbb{R}$ and suppose $f'(t)$ and $g'(t)$ exist. Then the following formulas are obtained.*

$$(af+bg)'(t) = af'(t) + bg'(t). \tag{6.3}$$

$$(fg)'(t) = f'(t)\,g(t) + f(t)\,g'(t). \tag{6.4}$$

The formula, (6.4) is referred to as the product rule.

If $g(t) \neq 0$,

$$\left(\frac{f}{g}\right)'(t) = \frac{f'(t)\,g(t) - g'(t)\,f(t)}{g^2(t)}. \tag{6.5}$$

Formula (6.5) is referred to as the quotient rule.

If f is differentiable at ct where $c \neq 0$, Then letting $g(t) \equiv f(ct)$,

$$g'(t) = cf'(ct). \tag{6.6}$$

Written with a slight abuse of notation,

$$(f(ct))' = cf'(ct). \tag{6.7}$$

If f is differentiable on (a,b) and if $g(t) \equiv f(t+c)$, then g is differentiable on $(a-c, b-c)$ and

$$g'(t) = f'(t+c). \tag{6.8}$$

Written with a slight abuse of notation,

$$(f(t+c))' = f'(t+c) \tag{6.9}$$

For p an integer and $f\,'(t)$ exists, let $g_p(t) \equiv f(t)^p$. Then

$$(g_p)'(t) = pf(t)^{p-1} f'(t). \tag{6.10}$$

(In the case where $p < 0$, assume $f(t) \neq 0$.)

Written with a slight abuse of notation, an easy to remember version of (6.10) says

$$(f(t)^p)' = pf(t)^{p-1} f'(t).$$

Proof: The first formula is left for you to prove. Consider the second, (6.4).

$$\begin{aligned}\frac{fg(t+h)-fg(t)}{h} &= \frac{f(t+h)g(t+h)-f(t+h)g(t)}{h}+\frac{f(t+h)g(t)-f(t)g(t)}{h}\\ &= f(t+h)\frac{(g(t+h)-g(t))}{h}+\frac{(f(t+h)-f(t))}{h}g(t)\end{aligned}$$

Taking the limit as $h \to 0$ and using Theorem 6.2.2 to conclude $\lim_{h\to 0} f(t+h) = f(t)$, it follows from Theorem 5.9.4 that (6.4) follows. Next consider the quotient rule.

$$\begin{aligned}h^{-1}\left(\frac{f}{g}(t+h)-\frac{f}{g}(t)\right) &= \frac{f(t+h)g(t)-g(t+h)f(t)}{hg(t)g(t+h)}\\ &= \frac{f(t+h)(g(t)-g(t+h))+g(t+h)(f(t+h)-f(t))}{hg(t)g(t+h)}\\ &= \frac{-f(t+h)}{g(t)g(t+h)}\frac{(g(t+h)-g(t))}{h}+\frac{g(t+h)}{g(t)g(t+h)}\frac{(f(t+h)-f(t))}{h}\end{aligned}$$

and from Theorem 5.9.4 on Page 110,

$$\left(\frac{f}{g}\right)'(t) = \frac{g(t)f'(t)-g'(t)f(t)}{g^2(t)}.$$

Now consider Formula (6.6).

$$\begin{aligned}(g(t+h)-g(t))h^{-1} &= h^{-1}(f(ct+ch)-f(ct))\\ &= c\frac{f(ct+ch)-f(ct)}{ch}\\ &= c\frac{f(ct+h_1)-f(ct)}{h_1}\end{aligned}$$

where $h_1 = ch$. Then $h_1 \to 0$ if and only if $h \to 0$ and so taking the limit as $h \to 0$ yields

$$g'(t) = cf'(ct)$$

as claimed. Formulas (6.7) and (6.8) are left as an exercise.

First consider (6.10) in the case where p equals a nonnegative integer. If $p = 0$, (6.10) holds because $g_0(t) = 1$ and so by Example 6.2.3,

$$g_0'(t) = 0 = 0(f(t))^{-1} f'(t).$$

Next suppose (6.10) holds for p an integer. Then

$$(g_{p+1}(t)) = f(t)g_p(t)$$

and so by the product rule,

$$\begin{aligned}g_{p+1}'(t) &= f'(t)g_p(t)+f(t)g_p'(t)\\ &= f'(t)(f(t))^p + f(t)\left(pf(t)^{p-1}f'(t)\right)\\ &= (p+1)f(t)^p f'(t).\end{aligned}$$

If the formula holds for some integer, p then it holds for $-p$. Here is why.

$$g_{-p}(t) = g_p(t)^{-1}$$

and so

$$\frac{g_{-p}(t+h)-g_{-p}(t)}{h}=\left(\frac{g_p(t)-g_p(t+h)}{h}\right)\left(\frac{1}{g_p(t)\,g_p(t+h)}\right).$$

Taking the limit as $h\to 0$ and using the formula for p,

$$\begin{aligned}g'_{-p}(t)&=-pf(t)^{p-1}f'(t)(f(t))^{-2p}\\&=-p(f(t))^{-p-1}f'(t).\end{aligned}$$

This proves the theorem.

Example 6.2.7 *Let $p(x)=3+5x+6x^2-7x^3$. Find $p'(x)$.*

From the above theorem, and abusing the notation,

$$\begin{aligned}p'(x)&=\left(3+5x+6x^2-7x^3\right)'\\&=3'+(5x)'+\left(6x^2\right)'+\left(-7x^3\right)'\\&=0+5+(6)(2)(x)(x)'+(-7)(3)\left(x^2\right)(x)'\\&=5+12x-21x^2.\end{aligned}$$

Note the process is to take the exponent and multiply by the coefficient and then make the new exponent one less in each term of the polynomial in order to arrive at the answer. This is the general procedure for differentiating a polynomial as shown in the next example.

Example 6.2.8 *Let a_k be a number for $k=0,1,\cdots,n$ and let $p(x)=\sum_{k=0}^n a_kx^k$. Find $p'(x)$.*

Use Theorem 6.2.6

$$\begin{aligned}p'(x)&=\left(\sum_{k=0}^n a_kx^k\right)'=\sum_{k=0}^n a_k\left(x^k\right)'\\&=\sum_{k=0}^n a_kkx^{k-1}(x)'=\sum_{k=0}^n a_kkx^{k-1}\end{aligned}$$

Example 6.2.9 *Find the derivative of the function $f(x)=\frac{x^2+1}{x^3}$.*

Use the quotient rule

$$f'(x)=\frac{2x\left(x^3\right)-3x^2\left(x^2+1\right)}{x^6}=-\frac{1}{x^4}\left(x^2+3\right)$$

Example 6.2.10 *Let $f(x)=\left(x^2+1\right)^4\left(x^3\right)$. Find $f'(x)$.*

Use the product rule and (6.10). Abusing the notation for the sake of convenience,

$$\begin{aligned}\left(\left(x^2+1\right)^4\left(x^3\right)\right)'&=\left(\left(x^2+1\right)^4\right)'\left(x^3\right)+\left(x^3\right)'\left(\left(x^2+1\right)^4\right)\\&=4\left(x^2+1\right)^3(2x)\left(x^3\right)+3x^2\left(x^2+1\right)^4\\&=4x^4\left(x^2+1\right)^3+3x^2\left(x^2+1\right)^4\end{aligned}$$

Example 6.2.11 *Let $f(x)=x^3/\left(x^2+1\right)^2$. Find $f'(x)$.*

Use the quotient rule to obtain

$$f'(x) = \frac{3x^2\left(x^2+1\right)^2 - 2\left(x^2+1\right)(2x)x^3}{\left(x^2+1\right)^4} = \frac{3x^2\left(x^2+1\right)^2 - 4x^4\left(x^2+1\right)}{\left(x^2+1\right)^4}$$

Obviously, one could consider taking the derivative of the derivative and then the derivative of that and so forth. The main thing to consider about this is the notation. The second derivative is denoted with two primes.

Example 6.2.12 *Let* $f(x) = x^3 + 2x^2 + 1$. *Find* $f''(x)$ *and* $f'''(x)$.

To find $f''(x)$ take the derivative of the derivative. Thus $f'(x) = 3x^2 + 4x$ and so $f''(x) = 6x + 4$. Then $f'''(x) = 6$.

When high derivatives are taken, say the 5^{th} derivative, it is customary to write $f^{(5)}(t)$ putting the number of derivatives in parentheses.

6.3 Exercises With Answers

1. For $f(x) = \frac{x^3+6x+2}{x^2+2}$, find $f'(x)$.

   Answer:

   $-\frac{-x^4-12+4x}{(x^2+2)^2}$

2. For $f(x) = \left(3x^3+6x+3\right)\left(3x^2+3x+9\right)$, find $f'(x)$.

   Answer:

   $\left(9x^2+6\right)\left(3x^2+3x+9\right) + \left(3x^3+6x+3\right)(6x+3)$

3. For $f(x) = \sqrt{5x^2+1}$, find $f'(x)$ from the definition of the derivative.

   Answer:

   $\frac{5x}{\sqrt{(5x^2+1)}}$

4. For $f(x) = \sqrt[3]{6x^2+1}$, find $f'(x)$ from the definition of the derivative.

   Answer:

   $\frac{4x}{\left(\sqrt[3]{(6x^2+1)}\right)^2}$

5. For $f(x) = (-5x+2)^9$, find $f'(x)$ from the definition of the derivative. **Hint:** You might use the formula

   $b^n - a^n = (b-a)\left(b^{n-1} + b^{n-2}a + \cdots + a^{n-2}b + b^{n-1}\right)$

   Answer:

   $-45(-5x+2)^8$

6. Let $f(x) = (x+5)^2 \sin(1/(x+5)) + 6(x-2)(x+5)$ for $x \neq -5$ and define $f(-5) \equiv 0$. Find $f'(-5)$ from the definition of the derivative if this is possible. **Hint:** Note that $|\sin(z)| \leq 1$ for any real value of z.

   Answer:

   -42

6.4 Exercises

1. Find derivatives of the following functions.
   (a) $3x^2 + 4x^3 - 7x + 11$
   (b) $\frac{1}{x^3+1}$
   (c) $\left(x^4 + x + 5\right)\left(x^5 + 7x\right)$
   (d) $\frac{x^3+x}{x^2+1}$
   (e) $\left(x^2 + 2\right)^4$
   (f) $\left(x^3 + 2x + 1\right)^{-1}$
2. For $f(x) = -3x^7 + 4x^5 + 2x^3 + x^2 - 5x$, find $f^{(3)}(x)$.
3. For $f(x) = -x^7 + x^5 + x^3 - 2x$, find $f^{(3)}(x)$.
4. Find $f'(x)$ for the given functions
   (a) $f(x) = \frac{3x^3-x-1}{3x^2+1}$
   (b) $f(x) = \frac{3x^3+3x-1}{x^2+1}$
   (c) $f(x) = \frac{\left(x^2+5\right)^4}{(x^2+1)^2}$
   (d) $f(x) = \left(4x + 7x^3\right)^6$
   (e) $f(x) = \left(-2x^3 + x + 3\right)^3\left(-2x^2 + 3x - 1\right)$
5. For $f(x) = \sqrt{4x^2+1}$, find $f'(x)$ from the definition of the derivative.
6. For $f(x) = \sqrt[3]{3x^2+1}$, find $f'(x)$ from the definition of the derivative. **Hint:** You might use
   $b^3 - a^3 = \left(b^2 + ab + a^2\right)(b - a)$ for $b = \sqrt[3]{3(x+h)^2 + 1}$ and $a = \sqrt[3]{3x^2+1}$.
7. For $f(x) = (-3x + 5)^5$, find $f'(x)$ from the definition of the derivative. **Hint:** You might use the formula
   $b^n - a^n = (b - a)\left(b^{n-1} + b^{n-2}a + \cdots + a^{n-2}b + b^{n-1}\right)$
8. Let $f(x) = (x+2)^2 \sin(1/(x+2)) + 6(x-5)(x+2)$ for $x \neq -2$ and define $f(-2) \equiv 0$. Find $f'(-2)$ from the definition of the derivative if this is possible. **Hint:** Note that $|\sin(z)| \leq 1$ for any real value of z.
9. Let $f(x) = (x+5)\sin(1/(x+5))$ for $x \neq -5$ and define $f(-5) \equiv 0$. Show $f'(-5)$ does not exist. **Hint:** Verify that $\lim_{h\to 0} \sin(1/h)$ does not exist and then explain why this shows $f'(-5)$ does not exist.
10. Suppose f is a continuous function and $y = mx + b$ is the equation of a straight line with slope m which intersects the graph of f only at the point $(x, f(x))$. Does it follow that $f'(x) = m$?
11. Give an example of a function, f which is not continuous but for which $\lim_{h\to 0} \frac{f(x+h)-f(x-h)}{2h}$ exists for every x.

12. Suppose f is defined on (a,b) and $f'(x)$ exists for some $x \in (a,b)$. Show $f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x-h)}{2h}$.

13. Let $f(x) = x^2$ if x is rational and $f(x) = 0$ if x is irrational. Show that $f'(0) = 0$ but that f fails to be continuous at every other point.

14. Suppose $|f(x)| \leq x^2$ for all $x \in (-1,1)$. Let $h(x) = x + f(x)$. Find $h'(0)$.

15. Find a formula for the derivative of a product of three functions. Generalize to a product of n functions where n is a positive integer.

16. From Corollary 3.5.4 on Page 62 the following inequality holds.

$$\sin x + (1 - \cos x) \geq x \geq \sin x. \tag{6.11}$$

Using this inequality, show

$$\lim_{x\to 0} \frac{\sin x}{x} = 1.$$

Hint: For $x > 0$, divide both sides of the inequality by $\sin x$. This yields

$$1 + \frac{1-\cos x}{x} \geq \frac{x}{\sin x} \geq 1.$$

Now

$$0 \leq \frac{1-\cos x}{x} = \frac{1-\cos^2 x}{x(1+\cos x)} = \frac{\sin^2 x}{x(1+\cos x)} \leq \frac{\sin x}{(1+\cos x)}. \tag{6.12}$$

If $x < 0$, $\frac{\sin s}{x} = \frac{\sin(-x)}{(-x)}$ and $-x > 0$.

17. Show that $\lim_{h\to 0}\left(\frac{1-\cosh}{h}\right) = 0$. **Hint:** For $h > 0$, consider the inequality (6.12) with x replaced with h.

18. Show using Problems 16 and 17 and the definition of the derivative that $\sin'(x) = \cos x$. Also show $\cos'(x) = -\sin(x)$. **Hint:** Just right down the difference quotient and use the limits in these problems.

$$\frac{\sin(x+h)-\sin x}{h} = \frac{\sin(x)\cos(h)+\cos(x)\sin(h)-\sin(x)}{h}, \text{ etc.}$$

19. Now that you know $\sin'(x) = \cos(x)$, use the definition of derivative to find $f'(x)$ where $f(x) = \sin(3x)$. What about $g'(x)$ where $g(x) = \cos(3x)$? Now show that both $\cos(\omega t)$ and $\sin(\omega t)$ are solutions to the differential equation, $y'' + \omega^2 y = 0$. This is the equation for undamped oscillations and it will be discussed more later.

6.5 Local Extrema

When you are on top of a hill, you are at a local maximum although there may be other hills higher than the one on which you are standing. Similarly, when you are at the bottom of a valley, you are at a local minimum even though there may be other valleys deeper than the one you are in. The word, "local" is applied to the situation because if you confine your attention only to points close to your location, you are indeed at either the top or the bottom.

Definition 6.5.1 *Let $f : D(f) \to \mathbb{R}$ where here $D(f)$ is only assumed to be some subset of $\mathbb{R}$. Then $x \in D(f)$ is a local minimum (maximum) if there exists $\delta > 0$ such that whenever $y \in (x-\delta, x+\delta) \cap D(f)$, it follows $f(y) \geq (\leq) f(x)$.*

Derivatives can be used to locate local maximums and local minimums.

Theorem 6.5.2 *Suppose $f : (a,b) \to \mathbb{R}$ and suppose $x \in (a,b)$ is a local maximum or minimum. Then $f'(x) = 0$.*

Proof: Suppose x is a local maximum. If $h > 0$ and is sufficiently small, then $f(x+h) \leq f(x)$ and so from Theorem 5.9.4 on Page 110,

$$f'(x) = \lim_{h \to 0+} \frac{f(x+h) - f(x)}{h} \leq 0.$$

Similarly,

$$f'(x) = \lim_{h \to 0-} \frac{f(x+h) - f(x)}{h} \geq 0.$$

The case when x is local minimum is similar. This proves the theorem.

Definition 6.5.3 *Points where the derivative of a function equals zero are called critical points. It is also customary to refer to points where the derivative of a function does not exist as a critical points.*

Example 6.5.4 *It is desired to find two positive numbers whose sum equals 16 and whose product is to be a large as possible.*

The numbers are x and $16 - x$ and $f(x) = x(16 - x)$ is to be made as large as possible. The value of x which will do this would be a local maximum so by Theorem 6.5.2 the procedure is to take the derivative of f and find values of x where it equals zero. Thus $16 - 2x = 0$ and the only place this occurs is when $x = 8$. Therefore, the two numbers are 8 and 8.

Example 6.5.5 *A farmer wants to fence a rectangular piece of land next to a straight river. What are the dimensions of the largest rectangle if there are exactly 600 meters of fencing available.*

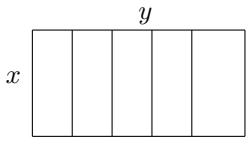
The two sides perpendicular to the river have length x and the third side has length $y = (600 - 2x)$. Thus the function to be maximized is $f(x) = 2x(600 - x) = 1200x - 2x^2$. Taking the derivative and setting it equal to zero gives

$$f'(x) = 1200 - 4x = 0$$

and so $x = 300$. Therefore, the desired dimensions are 300×600.

Example 6.5.6 *A rectangular playground is to be enclosed by a fence and divided in* 5 *pieces by* 4 *fences parallel to one side of the playground.* 1704 *feet of fencing is used. Find dimensions of the playground which will have the largest total area.*

Let x denote the length of one of these dividing fences and let y denote the length of the playground as shown in the following picture.



Thus $6x+2y=1704$ so $y=\frac{1704-6x}{2}$ and the function to maximize is

$$f(x)=x\left(\frac{1704-6x}{2}\right)=x(852-3x).$$

Therefore, to locate the value of x which will make $f(x)$ as large as possible, take $f'(x)$ and set it equal to zero.

$$852-6x=0$$

and so $x=142$ feet and $y=\frac{1704-6\times142}{2}=426$ feet.

Revenue is defined to be the amount of money obtained in some transaction. Profit is defined as the revenue minus the costs.

Example 6.5.7 *Francine, the manager of Francine's Fancy Shakes finds that at* \$4, *demand for her milk shakes is* 900 *per day. For each* \$.30 *increase in price, the demand decreases by* 50. *Find the price and the quantity sold which maximizes revenue.*

Let x be the number of \$.30 increases. Then the total number sold is $(900-50x)$. The revenue is

$$R(x)=(900-50x)(4+.3x).$$

Then to maximize it, $R'(x)=70-30x=0$. The solution is $x=2.33$ and so the optimum price is at \$4.70 and the number sold will be 783.

Example 6.5.8 *Sam, the owner of Spider Sam's Tarantulas and Creepy Critters finds he can sell* 6 *tarantulas every day at the regular price of* \$30 *each. At his last spider celebration sale he reduced the price to* \$24 *and was able to sell* 12 *tarantulas every day. He has to pay* \$.05 *per day to exhibit a tarantula and his fixed costs are* \$30 *per day, mainly to maintain the thousands of tarantulas he keeps on his tarantula breeding farm. What price should he charge to maximize his profit.*

He assumes the demand for tarantulas is a linear function of price. Thus if y is the number of tarantulas demanded at price x, it follows $y=36-x$. Therefore, the revenue for price x equals $R(x)=(36-x)x$. Now you have to subtract off the costs to get the profit. Thus

$$P(x)=(36-x)x-(36-x)(.05)-30.$$

It follows the profit is maximized when $P'(x)=0$ so $-2.0x+36.05=0$ which occurs when $x=\$18.025$. Thus Sam should charge about \$18 per tarantula.

Exercise 6.5.9 *Lisa, the owner of Lisa's gags and gadgets sells* 500 *whoopee cushions per year. It costs* \$.25 *per year to store a whoopee cushion. To order whoopee cushions it costs* \$8 *plus* \$.90 *per cushion. How many times a year and in what lot size should whoopee cushions be ordered to minimize inventory costs?*

Let x be the times per year an order is sent for a lot size of $\frac{500}{x}$. If the demand is constant, it is reasonable to suppose there are about $\frac{500}{2x}$ whoopee cushions which have to be stored. Thus the cost to store whoopee cushions is $\frac{125.0}{2x}=.25\left(\frac{500}{2x}\right)$. Each time an order is made for a lot size of $\frac{500}{x}$ it costs $8+\frac{450.0}{x}=8+.9\left(\frac{500}{x}\right)$ and this is done x times a year. Therefore, the total inventory cost is $x\left(8+\frac{450.0}{x}\right)+\frac{125.0}{2x}=C(x)$. The problem is to minimize $C(x)=x\left(8+\frac{450.0}{x}\right)+\frac{125.0}{2x}$. Taking the derivative yields

$$C'(x)=\frac{16x^2-125}{2x^2}$$

and so the value of x which will minimize $C(x)$ is $\frac{5}{4}\sqrt{5} = 2.79\cdots$ and the lot size is $\frac{500}{\frac{5}{4}\sqrt{5}} = 178..88$. Of course you would round these numbers off. Order 179 whoopee cushions 3 times a year.

6.6 Exercises With Answers

1. Find the x values of the critical points of the function $f(x) = x^2 - x^3$.

   Answer:

   $\frac{2}{3}, 0$

2. Find the x values of the critical points of the function $f(x) = \sqrt{3x^2 - 6x + 6}$.

   Answer:

   1

3. Find the extreme points of the function, $f(x) = 2x^2 - 20x + 55$ and tell whether the extreme point is a maximum or a minimum.

   Answer:

   The extremum is at $x = 5$. It is a maximum.

4. Find the extreme points of the function, $f(x) = x + \frac{9}{x}$ and tell whether the extreme point is a local maximum or a local minimum or neither.

   Answer:

   The extrema are at $x = \pm 3$. The one at 3 is a local minimum and the one at -3 is a local maximum.

5. A rectangular pasture is to be fenced off beside a river with no need of fencing along the river. If there is900 yards of fencing material, what are the dimensions of the largest possible pasture that can be enclosed?

   Answer:

   225×45

6. A piece of property is to be fenced on the front and two sides. Fencing for the sides costs \$3.50 per foot and fencing for the front costs \$5.60 per foot. What are the dimensions of the largest such rectangular lot if the available money is \$840.0?

   Answer:

   60×75.

7. In a particular apartment complex of 120 units, it is found that all units remain occupied when the rent is \$300 per month. For each \$30 increase in the rent, one unit becomes vacant, on the average. Occupied units require \$60 per month for maintenance, while vacant units require none. Fixed costs for the buildings are \$30 000 per month. What rent should be charged for maximum profit and what is the maximum profit?

   Answer:

   Need to maximize $f(x) = (300 + 30x)(120 - x) - 37\,200 + 60x$ for $x \in [0, 120]$ where x is the number of \$30 increases in rent.

   \$92 880 when the rent is \$1980.

8. A picture is 5 feet high and the eye level of an observer is 2 feet below the bottom edge of the picture. How far from the picture should the observer stand if he wants to maximize the angle subtended by the picture?

   Answer:

   Let the angle subtended by the picture be θ and let α denote the angle between a horizontal line from the observer's eye to the wall and the line between the observer's eye and the base of the picture. Then letting x denote the distance between the wall and the observer's eye, $7 = x\tan(\theta+\alpha) = x\left(\frac{\tan\theta+\tan\alpha}{1-\tan\theta\tan\alpha}\right) = x\left(\frac{\tan\theta+\frac{2}{x}}{1-\tan\theta\left(\frac{2}{x}\right)}\right)$. The problem is equivalent to maximizing $\tan\theta$ so denote this by z and solve for it. Thus $x\left(\frac{z+\frac{2}{x}}{1-z\left(\frac{2}{x}\right)}\right) = 7$ and so $z = 5\frac{x}{14+x^2}$. It follows $\frac{dz}{dx} = 5\frac{14-x^2}{(14+x^2)^2}$ and setting this equal to zero, $x = \sqrt{14}$.

9. Find the point on the curve, $y = \sqrt{81-6x}$ which is closest to $(0,0)$.

   Answer:

   $\left(3,\sqrt{63}\right)$

10. A street is 200 feet long and there are two lights located at the ends of the street. One of the lights is $\frac{27}{8}$ times as bright as the other. Assuming the brightness of light from one of these street lights is proportional to the brightness of the light and the reciprocal of the square of the distance from the light, locate the darkest point on the street.

    Answer:

    80 feet from one light and 120 feet from the other.

11. Two cities are located on the same side of a straight river. One city is at a distance of 3 miles from the river and the other city is at a distance of 8 miles from the river. The distance between the two points on the river which are closest to the respective cities is 40 miles. Find the location of a pumping station which is to pump water to the two cities which will minimize the length of pipe used.

    Answer:

    $\frac{120}{11}$ miles from the point on the river closest to the city which is at a distance of 3 miles from the river.

6.7 Exercises

1. If $f'(x) = 0$, is it necessary that x is either a local minimum or local maximum? **Hint:** Consider $f(x) = x^3$.

2. Two positive numbers add to 32. Find the numbers if their product is to be as large as possible.

3. The product of two positive numbers equals 16. Find the numbers if their sum is to be as small as possible.

4. The product of two positive numbers equals 16. Find the numbers if twice the first plus three times the second is to be as small as possible.

5. Theodore, the owner of Theodore's tarantulas finds he can sell 6 tarantulas at the regular price of $20 each. At his last spider celebration day sale he reduced the price to $14 and was able to sell 14 tarantulas. He has to pay $.0 5 per day to maintain a tarantula and his fixed costs are $30 per day. What price should he charge to maximize his profit.

6. Lisa, the owner of Lisa's gags and gadgets sells 500 whoopee cushions per year. It costs $. 25 per year to store a whoopee cushion. To order whoopee cushions it costs $2 plus $. 25 per cushion. How many times a year and in what lot size should whoopee cushions be ordered to minimize inventory costs?

7. A continuous function, f defined on $[a,b]$ is to be maximized. It was shown above in Theorem 6.5.2 that if the maximum value of f occurs at $x \in (a,b)$, and if f is differentiable there, then $f'(x) = 0$. However, this theorem does not say anything about the case where the maximum of f occurs at either a or b. Describe how to find the point of $[a,b]$ where f achieves its maximum. Does f have a maximum? Explain.

8. Find the maximum and minimum values and the values of x where these are achieved for the function, $f(x) = x + \sqrt{25 - x^2}$.

9. A piece of wire of length L is to be cut in two pieces. One piece is bent into the shape of an equilateral triangle and the other piece is bent to form a square. How should the wire be cut to maximize the sum of the areas of the two shapes? How should the wire be bent to minimize the sum of the areas of the two shapes? **Hint:** Be sure to consider the case where all the wire is devoted to one of the shapes separately. This is a possible solution even though the derivative is not zero there.

10. A cylindrical can is to be constructed of material which costs 3 cents per square inch for the top and bottom and only 2 cents per square inch for the sides. The can needs to hold 90π cubic inches. Find the dimensions of the cheapest can. **Hint:** The volume of a cylinder is $\pi r^2 h$ where r is the radius of the base and h is the height. The area of the cylinder is $2\pi r^2 + 2\pi rh$.

11. A rectangular sheet of tin has dimensions 10 cm. by 20 cm. It is desired to make a topless box by cutting out squares from each corner of the rectangular sheet and then folding the rectangular tabs which remain. Find the volume of the largest box which can be made in this way.

12. Let $f(x) = \frac{1}{3}x^3 - x^2 - 8x$ on the interval $[-1, 10]$. Find the point of $[-1, 10]$ at which f achieves its minimum.

13. A rectangular garden 200 square feet in area is to be fenced off against rabbits. Find the least possible length of fencing if one side of the garden is already protected by a barn.

14. A feed lot is to be enclosed by a fence and divided in 5 pieces by 4 fences parallel to one side. 1272 feet of fencing is used. Find dimensions of the feed lot which will have the largest total area.

15. Find the dimensions of the largest rectangle that can be inscribed in a semicircle of radius r where $r = 8$.

16. Find the dimensions of the largest rectangle that can be inscribed in the ellipse, $\frac{x^2}{9} + \frac{y^2}{4} = 1$.

17. A smuggler wants to fit a small cylindrical vial inside a hollow rubber ball with a eight inch diameter. Find the volume of the largest vial that can fit inside the ball. The volume of a cylinder equals $\pi r^2 h$ where h is the height and r is the radius.

18. A function, f, is said to be odd if $f(-x) = f(x)$ and a function is said to be even if $f(-x) = f(x)$. Show that if f' is even, then f is odd and if f' is odd, then f is even. Sketch the graph of a typical odd function and a typical even function.

19. Recall sin is an odd function and cos is an even function. Determine whether each of the trig functions is odd, even or neither.

20. Find the x values of the critical points of the function $f(x) = 3x^2 - 5x^3$.

21. Find the x values of the critical points of the function $f(x) = \sqrt{3x^2 - 6x + 8}$.

22. Find the extreme points of the function, $f(x) = x + \frac{25}{x}$ and tell whether the extreme point is a local maximum or a local minimum or neither.

23. A piece of property is to be fenced on the front and two sides. Fencing for the sides costs \$3.50 per foot and fencing for the front costs \$5.60 per foot. What are the dimensions of the largest such rectangular lot if the available money is \$1400?

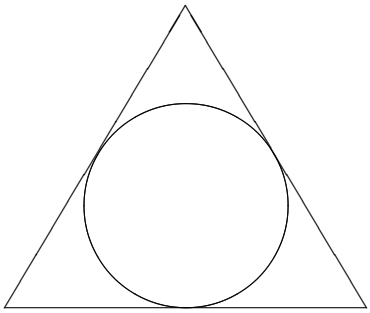
24. In a particular apartment complex of 200 units, it is found that all units remain occupied when the rent is \$400 per month. For each \$40 increase in the rent, one unit becomes vacant, on the average. Occupied units require \$80 per month for maintenance, while vacant units require none. Fixed costs for the buildings are \$20 000 per month. What rent should be charged for maximum profit and what is the maximum profit?

25. A picture is 9 feet high and the eye level of an observer is 2 feet below the bottom edge of the picture. How far from the picture should the observer stand if he wants to maximize the angle subtended by the picture?

26. Find the point on the curve, $y = \sqrt{25 - 2x}$ which is closest to $(0, 0)$.

27. A street is 200 feet long and there are two lights located at the ends of the street. One of the lights is $\frac{1}{8}$ times as bright as the other. Assuming the brightness of light from one of these street lights is proportional to the brightness of the light and the reciprocal of the square of the distance from the light, locate the darkest point on the street.

28. Find the volume of the smallest right circular cone which can be circumscribed about a sphere of radius 4 inches.



29. Two cities are located on the same side of a straight river. One city is at a distance of 3 miles from the river and the other city is at a distance of 8 miles from the river. The distance between the two points on the river which are closest to the respective cities is 40 miles. Find the location of a pumping station which is to pump water to the two cities which will minimize the length of pipe used.

30. A park ranger needs to get to a fire observation tower which is one mile from a long straight road in a dense forest.The point on the road closest to the observation tower is 10 miles down the road on which the park ranger is standing. Knowing that he can walk at 4 miles per hour on the road but only one mile per hour in the forest, how far down the road should he walk before entering the forest, in order to minimize the travel time?

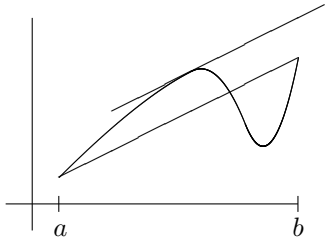
31. A hungry spider is located on the wall four feet off the floor directly above a point four feet from the corner of the room. This is a Daring Jumping Spider[1], not a lazy web spinner who just sits in the web and waits for its prey. This kind of spider stalks its dinner like a lion hunting an impala. At a point on the floor which is 6 feet from the wall and 8 feet from the corner is a possible dinner, a plump juicy fly temporarily distracted as it slurps on a suculent morcel of rotting meat. What path should the hungry spider follow? Describe a way to do this problem geometrically without using any calculus.

6.8 Mean Value Theorem

The mean value theorem is one of the most important theorems about the derivative. The best versions of many other theorems depend on this fundamental result. The mean value theorem says that under suitable conditions, there exists a point in (a,b), x, such that $f'(x)$ equals the slope of the secant line,

$$\frac{f(b)-f(a)}{b-a}.$$

The following picture is descriptive of this situation.



This theorem is an existence theorem and like the other existence theorems in analysis, it depends on the completeness axiom. The following is known as Rolle's[2] theorem.

Theorem 6.8.1 *Suppose $f:[a,b]\to\mathbb{R}$ is continuous,*

$$f(a)=f(b),$$

[1] These are very beautiful spiders if you don't look too close. They are small furry and black with white markings. You sometimes see them running along walls and ceilings. Like many spiders they can climb even very slippery surfaces like glass with no difficulties.

[2] Rolle is remembered for Rolle's theorem and not for anything else he did. Ironically, he did not like calculus.

and

$$f : (a,b) \to \mathbb{R}$$

has a derivative at every point of (a,b). *Then there exists* $x \in (a,b)$ *such that* $f'(x) = 0$.

Proof: Suppose first that $f(x) = f(a)$ for all $x \in [a,b]$. Then any $x \in (a,b)$ is a point such that $f'(x) = 0$. If f is not constant, either there exists $y \in (a,b)$ such that $f(y) > f(a)$ or there exists $y \in (a,b)$ such that $f(y) < f(b)$. In the first case, the maximum of f is achieved at some $x \in (a,b)$ and in the second case, the minimum of f is achieved at some $x \in (a,b)$. Either way, Theorem 6.5.2 on Page 140 implies $f'(x) = 0$. This proves Rolle's theorem.

The next theorem is known as the Cauchy mean value theorem.

Theorem 6.8.2 *Suppose* f, g *are continuous on* $[a,b]$ *and differentiable on* (a,b). *Then there exists* $x \in (a,b)$ *such that*

$$f'(x)(g(b) - g(a)) = g'(x)(f(b) - f(a)).$$

Proof: Let

$$h(x) \equiv f(x)(g(b) - g(a)) - g(x)(f(b) - f(a)).$$

Then letting $x = a$ and then letting $x = b$, a short computation shows $h(a) = h(b)$. Also, h is continuous on $[a,b]$ and differentiable on (a,b). Therefore Rolle's theorem applies and there exists $x \in (a,b)$ such that

$$h'(x) = f'(x)(g(b) - g(a)) - g'(x)(f(b) - f(a)) = 0.$$

This proves the theorem.

The usual mean value theorem, sometimes called the Lagrange mean value theorem, illustrated by the above picture is obtained by letting $g(x) = x$.

Corollary 6.8.3 *Let* f *be continuous on* $[a,b]$ *and differentiable on* (a,b). *Then there exists* $x \in (a,b)$ *such that* $f(b) - f(a) = f'(x)(b - a)$.

Corollary 6.8.4 *Suppose* $f'(x) = 0$ *for all* $x \in (a,b)$ *where* $a \geq -\infty$ *and* $b \leq \infty$. *Then* $f(x) = f(y)$ *for all* $x, y \in (a,b)$. *Thus* f *is a constant.*

Proof: If this is not true, there exists x_1 and x_2 such that $f(x_1) \neq f(x_2)$. Then by the mean value theorem,

$$0 \neq \frac{f(x_1) - f(x_2)}{x_1 - x_2} = f'(z)$$

for some z between x_1 and x_2. This contradicts the hypothesis that $f'(x) = 0$ for all x. This proves the theorem.

Corollary 6.8.5 *Suppose* $f'(x) > 0$ *for all* $x \in (a,b)$ *where* $a \geq -\infty$ *and* $b \leq \infty$. *Then* f *is strictly increasing on* (a,b). *That is, if* $x < y$, *then* $f(x) < f(y)$. *If* $f'(x) \geq 0$, *then* f *is increasing in the sense that whenever* $x < y$ *it follows that* $f(x) \leq f(y)$.

Proof: Let $x < y$. Then by the mean value theorem, there exists $z \in (x,y)$ such that

$$0 < f'(z) = \frac{f(y) - f(x)}{y - x}.$$

Since $y > x$, it follows $f(y) > f(x)$ as claimed. Replacing $<$ by $\leq$ in the above equation and repeating the argument gives the second claim.

Corollary 6.8.6 *Suppose $f'(x) < 0$ for all $x \in (a,b)$ where $a \geq -\infty$ and $b \leq \infty$. Then f is strictly decreasing on (a,b). That is, if $x < y$, then $f(x) > f(y)$. If $f'(x) \leq 0$, then f is decreasing in the sense that for $x < y$, it follows that $f(x) \geq f(y)$*

Proof: Let $x < y$. Then by the mean value theorem, there exists $z \in (x,y)$ such that

$$0 > f'(z) = \frac{f(y) - f(x)}{y - x}.$$

Since $y > x$, it follows $f(y) < f(x)$ as claimed. The second claim is similar except instead of a strict inequality in the above formula, you put $\geq$.

6.9 Exercises

1. Sally drives her Saturn over the 110 mile toll road in exactly 1.3 hours. The speed limit on this toll road is 70 miles per hour and the fine for speeding is 10 dollars per mile per hour over the speed limit. How much should Sally pay?

2. Two cars are careening down a freeway weaving in and out of traffic. Car A passes car B and then car B passes car A as the driver makes obscene gestures. This infuriates the driver of car A who passes car B while firing his handgun at the driver of car B. Show there are at least two times when both cars have the same speed. Then show there exists at least one time when they have the same acceleration. The acceleration is the derivative of the velocity.

3. Show the cubic function, $f(x) = 5x^3 + 7x - 18$ has only one real zero.

4. Suppose $f(x) = x^7 + |x| + x - 12$. How many solutions are there to the equation, $f(x) = 0$?

5. Let $f(x) = |x-7| + (x-7)^2 - 2$ on the interval $[6,8]$. Then $f(6) = 0 = f(8)$. Does it follow from Rolle's theorem that there exists $c \in (6,8)$ such that $f'(c) = 0$? Explain your answer.

6. Suppose f and g are differentiable functions defined on $\mathbb{R}$. Suppose also that it is known that $|f'(x)| > |g'(x)|$ for all x and that $|f'(t)| > 0$ for all t. Show that whenever $x \neq y$, it follows $|f(x) - f(y)| > |g(x) - g(y)|$. **Hint:** Use the Cauchy mean value theorem, Theorem 6.8.2.

7. Show that, like continuous functions, functions which are derivatives have the intermediate value property. This means that if $f'(a) < 0 < f'(b)$ then there exists $x \in (a,b)$ such that $f'(x) = 0$. **Hint:** Argue the minimum value of f occurs at an interior point of $[a,b]$.

8. Consider the function
$$f(x) \equiv \begin{cases} 1 \text{ if } x \geq 0 \\ -1 \text{ if } x < 0 \end{cases}.$$
Is it possible that this function could be the derivative of some function? Why?

9. Suppose $a \in I$, an open interval and that a function f, defined on I has $n+1$ derivatives. Then for each $m \leq n$ the following formula holds for $x \in I$.
$$f(x) = \sum_{k=0}^{m} f^{(k)}(a) \frac{(x-a)^k}{k!} + f^{(m+1)}(y) \frac{(x-a)^{m+1}}{(m+1)!} \tag{6.13}$$

where y is some point between x and a. Note that if $n = 0$, this reduces to the Lagrange form of the mean value theorem so the formula holds if $m = 0$. Suppose it holds for some $0 \leq m-1 < n$. The task is to show then that it also holds for m. It will then follow that the above formula will hold for all $m \leq n$ as claimed. This formula is very important. The last term is called the Lagrange form of the remainder in Taylor series. It will be discussed carefully later. For now, try to prove the formula using the following steps. If the formula holds for $m-1$, then you can apply it to f'.

$$f'(x) - \sum_{k=0}^{m-1} \frac{f^{(k+1)}(a)}{k!}(x-a)^k = \frac{f^{(m+1)}(y)}{m!}(x-a)^m. \tag{6.14}$$

Now let $g(x) = f(x) - \sum_{k=0}^{m} \frac{f^{(k)}(a)}{k!}(x-a)^k$ and let $h(x) = (x-a)^{m+1}$. Now from the Cauchy mean value theorem there exists z between x and a such that

$$\frac{g(x)-g(a)}{h(x)-h(a)} = \frac{g'(z)}{h'(z)} = \frac{f'(z) - \sum_{k=1}^{m} \frac{f^{(k)}(a)}{(k-1)!}(z-a)^{k-1}}{(m+1)(z-a)^m}.$$

Now explain why $\sum_{k=1}^{m} \frac{f^{(k)}(a)}{(k-1)!}(z-a)^{k-1} = \sum_{k=0}^{m-1} \frac{f^{(k+1)}(a)}{k!}(z-a)^k$ and then use (6.14). Explain why this implies

$$\frac{f(x) - \sum_{k=0}^{m} \frac{f^{(k)}(a)}{k!}(x-a)^k}{(x-a)^{m+1}} = \frac{\frac{f^{(m+1)}(y)}{m!}(z-a)^m}{(m+1)(z-a)^m}$$

which yields the desired formula for m. A way to think of (6.13) is as a generalized mean value theorem. Note that the key result which made it work was the Cauchy mean value theorem.

6.10 Curve Sketching

The theorems and corollaries given above can be used to aid in sketching the graphs of functions. The second derivative will also help in determining the shape of the function.

Definition 6.10.1 *A differentiable function, f, defined on an interval, (a,b), is concave up if f' is an increasing function. A differentiable function defined on an interval, (a,b), is concave down if f' is a decreasing function. A point where the graph of the function changes from being concave up to concave down or from concave down to concave up is called an inflection point.*

From the geometric description of the derivative as the slope of a tangent line to the graph of the function, to say the derivative is an increasing function means that as you move from left to right, the slopes of the lines tangent to the graph of f become larger. Thus the graph of the function is bent up in the shape of a smile. It may also help to think of it as a cave when you view it from above, hence the term concave up. If the derivative is decreasing, it follows that as you move from left to right the slopes of the lines tangent to the graph of f become smaller. Thus the graph of the function is bent down in the form of a frown. It is concave down because it is like a cave when viewed from beneath. The following theorem will give a convenient criterion in terms of the second derivative for finding whether a function is concave up or concave down. The term, concavity, is used to refer to this property. Thus you determine the concavity of a function when you find whether it is concave up or concave down.

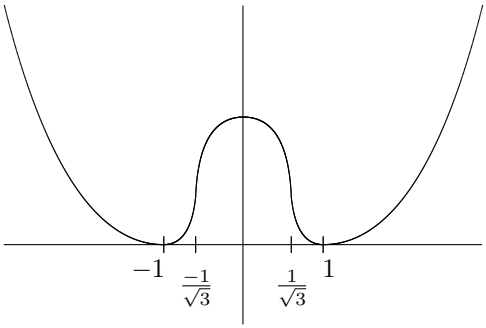
Theorem 6.10.2 *Suppose $f''(x) > 0$ for $x \in (a,b)$. Then f is concave up on (a,b). Suppose $f''(x) < 0$ on (a,b). Then f is concave down.*

Proof: This follows immediately from Corollaries 6.8.6 and 6.8.5 applied to the first derivative. The following picture may help in remembering this.

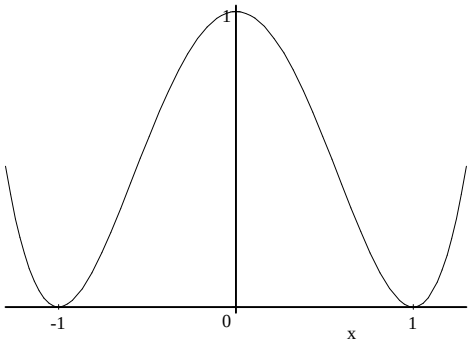
In this picture, the plus signs and the smile on the left correspond to the second derivative being positive. The smile gives the way in which the graph of the function is bent. In the second face, the minus signs correspond to the second derivative being negative. The frown gives the way in which the graph of the function is bent.

Example 6.10.3 *Sketch the graph of the function, $f(x) = \left(x^2-1\right)^2 = x^4 - 2x^2 + 1$*

Take the derivative of this function, $f'(x) = 4x^3 - 4x = 4x(x-1)(x+1)$ which equals zero at $-1, 0$, and 1. It is positive on $(-1,0)$, and $(1,\infty)$ and negative on $(0,1)$ and $(-\infty,-1)$. Therefore, $x = 0$ corresponds to a local maximum and $x = -1$ and $x = 1$ correspond to local minimums. The second derivative is $f''(x) = 12x^2 - 4$ and this equals zero only at the points $-1/\sqrt{3}$ and $1/\sqrt{3}$. The second derivative is positive on the intervals $\left(1/\sqrt{3},\infty\right)$ and $\left(-\infty,-1/\sqrt{3}\right)$ so the function, f is smiling on these intervals. The second derivative is negative on the interval $\left(-1/\sqrt{3},1/\sqrt{3}\right)$ and so the original function is frowning on this interval. This describes in words the qualitative shape of the function. It only remains to draw a picture which incorporates this description. The following is such a sketch. It is not intended to be an accurate drawing made to scale, only to be a qualitative picture of what was just determined.



A better graph of this function is the following, done by a computer algebra system. However the computer worked a lot harder.



In general, if you are interested in getting a nice graph of a function, you should use a computer algebra system. An effective way to accomplish your graphing is to go to the help menu and copy and paste an example from this menu changing it as needed. Both mathematica and Maple have good help menus. Keep in mind there are certain conventions which must be followed. For example to write x raised to the second power you enter x ^2. In Maple, you also need to place an asterisk between quantities which are multiplied since otherwise it will not know you are multiplying and won't work. There are also easy to use versions of Maple available which involve essentially pointing and clicking. You won't learn any calculus from playing with a computer algebra system but you might have a lot of fun.

6.11 Exercises

1. Sketch the graph of the function, $f(x) = x^3 - 3x + 1$ showing the intervals on which the function is concave up and down and identifying the intervals on which the function is increasing.

2. Find intervals on which the function, $f(x) = \sqrt{1-x^2}$ is increasing and intervals on which it is concave up and concave down. Sketch a graph of the function.

3. Sketch the graphs of $y = x^4, y = x^3$, and $y = -x^4$. What do these graphs tell you about the case when the second derivative equals zero?

4. Sketch the graph of $f(x) = 1/\left(1+x^2\right)$ showing the intervals on which the function is increasing or decreasing and the intervals on which the graph is concave up and concave down.

5. Sketch the graph of $f(x) = x/\left(1+x^2\right)$ showing the intervals on which the function is increasing or decreasing and the intervals on which the graph is concave up and concave down.

6. Show that inflection points can be identified by looking at those points where the second derivative equals zero but that not every point where the second derivative equals zero is an inflection point. **Hint:** For the last part consider $y = x^3$ and $y = x^4$.

7. Suppose $f''(x) = 0$ and $f'''(x) \neq 0$. Does it follow that x must be an inflection point?

8. Find all inflection points for the function, $f(x) = x^2/\left(1+x^2\right)$.

Some Important Special Functions

7.0.1 Outcomes

1. Understand the definition and derivatives of the circular functions.

2. Use the rules of exponents and find derivatives of the exponential functions and the logarithm functions.

3. Understand the number, e.

7.1 The Circular Functions

The Trigonometric functions are also called the circular functions. Thus this section will be on the functions, $\cos, \sin, \tan, \sec, \csc$, and $\cot$. The first thing to do is to give an important lemma. There are several approaches to this lemma. To see it done in terms of areas of a circular sector, see Apostol, [1], or almost any other calculus book. However, the book by Apostol has no loose ends in the presentation unlike most other books which use this approach. The proof given here is a modification of that found in Tierney, [27] and Rose, [22] and is based on arc length.

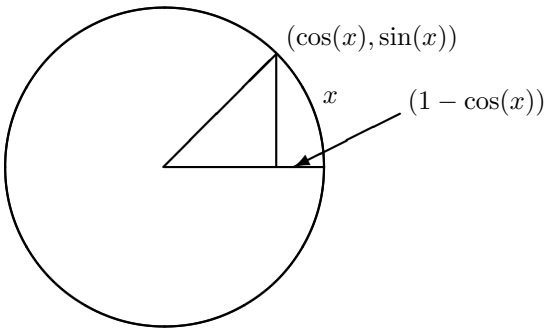
Lemma 7.1.1 *The following limits hold.*

$$\lim_{x\to 0} \frac{\sin x}{x} = 1 \tag{7.1}$$

$$\lim_{x\to 0} \frac{1-\cos x}{x} = 0 \tag{7.2}$$

Proof: First consider (7.1). In the following picture, it follows from Corollary 3.5.4 on Page 62 that for small positive x,

$$\sin x + (1-\cos x) \geq x \geq \sin x. \tag{7.3}$$



Now divide by $\sin x$ to get

$$1+\frac{1-\cos x}{|\sin x|}=1+\frac{1-\cos x}{\sin x}\geq\frac{x}{\sin x}\geq 1.$$

For small negative values of x, it is also true that

$$1+\frac{1-\cos x}{|\sin x|}\geq\frac{x}{\sin x}\geq 1.$$

(Why?) From the trig. identities, it follows that for all small values of x,

$$1+\frac{\sin^2 x}{|\sin x|\,(1+\cos x)}=1+\frac{|\sin x|}{(1+\cos x)}\geq\frac{x}{\sin x}\geq 1$$

and so from the squeezing theorem, Theorem 5.9.5 on Page 111,

$$\lim_{x\to 0}\frac{x}{\sin x}=1$$

and consequently, from the limit theorems,

$$\lim_{x\to 0}\frac{\sin x}{x}=\lim_{x\to 0}\frac{1}{\left(\frac{x}{\sin x}\right)}=1.$$

Finally,

$$\frac{1-\cos x}{x}=\frac{1-\cos^2 x}{x\,(1+\cos x)}=\sin x\frac{\sin x}{x}\frac{1}{1+\cos x}.$$

Therefore, from Theorem 5.5.1 on Page 102 which says $\lim_{x\to 0}\sin(x)=0$, and the limit theorems,

$$\lim_{x\to 0}\frac{1-\cos x}{x}=0.$$

This proves the Lemma.

With this, it is easy to find the derivative of sin. Using Lemma 7.1.1,

$$\begin{aligned}\lim_{h\to 0}\frac{\sin(x+h)-\sin x}{h} &= \lim_{h\to 0}\frac{\sin(x)\cos(h)+\cos(x)\sin(h)-\sin x}{h}\\ &= \lim_{h\to 0}\frac{(\sin x)(\cos(h)-1)}{h}+\cos x\frac{\sin(h)}{h}\\ &= \cos x.\end{aligned}$$

The derivative of cos can be found the same way. Alternatively,

$$\cos(x) = \sin(x + \pi/2)$$

and so

$$\begin{aligned} \cos'(x) &= \sin'(x + \pi/2) \\ &= \cos(x + \pi/2) \\ &= \cos x \cos(\pi/2) - \sin x \sin(\pi/2) \\ &= -\sin x. \end{aligned}$$

The following theorem is now obvious and the proofs of the remaining parts are left for you.

Theorem 7.1.2 *The derivatives of the trig. functions are as follows.*

$$\begin{aligned} \sin'(x) &= \cos x \\ \cos'(x) &= -\sin x \\ \tan'(x) &= \sec^2(x) \\ \cot'(x) &= -\csc^2(x) \\ \sec'(x) &= \sec x \tan x \\ \csc'(x) &= -\csc x \cot x \end{aligned}$$

Here are some examples of extremum problems which involve the use of the trig. functions.

Example 7.1.3 *Two hallways intersect at a right angle. One is* 5 *feet wide and the other is* 2 *feet wide. What is the length of the longest thin rod which can be carried horizontally from one hallway to the other?*

You must minimize the length of the rod which touches the inside corner of the two halls and extends to the outside walls. Letting θ be the angle between this rod and the outside wall for the hall having width 2, minimize

$$f(\theta) = \overbrace{2\csc\theta + 5\sec\theta}^{\text{length of rod}}.$$

Therefore, using the rules of differentiation,

$$f'(\theta) = \frac{2\cos^3\theta - 5\sin\theta + 5\sin\theta\cos^2\theta}{(\cos^2\theta)(-1+\cos^2\theta)} = 0$$

should be solved to get the angle where this length is as small as possible. Thus

$$2\cos^3\theta - 5\sin\theta + 5\sin\theta\cos^2\theta = 0.$$

and $2\cos^3\theta - 5\sin^3\theta = 0$ and so $\tan\theta = \frac{1}{5}\sqrt[3]{2}\left(\sqrt[3]{5}\right)^2$. Drawing a triangle, you see that at this value of θ, you have $\sec\theta = \frac{\sqrt{\left(\sqrt[3]{5}\right)^2+\left(\sqrt[3]{2}\right)^2}}{\sqrt[3]{5}}$ and $\csc\theta = \frac{\sqrt{\left(\sqrt[3]{5}\right)^2+\left(\sqrt[3]{2}\right)^2}}{\sqrt[3]{2}}$. Therefore, the minimum is obtained by substituting these values in to the equation for $f(\theta)$ yielding $\left(\sqrt{\left(\left(\sqrt[3]{5}\right)^2 + \left(\sqrt[3]{2}\right)^2\right)}\right)^3$.

Example 7.1.4 *A fence* 9 *feet high is* 2 *feet from a building. What is the length of the shortest ladder which will lean against the top of the fence and touch the building?*

Let θ be the angle of the ladder with the ground. Then the length of this ladder making this angle with the ground and leaning on the top of the fence while touching the building is

$$f(\theta) = \frac{9}{\sin\theta} + \frac{2}{\cos\theta}.$$

Then the final answer is $\left(\sqrt{\left(3\sqrt[3]{3} + \left(\sqrt[3]{2}\right)^2\right)}\right)^3$. The details are similar to the problem of the two hallways.

7.2 Exercises

1. Prove all parts of Theorem 7.1.2.
2. Prove $\tan'(x) = 1 + \tan^2(x)$.
3. Find and prove a formula for the derivative of $\sin^m(x)$ for m an integer.
4. Find the derivative of the function, $\sin^6(5x)$.
5. Find the derivative of the function, $\tan^7(4x)$.
6. Find the derivative of the function, $\frac{\sec^3(2x)}{\tan^3(3x)}$.
7. Find all intervals where $\sin(2x)$ is concave down.
8. Find the intervals where $\cos(3x)$ is increasing.
9. Two hallways intersect at a right angle. One is 3 feet wide and the other is 4 feet wide. What is the length of the longest thin rod which can be carried horizontally from one hallway to the other?
10. A fence 5 feet high is 2 feet from a building. What is the length of the shortest ladder which will lean against the top of the fence and touch the building?
11. Suppose $f(x) = A\cos\omega x + B\sin\omega x$. Show there exists an angle, ϕ such that $f(x) = \sqrt{A^2+B^2}\sin(\omega x + \phi)$. The number, $\sqrt{A^2+B^2}$ gives the "amplitude" and ϕ is called the "phase shift" while ω is called the "frequency". This is very important because it allows us to understand what is going on. The amplitude gives the height of the periodic function, f. **Hint:** Remember a point on the unit circle determines an angle. Write $f(x)$ in the form

$$\sqrt{A^2+B^2}\left(\frac{A}{\sqrt{A^2+B^2}}\cos\omega x + \frac{B}{\sqrt{A^2+B^2}}\sin\omega x\right)$$

and note that $\left(\frac{B}{\sqrt{A^2+B^2}}, \frac{A}{\sqrt{A^2+B^2}}\right)$ is a point on the unit circle.

12. Repeat Problem 11 but this time show $f(x) = \sqrt{A^2+B^2}\cos(\omega x + \phi)$. How could you find ϕ?
13. A kite is moving horizontally at the rate of 10 feet per second and is 100 feet high. How fast is the angle of elevation of the kite string changing when 200 feet of string have been let out?

14. A square picture is 6 feet high and is fastened to the wall with its lowest edge one foot above the eye level of an observer. Where should he stand to maximize the angle subtended by the picture at his eye?

15. A 100 foot tall lamp post with a light on top creates a shadow from a falling ball dropped from a height of 100 feet at a distance of 50 feet from the lamp post. Thus the distance the ball has fallen at time t is $16t^2$. Find the velocity of the shadow when the ball has dropped a distance of 64 feet. Also find the rate of change of the angle of elevation from the shadow to the light.

7.3 The Exponential And Log Functions

7.3.1 The Rules Of Exponents

As mentioned earlier, b^m means to multiply b by itself m times assuming m is a positive integer. $b^0 \equiv 1$ provided $b \neq 0$. In the case where $b = 0$ the symbol is undefined. If $m < 0$, b^m is defined as $\frac{1}{b^{-m}}$. Then the following algebraic properties are obtained. Be sure you understand these properties for x and y integers.

$$b^{x+y} = b^x b^y,\ (ab)^x = a^x b^x \tag{7.4}$$

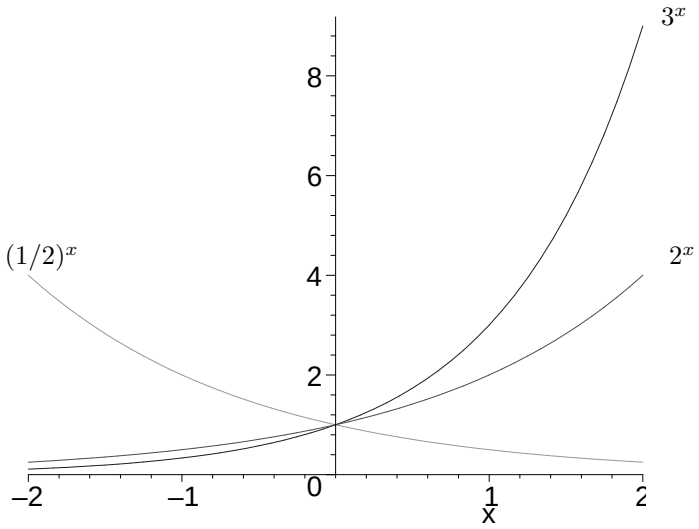
$$b^{xy} = (b^x)^y\ ,\ b^{-1} = \frac{1}{b} \tag{7.5}$$

These properties are called the rules of exponents.

When x and y are not integers, the meaning of b^x is no longer clear. For example, suppose $b = -1$ and $x = 1/2$. What exactly is meant by $(-1)^{1/2}$? Even in the case where $b > 0$ there are difficulties. If x is a rational number, m/n and $b > 0$ the symbol $b^{m/n}$ means $\sqrt[n]{b^m}$. That is its definition and it is a useful exercise for you to verify (7.4) and (7.5) hold with this definition. There are no mathematical questions about the existence of this number. To see this, consider Problem 7 on Page 108. The problem is not one of theory but of practicality. Could you use this definition to find $2^{\frac{1234567812345}{1234567812344}}$? Consider what you would do. First find the number $2^{1234567812345}$ and then $\cdots$? Can you find this number? It is just too big. However, a calculator can find $2^{\frac{1234567812345}{1234567812344}}$. It yields $2^{\frac{1234567812345}{1234567812344}} = 2.000\,000\,000\,001\,123$ as an approximate answer. Clearly something else must be going on. To make matters even worse, what would you do with $2^{\sqrt{2}}$? As mentioned earlier, $\sqrt{2}$ is irrational and so cannot be written as the quotient of two integers. These are serious difficulties and must be dealt with.

7.3.2 The Exponential Functions, A Wild Assumption

Using your calculator or a computer you can obtain graphs of the functions, $y = b^x$ for various choices of b. The following picture gives a few of these graphs.



These graphs suggest that if $b < 1$ the function, $y = b^x$ is decreasing while if $b > 1$, the function is increasing but just how was the calculator or computer able to draw those graphs? Also, do the laws of exponents continue to hold for all real values of x? The short answer is that they do and this is shown later but for now here is a wild assumption which glosses over these issues.

Wild Assumption 7.3.1 *For every $b > 0$ there exists a unique differentiable function $\exp_b(x) \equiv b^x$ valid for all real values of x such that (7.4) and (7.5) both hold for all $x, y \in \mathbb{R}$, $\exp_b(m/n) = \sqrt[n]{b^m}$ whenever m, n are integers, and $b^x > 0$ for all $x \in \mathbb{R}$. Furthermore, if $b \neq 1$ and $h \neq 0$, then $\exp_b(h) = b^h \neq 1$.*

Instead of writing $\exp_b(x)$ I will often write b^x and I will also be somewhat sloppy and regard b^x as the name of a function and not just as $\exp_b(x)$, a given function defined at x. This is done to conform with usual usage. Also, the last claim in Wild Assumption 7.3.1 follows from the first part of this assumption. See Problem 1. I want it to be completely clear that the Wild Assumption is just that. No reason for believing in such an assumption

has been given notwithstanding the pretty pictures drawn by the calculator. Later in the book, the wild assumption will be completely justified. Based on Wild Assumption 7.3.1 one can easily find out all about b^x.

Theorem 7.3.2 *Let* $\exp_b$ *be defined in Wild Assumption 7.3.1 for* $b > 0$. *Then there exists a unique number, denoted by* $\ln b$ *for* $b > 0$ *satisfying*

$$\exp_b'(x) = \ln b \exp_b(x). \tag{7.6}$$

Furthermore,

$$\ln(ab) = \ln(a) + \ln(b),\ \ln 1 = 0, \tag{7.7}$$

and for all $y \in \mathbb{R}$,

$$\ln(b^y) = y \ln b. \tag{7.8}$$

$$\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b) \tag{7.9}$$

The function, $x \to \ln x$ *is differentiable and defined for all* $x > 0$ *and*

$$\ln'(x) = \frac{1}{x}. \tag{7.10}$$

The function, $x \to \ln x$ *is one to one on* $(0,\infty)$. *Also,* ln *maps* $(0,\infty)$ *onto* $(-\infty,\infty)$.

Proof: First consider (7.6).

$$\begin{aligned}\lim_{h\to 0}\frac{\exp_b(x+h)-\exp_b(x)}{h} &= \lim_{h\to 0}\frac{b^{x+h}-b^x}{h}\\ &= \lim_{h\to 0}\left(\frac{b^h-1}{h}\right)b^x.\end{aligned}$$

The expression, $\lim_{h\to 0}\left(\frac{b^h-1}{h}\right)$ is assumed to exist thanks to Wild Assumption 7.3.1 and this is denoted by $\ln b$. This proves (7.6).

To verify (7.7), if $b = 1$ then $b^x = 1^x$ for all $x \in \mathbb{R}$. Now by (7.4) and (7.5),

$$1^x 1^x = 1^{x+x} = \left(1^2\right)^x = 1^x$$

and so, dividing both sides by 1^x, an operation justified by Wild Assumption 7.3.1, $1^x = 1$ for all $x \in \mathbb{R}$. Therefore, $\exp_1(x) = 1$ for all x and so $\exp_1'(x) = \ln 1 \exp_1(x) = 0$. Thus $\ln 1 = 0$ as claimed. Next, by the product rule and Wild Assumption 7.3.1,

$$\begin{aligned}\ln(ab)(ab)^x &= \left((ab)^x\right)'\\ &= \left(a^x b^x\right)' = \left(a^x\right)' b^x + a^x\left(b^x\right)'\\ &= (\ln a)\, a^x b^x + (\ln b)\, b^x a^x\\ &= [\ln a + \ln b](ab)^x.\end{aligned}$$

Therefore, $\ln(ab) = \ln a + \ln b$ as claimed.

Next consider (7.8). Keeping y fixed, consider the function $x \to b^{xy} = \left(b^y\right)^x$. Then,

$$\ln\left(b^y\right)\left(b^y\right)^x = \left(\left(b^y\right)^x\right)' = g'(x)$$

where $g(x) \equiv b^{xy}$. Therefore, using (6.7) on Page 134,

$$g'(x) = y \exp_b'(xy)$$

and so

$$\ln\left(b^y\right)\left(b^y\right)^x = y\exp_b'(xy) = y\left(\ln b\right)\left(b^{xy}\right) = y\left(\ln b\right)\left(b^y\right)^x.$$

Now dividing both sides by $\left(b^y\right)^x$ verifies (7.8).

To obtain (7.9) from this, note

$$\ln\left(\frac{a}{b}\right) = \ln\left(ab^{-1}\right) = \ln(a) + \ln\left(b^{-1}\right) = \ln(a) - \ln(b).$$

It remains to verify (7.10). From (6.2) and the continuity of 2^x,

$$\ln'(1) = \lim_{h\to 0}\frac{\ln\left(2^h\right) - \ln 1}{2^h - 1} = \lim_{h\to 0}\frac{h\ln 2}{2^h - 1} = \frac{\ln 2}{\ln 2} = 1.$$

$\ln 2 \neq 0$ because if it were, then $\left(2^x\right)' = (\ln 2)\,2^x = 0$ and by Corollary 6.8.4, this would imply 2^x is a constant function which it is not. Now the first part of this lemma implies

$$\begin{aligned}\ln'(x) &= \lim_{y\to x}\frac{\ln y - \ln x}{y - x} = \lim_{y\to x}\frac{1}{x}\frac{\ln\left(\frac{y}{x}\right) - \ln 1}{\left(\frac{y}{x}\right) - 1}\\ &= \frac{1}{x}\ln'(1) = \frac{1}{x}.\end{aligned}$$

It remains to verify ln is one to one. Suppose $\ln x = \ln y$. Then by the mean value theorem, there exists t between x and y such that $(1/t)(x - y) = \ln x - \ln y = 0$. Therefore, $x = y$ and this shows ln is one to one as claimed.

It only remains to verify that ln maps $(0,\infty)$ onto $(-\infty,\infty)$. By Wild Assumption 7.3.1 and the mean value theorem, Corollary 6.8.3, there exists $y \in (0,1)$ such that

$$0 < \frac{2^1 - 1}{1} = \ln(2)\,2^y$$

Since $2^y > 0$ it follows $\ln(2) > 0$ and so $x \to 2^x$ is strictly increasing. Therefore, by Corollary 6.8.3

$$\frac{2^1 - 1}{1} = \ln(2)\,2^y \leq \ln(2)\,2^1$$

and it follows that

$$\frac{1}{2} \leq \ln 2. \tag{7.11}$$

Also, from (7.7)

$$0 = \ln(2) + \ln\left(\frac{1}{2}\right) \geq \frac{1}{2} + \ln\left(\frac{1}{2}\right)$$

which shows that

$$\ln\left(\frac{1}{2}\right) \leq -\frac{1}{2}. \tag{7.12}$$

It follows from (7.11) and (7.12) that ln achieves values which are arbitrarily large and arbitrarily large in the negative direction. Therefore, by the intermediate value theorem, ln achieves all values.

More precisely, let $y \in \mathbb{R}$. Then choose n large enough that $\frac{n}{2} > y$ and $-\frac{n}{2} < y$. Then from (7.11) and (7.12)

$$\ln\left(\left(\frac{1}{2}\right)^n\right) \leq \frac{-n}{2} < y < \frac{n}{2} < \ln\left(2^n\right).$$

By the intermediate value theorem, there exists $x \in \left(\left(\frac{1}{2}\right)^n, 2^n\right)$ such that $\ln x = y$. This proves the theorem.

Example 7.3.3 *Find* $f'(x)$ *if* $f(x) = 7^{2x}$.
$7^{2x} = (7^x)^2$ *and so the derivative is* $2(7^x)(7^x)' = 2(7^x)\ln 7\,(7^x) = 2(\ln 7)\,7^{2x}$.

Example 7.3.4 *Find* $f'(x)$ *if* $f(x) = 7^{(x^2)}$.

Set up the difference quotient,

$$\begin{aligned}\frac{7^{(x+h)^2} - 7^{x^2}}{h} &= 7^{x^2}\left(\frac{7^{2xh+h^2}-1}{h}\right)\\ &= 7^{x^2}\left(\frac{7^{2xh+h^2}-1}{2xh+h^2}\right)\frac{2xh+h^2}{h}.\end{aligned}$$

Taking the limit as $h \to 0$ and using the definition of $\ln 7$, this limit equals $7^{x^2}\ln(7)\,2x$.

7.3.3 The Special Number, e

Since ln is one to one onto $\mathbb{R}$, it follows there exists a unique number, e such that $\ln(e) = 1$. Therefore,

$$\exp_e'(x) \equiv (e^x)' = \ln(e)\,e^x \equiv \ln(e)\exp_e(x) = \exp_e(x)$$

showing that $\exp_e$ has the remarkable property that it equals its own derivative. This wonderful number is called Euler's number and it can be shown to equal approximately 2.718 3. It is customary to write $\exp(x)$ for $\exp_e(x)$. Thus

$$\exp'(x) = \exp(x). \tag{7.13}$$

7.3.4 The Function $\ln|x|$

The function, ln is only defined on positive numbers. However, it is possible to write $\ln|x|$ whenever $x \neq 0$. What is the derivative of this function?

Corollary 7.3.5 *Let* $f(x) = \ln|x|$ *for* $x \neq 0$. *Then*

$$f'(x) = \frac{1}{x}.$$

Proof: If $x > 0$ the formula is just (7.10). Suppose then that $x < 0$. Then $\ln|x| = \ln(-x)$ so by (6.7) on Page 134,

$$\begin{aligned}(\ln|x|)' &= (\ln(-x))' = (\ln((-1)\,x))'\\ &= \frac{1}{-x}(-1) = \frac{1}{x}.\end{aligned}$$

This proves the corollary.

7.3.5 Logarithm Functions

Next a new function called $\log_b$ will be defined.

Definition 7.3.6 *For all* $b > 0$ *and* $b \neq 1$

$$\log_b(x) \equiv \frac{\ln x}{\ln b}. \tag{7.14}$$

Notice this definition implies (7.7) - (7.9) all hold with ln replaced with $\log_b$. Also, $\log_b : (0,\infty) \to \mathbb{R}$ is one to one and onto.

The fundamental relationship between the exponential function, b^x and $\log_b x$ is in the following proposition. This proposition shows this new function is $\log_b$ you may have studied in high school.

Proposition 7.3.7 *Let $b > 0$ and $b \neq 1$. Then for all $x > 0$,*

$$b^{\log_b x} = x, \tag{7.15}$$

and for all $y \in \mathbb{R}$,

$$\log_b b^y = y, \tag{7.16}$$

Also,

$$\log_b'(x) = \frac{1}{\ln b}\frac{1}{x}. \tag{7.17}$$

Proof: Formula (7.15) follows from (7.8).

$$\ln\left(b^{\log_b x}\right) = \log_b x \ln b = \ln x$$

and so, since ln is one to one, it follows (7.15) holds.

$$\log_b b^y \equiv \frac{\ln(b^y)}{\ln b} = \frac{y \ln b}{\ln b} = y$$

and this verifies (7.16). Formula (7.17) is obvious from (7.14).

The functions, $\log_b$ are only defined on positive numbers. However, it is possible to write $\log_b |x|$ whenever $x \neq 0$. What is the derivative of these functions?

Corollary 7.3.8 *Let $f(x) = \log_b |x|$ for $x \neq 0$. Then*

$$f'(x) = \frac{1}{(\ln b)\, x}.$$

Proof: If $x > 0$ the formula is just (7.17). Suppose then that $x < 0$. Then $\log_b |x| = \log_b (-x)$ so by (6.7) on Page 134,

$$\begin{aligned}(\log_b |x|)' &= (\log_b (-x))' = (\log_b ((-1)\, x))' \\ &= \frac{1}{-x \ln b}(-1) = \frac{1}{x \ln b}.\end{aligned}$$

This proves the corollary.

Example 7.3.9 *Using properties of logarithms, simplify the expression, $\log_3\left(\frac{1}{9}x\right)$.*

From (7.7) - (7.9),

$$\begin{aligned}\log_3\left(\frac{1}{9}x\right) &= \log_3\left(\frac{1}{9}\right) + \log_3(x) \\ &= \log_3\left(3^{-2}\right) + \log_3(x) = -2 + \log_3(x).\end{aligned}$$

Example 7.3.10 *Using properties of logarithms, solve $5^{x-1} = 3^{2x+2}$.*

Take ln of both sides. Thus $(x-1)\ln 5 = (2x+2)\ln 3$. Then solving this for x yields $x = \frac{\ln 5 + 2\ln 3}{\ln 5 - 2\ln 3}$.

Example 7.3.11 *Solve* $\log_3(x)+2=\log_9(x+3)$.

From the given equation,

$$3^{\log_3(x)+2}=3^{\log_9(x+3)}=9^{\frac{1}{2}(\log_9(x+3))}=9^{\log_9\sqrt{(x+3)}}$$

and so $9x=\sqrt{x+3}$. Therefore, $x=\frac{1+\sqrt{1+12\times 81}}{2(81)}=\frac{1}{162}+\frac{1}{162}\sqrt{973}$. In the use of the quadratic formula, only one solution was possible. (Why?)

Example 7.3.12 *Compare* $\ln(x)$ *and* $\log_e(x)$

Recall that

$$\log_e(x)\equiv\frac{\ln x}{\ln e}.$$

Since $\ln e=1$ from the definition of e, it follows $\log_e(x)=\ln x$. These logarithms are called natural logarithms.

Example 7.3.13 *Find the derivative of the function,* $f(x)=\log_5(x)$.

$$f'(x)=\left(\frac{\ln(x)}{\ln 5}\right)'=\frac{1}{\ln 5}\frac{1}{x}.$$

Example 7.3.14 *Find the derivative of* $f(x)=\ln\left(x^3\right)$.

$$f'(x)=\left(\ln\left(x^3\right)\right)'=(3\ln x)'=\frac{3}{x}.$$

7.4 Exercises

1. Prove the last part of the Wild Assumption follows from the first part of this assumption. That is, show that if $b\neq 1$, then $\exp_b(h)\neq 1$ if $h\neq 0$ follows from the first part. **Hint:** If $b^h=1$ for $h\neq 0$, show $b^x=1$ for all $x\in\mathbb{R}$.

2. Simplify

   (a) $\log_4(16x)$.
   (b) $\log_3\left(27x^3\right)$
   (c) $(\log_b a)(\log_a b)$ for a,b positive real numbers not equal to 1.

3. Simplify

   (a) $\log_3(\log_3(27)-\log_3(9))$
   (b) $\log_2(3)\log_3(2)$
   (c) $\log_b(x)\log_a(b)$
   (d) $\log_{10}\left(100000^{1/3}\right)$

4. Find the derivatives of the following functions. You may want to do this by looking at the definition of the derivative in some cases.

   (a) $\log_5\left(x^2\right)$
   (b) $\log_3(5x+1)$

(c) $\log_6(\sqrt{x})$

(d) $\log_3\left(\sqrt{x^2+1}\right)$. **Hint:** Use the definition of the derivative for this one.

5. Find the derivatives of the following functions. You may want to do this by looking at the definition of the derivative in some cases.

   (a) 3^{x^2}

   (b) 2^{3x+1}

   (c) 5^{2x+7}

   (d) 7^{x^3}

   (e) $3^{\sqrt{x^2+1}}$

6. Explain why the function $6^x(1-x\ln 6)$ is never larger than 1. **Hint:** Consider $f(x) = 6^x(1-x\ln 6)$ and find its maximum value.

7. Solve $\log_2(x)+3=\log_2(3x+8)$.

8. Solve $\log_4(x)+3=\log_2(x+8)$.

9. Solve the equation $5^{2x+9}=7^x$ in terms of logarithms.

10. Using properties of logarithms, simplify the expression, $\log_4\left(\frac{1}{64}x\right)$.

11. Using properties of logarithms, solve $4^{x-1}=3^{2x+2}$.

12. The Wild Assumption gave the existence of a function, b^x satisfying certain properties. Show there can be no more than one such function. **Hint:** Recall the rational numbers were dense in $\mathbb{R}$ and so one can obtain a rational number arbitrarily close to a given real number. Exploit this and the assumed continuity of $\exp_b$ to obtain uniqueness.

13. Prove the function, b^x is concave up and $\log_b(x)$ is concave down.

14. Using properties of logarithms and exponentials, solve $3+\ln(-3x)=4+\ln 3x^2$.

15. Let f be a differentiable function and suppose $f(a)\geq 0$ and that $f'(x)\geq 0$ for $x\geq a$. Show that $f(x)\geq f(a)$ for all $x\geq a$. **Hint:** Use the mean value theorem.

16. Let e be defined by $\ln(e)=1$ and suppose $e<x<y$. Find a relationship between x^y and y^x. **Hint:** Use Problem 15 and at some point consider the function $h(x)=\frac{\ln x}{x}$.

17. Suppose f is any function defined on the positive real numbers and $f'(x)=g(x)$ where g is an odd function. ($g(-x)=-g(x)$.) Show $(f(|x|))'=g(x)$.

18. You know $(e^x)'=e^x$. Use the definition of the derivative to verify that $(e^{ax})'=ae^{ax}$. Show that Ae^{ax} solves the differential equation, $y'=ay$ along with the initial condition, $y(0)=A$.

19. Suppose $y'=ay$ and $y(0)=A$. Does it follow that $y=Ae^{ax}$? You know from Problem 18 that Ae^{ax} is one possibility. Is it the only possibility? **Hint:** Consider $y(x)e^{-ax}$ and use the product rule and Problem 18 to show $(y(x)e^{-ax})'=0$. Then review the mean value theorem. What does the mean value theorem say about functions whose derivatives are always equal to zero?

20. Show from the definition of the derivative that $(\sin(ax)) = a\cos(ax)$ and $(\cos(ax))' = -a\sin(ax)$. With this and Problem 18 find the derivatives of the following functions using the rules of derivatives. You will need to use the product and maybe the quotient rule along with properties of exponents.

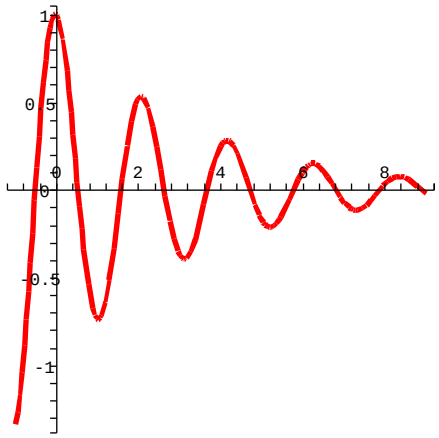
    (a) $\sin(3x)\,5^{6x}$

    (b) $\tan(2x)\,3^{3x+1}$

    (c) $\frac{\sin(3x)2^{5x}}{\cos(5x)3^{2x-1}}$

    (d) $\sin(2x)\cos(3x)\tan(4x)\,2^{3x}$

21. Using Problems 18 and 20 try to determine real numbers, b and a such that $y(t) = e^{bt}\cos at$ and $y(t) = e^{bt}\sin at$ both are solutions of the differential equation $y'' + 2\alpha y' + \beta^2 y = 0$ given that $\alpha^2 - \beta^2 < 0$. This is the equation of damped oscillation and will be discussed more carefully later. Typically y measures some sort of displacement from an equilibrium position and t represents time. Here is a picture of the graph of $y(t) = e^{-.3t}\cos(3t)$.



Properties And Applications Of Derivatives

8.0.1 Outcomes

1. Understand the proof of the chain rule and use the chain rule to differentiate composite functions.

2. Use the chain rule to find the derivative of an inverse function and to do implicit differentiation.

3. Understand the proof that inverse functions of differentiable functions can also be differentiated.

4. Understand the function x^r for r a real number and be able to differentiate function which involve raising to a real exponent.

5. Understand and use logarithmic differentiation.

6. Understand and use the inverse trigonometric functions and the inverse hyperbolic functions.

7. Understand and use L'Hôpital's rule.

8. Understand and use compound interest.

9. Understand and use the chain rule to solve related rates problems.

10. Use the derivative to set up and solve optimization problems.

11. Use the Newton Raphson method to find solutions to nonlinear equations.

8.1 The Chain Rule And Derivatives Of Inverse Functions

8.1.1 The Chain Rule

The chain rule is one of the most important of differentiation rules. Special cases of it are in Theorem 6.2.6. Now it is time to consider the theorem in full generality.

Theorem 8.1.1 *Suppose $f : (a,b) \to (c,d)$ and $g : (c,d) \to \mathbb{R}$. Also suppose that $f'(x)$ exists and that $g'(f(x))$ exists. Then $(g \circ f)'(x)$ exists and*

$$(g \circ f)'(x) = g'(f(x)) f'(x).$$

Proof: Define

$$H(h) \equiv \begin{cases} \frac{g(f(x+h))-g(f(x))}{f(x+h)-f(x)} \text{ if } f(x+h) - f(x) \neq 0 \\ g'(f(x)) \text{ if } f(x+h) - f(x) = 0 \end{cases}.$$

Then for $h \neq 0$,

$$\frac{g(f(x+h)) - g(f(x))}{h} = H(h) \frac{f(x+h) - f(x)}{h}.$$

Note that $\lim_{h\to 0} H(h) = g'(f(x))$ due to Theorems 5.9.6 on Page 112 and 6.2.2 on Page 133. Therefore, taking the limit and using Theorem 5.9.4,

$$\lim_{h\to 0} \frac{g(f(x+h)) - g(f(x))}{h} = g'(f(x)) f'(x).$$

This proves the chain rule.

Example 8.1.2 *Let $f(x) = \ln\left|\ln\left(x^4+1\right)\right|$. Find $f'(x)$.*

From the chain rule,

$$\begin{aligned} f'(x) &= \ln'\left(\ln\left(x^4+1\right)\right)\left(\ln\left(x^4+1\right)\right)' \\ &= \frac{1}{\ln(x^4+1)} \frac{1}{x^4+1}\left(x^4\right)' \\ &= \left(\frac{1}{\ln(x^4+1)}\right)\left(\frac{1}{x^4+1}\right)\left(4x^3\right). \end{aligned}$$

Example 8.1.3 *Let $f(x) = (2+\ln|x|)^3$. Find $f'(x)$.*

Use the chain rule again. Thus

$$\begin{aligned} f'(x) &= 3(2+\ln|x|)^2 (2+\ln|x|)' \\ &= \frac{3}{x}(2+\ln|x|)^2. \end{aligned}$$

8.1.2 Implicit Differentiation And Derivatives Of Inverse Functions

Sometimes a function is not given explicitly in terms of a formula. For example, you might have $x^2+y^2=4$. This relation defines y as a function of x near a given point such as $(0,1)$. Near this point, $y = \sqrt{4-x^2}$. Near the point, $(0,-1)$, you have $y = -\sqrt{4-x^2}$. Near the point, $(1,0)$, you can't solve for y in terms of x but you can solve for x in terms of y. Thus near $(1,0)$, $x = \sqrt{4-y^2}$. This was a simple example but in general, you can't use algebra to solve for one of the variables in terms of the others even if the relation defines that variable as a function of the others. Here is an example in which, even though it is impossible to find $y(x)$ you can still find the derivative of y. The procedure by which this is accomplished is nothing more than the chain rule and other rules of differentiation.

Example 8.1.4 *Suppose y is a differentiable function of x and $y^3+2yx = x^3+7+\ln|y|$. Find $y'(x)$.*

This illustrates the technique of implicit differentiation. If you believe y is some differentiable function of x, then you can differentiate both sides with respect to x and write, using the chain rule and product rule.

$$3y^2y' + 2xy' + 2y = 3x^2 + \frac{y'}{y}.$$

Now you can solve for y' and obtain $y' = -\frac{2y-3x^2}{3y^3+2xy-1}y$.

Of course there are significant mathematical considerations which are being ignored when it is assumed y is a differentiable function of x. It turns out that for problems like this, the equation relating x and y actually does define y as a differentiable function of x near points where it makes sense to formally solve for y' as just done. The theorems which give this justification are called the implicit and inverse function theorems. They are some of the most profound theorems in mathematics and are topics for advanced calculus. The interested reader should consult the book by Rudin, [23] for this and generalizations of all the hard theorems given in this book. One case is of special interest in which $y = f(x)$ and it is desired to find $\frac{dx}{dy}$ or in other words, the derivative of the inverse function.

It happens that if f is a differentiable one to one function defined on an interval, $[a,b]$, and $f'(x)$ exists and is non zero then the inverse function, f^{-1} has a derivative at the point $f(x)$. Recall that f^{-1} is defined according to the formula

$$f^{-1}(f(x)) = x.$$

Definition 8.1.5 *Let $f : [a,b] \to \mathbb{R}$ be a continuous function. Define*

$$f'(a) \equiv \lim_{x\to a+} \frac{f(x)-f(a)}{x-a},\ f'(b) \equiv \lim_{x\to b-} \frac{f(x)-f(b)}{x-b}.$$

Recall the notation $x \to a+$ means that only $x > a$ are considered in the definition of limit. The notation $x \to b-$ is defined similarly. Thus, this definition includes the derivative of f at the endpoints of the interval and to save notation,

$$f'(x_1) \equiv \lim_{x\to x_1} \frac{f(x)-f(x_1)}{x-x_1}$$

where it is understood that x is always in $[a,b]$.

Theorem 8.1.6 *Let $f : [a,b] \to \mathbb{R}$ be continuous and one to one. Suppose $f'(x_1)$ exists for some $x_1 \in [a,b]$ and $f'(x_1) \neq 0$. Then $\left(f^{-1}\right)'(f(x_1))$ exists and is given by the formula, $\left(f^{-1}\right)'(f(x_1)) = \frac{1}{f'(x_1)}$.*

Proof: By Lemma 5.7.4, and Corollary 5.7.6 on Page 105 f is either strictly increasing or strictly decreasing and f^{-1} is continuous. Therefore there exists $\eta > 0$ such that if $0 < |f(x_1) - f(x)| < \eta$, then

$$0 < |x_1 - x| = \left|f^{-1}(f(x_1)) - f^{-1}(f(x))\right| < \delta$$

where δ is small enough that for $0 < |x_1 - x| < \delta$,

$$\left|\frac{x-x_1}{f(x)-f(x_1)} - \frac{1}{f'(x_1)}\right| < \varepsilon.$$

It follows that if $0 < |f(x_1) - f(x)| < \eta$,

$$\left|\frac{f^{-1}(f(x)) - f^{-1}(f(x_1))}{f(x)-f(x_1)} - \frac{1}{f'(x_1)}\right| = \left|\frac{x-x_1}{f(x)-f(x_1)} - \frac{1}{f'(x_1)}\right| < \varepsilon$$

Therefore, since $\varepsilon > 0$ is arbitrary,

$$\lim_{y \to f(x_1)} \frac{f^{-1}(y) - f^{-1}(f(x_1))}{y - f(x_1)} = \frac{1}{f'(x_1)}$$

and this proves the theorem.

The following obvious corollary comes from the above by not bothering with end points.

Corollary 8.1.7 *Let $f : (a,b) \to \mathbb{R}$ be continuous and one to one. Suppose $f'(x_1)$ exists for some $x_1 \in (a,b)$ and $f'(x_1) \neq 0$. Then $\left(f^{-1}\right)'(f(x_1))$ exists and is given by the formula, $\left(f^{-1}\right)'(f(x_1)) = \frac{1}{f'(x_1)}$.*

This is one of those theorems which is very easy to remember if you neglect the difficult questions and simply focus on formal manipulations. Consider the following.

$$f^{-1}(f(x)) = x.$$

Now use the chain rule on both sides to write

$$\left(f^{-1}\right)'(f(x))\, f'(x) = 1,$$

and then divide both sides by $f'(x)$ to obtain

$$\left(f^{-1}\right)'(f(x)) = \frac{1}{f'(x)}.$$

Of course this gives the conclusion of the above theorem rather effortlessly and it is formal manipulations like this which aid many of us in remembering formulas such as the one given in the theorem.

Example 8.1.8 *Let $f(x) = \ln\left(1+x^2\right) + x^3 + 7$. Show that f has an inverse and find $\left(f^{-1}\right)'(7)$.*

I am not able to find a formula for the inverse function. This is typical in useful applications so you need to get used to this idea. The methods of algebra are insufficient to solve hard problems in analysis. You need something more. The question is to determine whether f has an inverse. To do this,

$$\begin{aligned} f'(x) &= \frac{2x}{1+x^2} + 3x^2 + 7 \\ &> -1 + 3x^2 + 7 \\ &> 6 > 0. \end{aligned}$$

By Corollary 6.8.5 on Page 148, this function is strictly increasing on $\mathbb{R}$ and so it has an inverse function although I have no idea how to find an explicit formula for this inverse function. However, I can see that $f(0) = 7$ and so by the formula for the derivative of an inverse function,

$$\begin{aligned} \left(f^{-1}\right)'(7) = \left(f^{-1}\right)'(f(0)) &= \frac{1}{f'(0)} \\ &= \frac{1}{7}. \end{aligned}$$

Example 8.1.9 *Suppose $f(a) = 0$ and $f'(x) = \sqrt{1 + x^4 + \ln(1+x^2)}$. Find $\left(f^{-1}\right)'(0)$.*

The function, f is one to one because it is strictly increasing due to the fact that its derivative is positive for all x. As in the last example, I have no idea how to find a formula for f^{-1} but I do see that $f(a) = 0$ and so

$$\left(f^{-1}\right)'(0) = \left(f^{-1}\right)'(f(a)) = \frac{1}{f'(a)} = \frac{1}{\sqrt{1+a^4+\ln(1+a^2)}}.$$

The chain rule has a particularly attractive form in Leibniz's notation. Suppose $y = g(u)$ and $u = f(x)$. Thus $\mathbf{y} = g \circ f(x)$. Then from the above theorem

$$\begin{aligned}(g \circ f)'(x) &= g'(f(x))\, f'(x) \\ &= g'(u)\, f'(x)\end{aligned}$$

or in other words,

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}.$$

Notice how the $du's$ cancel. This particular form is a very useful crutch and is used extensively in applications.

8.2 Exercises

1. In each of the following, find $\frac{dy}{dx}$.

   (a) $y = e^{\sin x}$

   (b) $y = \sqrt{7 + x^2 + \sin x}$

   (c) $y = \ln\left(x^2+1\right)$

   (d) $y = \sin\left(\ln\left(x^2+1\right)\right)$

   (e) $y = \ln(\sin(x)+3)$

   (f) $y = \ln(\tan(x))$

   (g) $y = \sin^2(\ln(\tan(x)))$

   (h) $y = \sin\left((x+\tan x)^6\right)$

   (i) $y = \ln\left(\sin\left(x^2+7\right)\right)$

   (j) $y = \tan\left(\cos\left(x^2\right)\right)$

   (k) $y = \log_2(\sin(x)+6)$

   (l) $y = \sin\left(\log_3\left(x^2+1\right)\right)$

   (m) $y = \frac{\sqrt{x^3+7}}{\sqrt{\sin(x)+4}}$

   (n) $y = 3^{\tan(\sin(x))}$

   (o) $y = \left(\frac{x^2+2x}{\tan(x^2+1)}\right)^6$

2. In each of the following, assume the relation defines y as a function of x for values of x and y of interest and use the process of implicit differentiation to find $y'(x)$.

   (a) $xy^2 + \sin(y) = x^3 + 1$

   (b) $y^3 + x\cos\left(y^2\right) = x^4$

(c) $y\cos(x) = \tan(y)\cos\left(x^2\right) + 2$

(d) $\left(x^2+y^2\right)^6 = x^3y + 3$

(e) $\frac{xy^2+y}{y^5+x} + \cos(y) = 7$

(f) $\sqrt{x^2+y^4}\sin(y) = 3x$

(g) $y^3\sin(x) + y^2x^2 = 2^{x^2}y + \ln|y|$

(h) $y^2\sin(y)\,x + \log_3(xy) = y^2 + 11$

(i) $\sin\left(x^2+y^2\right) + \sec(xy) = e^{x+y} + y2^y + 2$

(j) $\sin\left(\tan\left(xy^2\right)\right) + y^3 = 16$

(k) $\cos(\sec(\tan(y))) + \ln(5+\sin(xy)) = x^2y + 3$

3. In each of the following, assume the relation defines y as a function of x for values of x and y of interest and use the process of implicit differentiation to show y satisfies the given differential equation.

(a) $x^2y + \sin y = 7,\ \left(x^2+\cos y\right)y' + 2xy = 0.$

(b) $x^2y^3 + \sin\left(y^2\right) = 5,\ 2xy^3 + \left(3x^2y^2 + 2\left(\cos\left(y^2\right)\right)y\right)y' = 0.$

(c) $y^2\sin(y) + xy = 6,\ \left(2y(\sin(y)) + y^2(\cos(y)) + x\right)y' + y = 0.$

(d) $\tan\left(x^2+y\right)+x^x = 7,\ \left(1+\tan^2\left(x^2+y\right)\right)y'+\left(1+\tan^2\left(x^2+y\right)\right)2x+x^x(\ln x+1) = 0.$

(e) $y^{x^2+\sin(x)} = 5,\ y^{\left(x^2+\sin x\right)}\left((2x+\cos x)\ln(y) + \left(x^2+\sin x\right)\frac{y'}{y}\right) = 0.$

(f) $y^{x^2+y} = 2,\ y^{\left(x^2+y\right)}\left((2x+y')\ln(y) + \left(x^2+y\right)\frac{y'}{y}\right) = 0.$

4. Show that if $D(g) \subseteq U \subseteq D(f)$, and if f and g are both one to one, then $f\circ g$ is also one to one.

5. Using Problem 4 show that the following functions are one to one and find the derivative of the inverse function at the indicated point.

(a) $y = e^{x^3+1}, e^2$

(b) $y = \left(x^3+7x+1\right)^3, 0$

(c) $y = \tan\left(x^3+\frac{\pi}{4}\right), 1$

(d) $y = \tan\left(-x^5+\frac{\pi}{4}\right), 1$

(e) $y = 2^{5x+\sin(x)}, 5\left(\frac{\pi}{2}\right)+1$

8.3 The Function x^r For r A Real Number

Theorem 7.3.2 on Page 159 says that for $x > 0$, and for r a real number,

$$\ln\left(x^r\right) = r\ln(x) \tag{8.1}$$

By this theorem, it also follows that $\ln^{-1}:\mathbb{R}\to(0,\infty)$ exists. Then by Corollary 8.1.7 $\ln^{-1}$ is differentiable. Also, for all $x\in\mathbb{R}$,

$$\ln\left(\ln^{-1}(x)\right) = x,\ \ln\left(\exp_e(x)\right) = \ln\left(e^x\right) = x\ln e = x.$$

Since ln is one to one, $\ln^{-1}(x) = \exp_e(x) = \exp(x)$. Thus

$$\exp(\ln x) = \ln^{-1}(\ln(x)) = x,\ \ln(\exp(y)) = y. \tag{8.2}$$

From (7.13) on Page 161, $\exp'(x) = \exp(x)$. Recall why this was. From the definition of ln and the fact $\ln e = 1$,

$$\exp'(x) = \exp_e'(x) \equiv \ln(e)\exp_e(x) = \exp(x).$$

With this understanding, it becomes possible to find derivatives of functions raised to arbitrary real powers. First, note that upon taking exp of both sides of (8.1) and using (8.2),

$$x^r = \exp(r\ln x). \tag{8.3}$$

Theorem 8.3.1 *For $x > 0, (x^r)' = rx^{r-1}$.*

Proof: Differentiate both sides of (8.3) using the chain rule. From the Wild Assumption on Page 159, in particular, the part about the validity of the laws of exponents,

$$(x^r)' = \exp'(r\ln x)\frac{r}{x} = \exp(r\ln x)\frac{r}{x} = \frac{r}{x}x^r = rx^{r-1}. \tag{8.4}$$

and this shows from (8.4) that $(x^r)' = rx^{r-1}$ as claimed.

Example 8.3.2 *Suppose $f(x)$ is a non zero differentiable function. Find the derivative of $|f(x)|^r$.*

From (8.3),

$$|f(x)|^r = \exp(r\ln|f(x)|).$$

Therefore,

$$\begin{aligned}(|f(x)|^r)' &= \exp(r\ln|f(x)|)(r\ln|f(x)|)' \\ &= |f(x)|^r r\frac{f'(x)}{f(x)} = r|f(x)|^{r-2}f(x)f'(x).\end{aligned}$$

8.3.1 Logarithmic Differentiation

Example 8.3.3 *Let $f(x) = \left(1+x^2\right)^x$. Find $f'(x)$.*

One way to do this is to take ln of both sides and use the chain rule to differentiate both sides with respect to x. Thus

$$\ln(f(x)) = x\ln\left(1+x^2\right)$$

and so, taking the derivative of both sides, using the chain and product rules,

$$\frac{f'(x)}{f(x)} = \frac{2x^2}{1+x^2} + \ln\left(1+x^2\right).$$

Then solve for $f'(x)$ to obtain

$$f'(x) = \left(1+x^2\right)^x\left(\frac{2x^2}{1+x^2} + \ln\left(1+x^2\right)\right).$$

This process is called logarithmic differentiation.

Example 8.3.4 *Let* $f(x) = \frac{\sqrt[3]{x^3+\sin(x)}}{\sqrt[6]{x^4+2x}}$. *Find* $f'(x)$.

You could use the quotient and chain rules but it is easier to use logarithmic differentiation.

$$\ln(f(x)) = \frac{1}{3}\ln\left(x^3+\sin(x)\right) - \frac{1}{6}\ln\left(x^4+2x\right).$$

Differentiating both sides,

$$\frac{f'(x)}{f(x)} = \frac{1}{3}\left(\frac{3x^2+\cos x}{x^3+\sin x}\right) - \frac{1}{6}\frac{4x^3+2}{x^4+2x}.$$

Therefore, the answer is

$$f'(x) = \frac{\sqrt[3]{x^3+\sin(x)}}{\sqrt[6]{x^4+2x}}\left(\frac{1}{3}\left(\frac{3x^2+\cos x}{x^3+\sin x}\right) - \frac{1}{6}\frac{4x^3+2}{x^4+2x}\right).$$

I think you can see the advantage of doing it this way over using the quotient rule.

8.4 Exercises

1. Let $f(x) \equiv x^3+1$. Find $f^{-1}(y)$. Now find $\left(f^{-1}\right)'(1)$.
2. Let $f(x) \equiv x^3+7x+3$. Explain why f has an inverse. Find $\left(f^{-1}\right)'(3)$.
3. Derive the quotient rule from the product rule and the chain rule. This shows you don't need to remember the wretched quotient rule if you don't want to. It follows from two rules which you cannot survive without.
4. What is wrong with the following "proof" of the chain rule? Here $g'(f(x))$ exists and $f'(x)$ exists.

$$\lim_{h\to 0}\frac{g(f(x+h))-g(f(x))}{h}$$

$$= \lim_{h\to 0}\frac{g(f(x+h))-g(f(x))}{f(x+h)-f(x)}\frac{f(x+h)-f(x)}{h}$$
$$= g'(f(x))f'(x).$$

5. Is the derivative of a function always continuous? **Hint:** Consider a differentiable function f which is periodic of period 1 and non constant. (Periodic of period 1 means $f(x+1) = f(x)$ for all $x \in \mathbb{R}$.) Now consider

$$h(x) = \begin{cases} x^2 f\left(\frac{1}{x}\right) \text{ if } x \neq 0 \\ 0 \text{ if } x = 0 \end{cases}.$$

   Show $h'(0) = 0$. What is $h'(x)$ for $x \neq 0$? Is h' also periodic of period 1?

6. Let $f(x) = x^3+1$. Find an explicit formula for f^{-1} and use it to compute $\left(f^{-1}\right)'(9)$. Then use the formula given in the theorem of this section to see you get the same answer.
7. Find the derivatives of the following functions.

   (a) $\sin\left(x^2\right)\ln\left(x^2+1\right)$

(b) $\ln\left(1+x^2\right)$

(c) $\left(x^3+1\right)^6 \sin\left(x^2+7\right)$

(d) $\ln\left(\left(x^3+1\right)^6 \sin\left(x^2+7\right)\right)$

(e) $\tan\left(\sec\left(\sin\left(x^2+1\right)\right)\right)$

(f) $\left(\sin^2\left(x^2+5\right)\right)^{\sqrt{7}}$

8. Use (8.3) or logarithmic differentiation to differentiate the following functions.

(a) $\left(2+\sin\left(x^2+6\right)\right)^{\tan x}$

(b) x^x

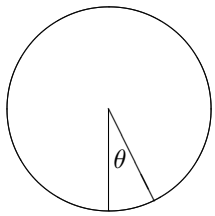
(c) $\left(x^x\right)^x$

(d) $\left(\tan^2\left(x^4+4\right)+1\right)^{\cos x}$

(e) $\left(\sin^2(x)\right)^{\tan x}$

9. A search light at a prison revolves five times every minute and is located 200 yards from a long straight wall. How fast in yards per minute is the light moving along the wall at a distance of 100 yards from the point of the wall closest to the light? If you are an escaping convict, where is the most dangerous location on this wall?

10. A circular disk of paper has a circular sector removed and then the edges are joined to form a cone. What is the angle of the removed circular sector which will create the cone of largest volume?



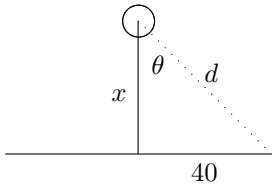
11. Let θ be the angle between the two equal sides of an iscoceles triangle. Suppose the sum of the lengths of the sides of this triangle is L. Find the value of θ which will maximize the area of the triangle.

12. A window has total perimeter equal to L. It consists of a square surmounted by an isoceles triangle. Find the dimensions which will make the largest window.

13. A car is proceeding down a road near Crystal Falls Michigan at 60 miles per hour. Fifty yards in front of the car there is a suicidal deer standing 50 feet from the road[1]. Consider the angle formed by the road and the line of sight to the deer. How fast is this angle changing? What happens to the rate of change of this angle as the driver gets closer to the deer?

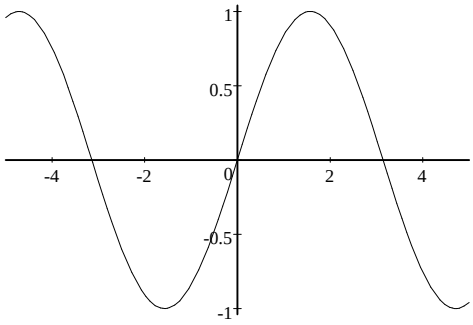
[1] Crystal Falls Michigan is in the upper peninsula. It is a very beautiful place, especially in the fall when the leaves change colors but it is hazardous to drive there, especially at dusk, because of suicidal deer which jump out in front of cars unexpectedly. These are very large dear.

14. A light pole with a light on top of it is intended to illuminate the edge of a circle of radius 40 feet centered at the base of the light pole. The brightness of the illumination is of the form $k\frac{\cos\theta}{d^2}$ where θ and d are given in the following picture. Find the height, x which will result in the brightest illumination at the edge of this circle.



8.5 The Inverse Trigonometric Functions

It is desired to consider the inverse trigonometric functions. Graphing the function $y = \sin x$, is clear sin is not one to one on $\mathbb{R}$ and so it is not possible to define an inverse function.



However, a little thing like this will not prevent the definition of useful inverse trig. functions. Observe the function, sin, is one to one on the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ shown in the above picture as the interval containing zero on which the function climbs from -1 to 1. Also note that for y on this interval, $\cos(y) \geq 0$. Now the arcsin function is defined as the inverse of the sin when its domain is restricted to $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. In words, $\arcsin(x)$ is defined to be the angle whose sine is x which lies in $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. From Theorem 8.1.6 on Page 169 about the derivative of the inverse function, the derivative of $x \to \arcsin(x)$ exists for all $x \in [-1, 1]$. The formula in this theorem could be used to find the derivative of arcsin but it is more useful to simply use that theorem to resolve the existence question and apply the chain rule to find the formula. It is a mistake to memorize too many formulas. Let $y = \arcsin(x)$ so $\sin(y) = x$. Now taking the derivative of both sides,

$$\cos(y)\, y' = 1$$

and so

$$y' = \frac{1}{\cos y} = \frac{1}{\sqrt{1 - \sin^2(y)}} = \frac{1}{\sqrt{1 - x^2}}.$$

The positive value of the square root is used because for $y \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, $\cos(y) \geq 0$. Thus

$$\frac{1}{\sqrt{1 - x^2}} = \arcsin'(x). \tag{8.5}$$

Next consider the inverse tangent function. You are aware that tan is periodic of period

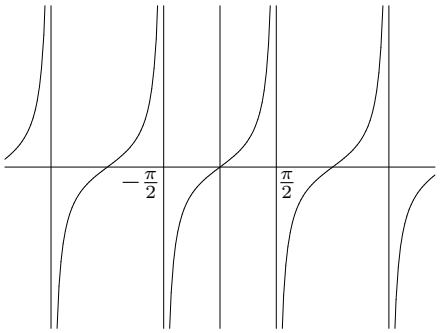
π because

$$\begin{aligned}\tan(x+\pi) &= \frac{\sin(x+\pi)}{\cos(x+\pi)} = \frac{\sin(x)\cos(\pi)+\cos(x)\sin(\pi)}{\cos(x)\cos(\pi)-\sin(x)\sin(\pi)} \\ &= \frac{-\sin(x)}{-\cos(x)} = \tan(x).\end{aligned}$$

Therefore, it is impossible to take the inverse of tan. However, tan is one to one on $\left(-\frac{\pi}{2},\frac{\pi}{2}\right)$ and

$$\lim_{x\to\frac{\pi}{2}-}\tan(x)=+\infty, \quad \lim_{x\to-\frac{\pi}{2}+}\tan(x)=-\infty$$

as shown in the following graph of $y=\tan(x)$ in which the vertical lines represent vertical asymptotes.



Therefore, $\arctan(x)$ for $x\in(-\infty,\infty)$ is defined according to the rule: $\arctan(x)$ is the angle whose tangent is x which is in $\left(-\frac{\pi}{2},\frac{\pi}{2}\right)$. By Theorem 8.1.6 arctan has a derivative. Therefore, letting $y=\arctan(x)$, $\tan(y)=x$ and by the chain rule,

$$\sec^2(y)\,y'=1.$$

Therefore,

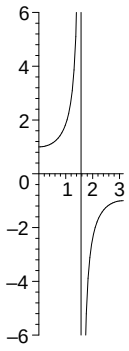
$$y'=\frac{1}{\sec^2(y)}=\frac{1}{1+\tan^2(y)}=\frac{1}{1+x^2}.$$

and so

$$\frac{1}{1+x^2}=\arctan'(x). \tag{8.6}$$

The inverse secant function can be defined similarly. There is no agreement on the best way to restrict the domain of sec. I will follow the way of doing it which is used in the book by Salas and Hille [25] recognizing that there are good reasons for doing it other ways also. The graph of sec is represented below on $[0,\frac{\pi}{2})\cup(\frac{\pi}{2},\pi]$. There is a vertical asymptote at $x=\frac{\pi}{2}$. Thus

$$\lim_{x\to\frac{\pi}{2}-}\sec(x)=+\infty, \quad \lim_{x\to\frac{\pi}{2}+}\sec(x)=-\infty$$



As in the case of arcsin and arctan, $\operatorname{arcsec}(x)$ is the angle whose secant is x which lies in $[0, \frac{\pi}{2}) \cup (\frac{\pi}{2}, \pi]$. Let $y = \operatorname{arcsec}(x)$ so $x = \sec(y)$ and using the chain rule,

$$1 = \sec(y)\tan(y)\, y'.$$

Now from the trig. identity, $1 + \tan^2(y) = \sec^2(y)$,

$$\begin{aligned} y' &= \frac{1}{\sec(y)\tan(y)} \\ &= \frac{1}{x\left(\pm\sqrt{x^2-1}\right)} \end{aligned}$$

and it is necessary to consider what to do with $\pm$. If $y \in [0, \frac{\pi}{2})$, both $x = \sec(y)$ and $\tan(y)$ are nonnegative and so in this case,

$$y' = \frac{1}{x\sqrt{x^2-1}} = \frac{1}{|x|\sqrt{x^2-1}}.$$

If $y \in (\frac{\pi}{2}, \pi]$, then $x = \sec(y) < 0$ and $\tan(y) \leq 0$ so

$$y' = \frac{1}{x(-1)\sqrt{x^2-1}} = \frac{1}{(-x)\sqrt{x^2-1}} = \frac{1}{|x|\sqrt{x^2-1}}.$$

Thus either way,

$$y' = \frac{1}{|x|\sqrt{x^2-1}}.$$

This yields the formula

$$\frac{1}{|x|\sqrt{x^2-1}} = \operatorname{arcsec}'(x). \tag{8.7}$$

As in the case of ln, there is an interesting and useful formula involving $\operatorname{arcsec}(|x|)$. For $x < 0$, this function equals $\operatorname{arcsec}(-x)$ and so by the chain rule, its derivative equals

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

If $x > 0$, this function equals $\operatorname{arcsec}(x)$ and its derivative equals

$$\frac{1}{|x|\sqrt{x^2-1}} = \frac{1}{x\sqrt{x^2-1}}$$

and so either way,

$$(\operatorname{arcsec}(|x|))' = \frac{1}{x\sqrt{x^2-1}}. \tag{8.8}$$

8.6 The Hyperbolic And Inverse Hyperbolic Functions

The hyperbolic functions are given by

$$\sinh(x) \equiv \frac{e^x - e^{-x}}{2}, \cosh(x) \equiv \frac{e^x + e^{-x}}{2},$$

and

$$\tanh(x) \equiv \frac{\sinh(x)}{\cosh(x)}.$$

The first of these is called the hyperbolic sine and the second the hyperbolic cosine. I imagine you can guess what the third is called. If you guessed "hyperbolic tangent" you got it right. The other hyperbolic functions are defined by analogy to the circular functions.

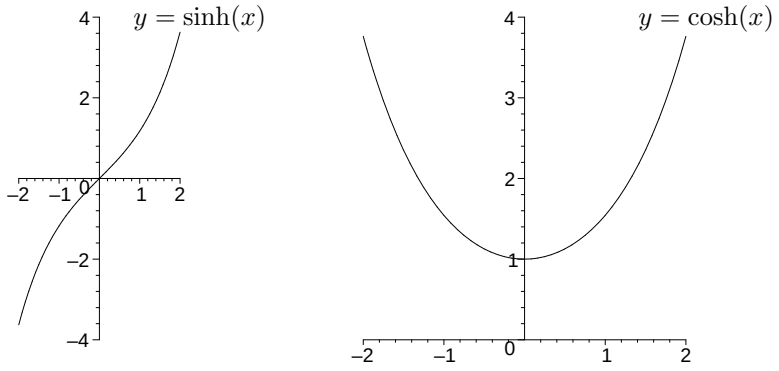
The reason these are called hyperbolic functions is that

$$\cosh^2 t - \sinh^2 t = 1 \tag{8.9}$$

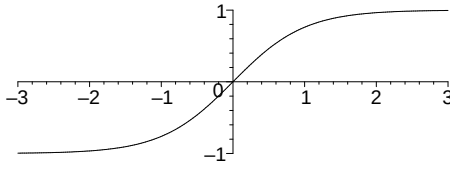
and so the point, $(\cosh t, \sinh t)$ is a point on the hyperbola whose equation is $x^2 - y^2 = 1$. This is not important but is the source for the term hyperbolic. Using the chain rule,

$$\cosh'(x) = \sinh(x), \ \sinh'(x) = \cosh x.$$

Also, you see that $\sinh(0) = 0$, $\cosh(0) = 1$ and that $\sinh(x) < 0$ if $x < 0$ while $\sinh(x) > 0$ for $x > 0$, but $\cosh(x) > 0$ for all x. Therefore, sinh is an increasing function, concave down for $x < 0$ and concave up for $x > 0$ because $\sinh''(x) = \sinh(x)$ while cosh is decreasing for $x < 0$ and increasing for $x > 0$. Since $\cosh''(x) = \cosh(x)$, it is concave up for all x. Thus the graphs of these functions are as follows.



Also, you can use the graph of the function $x \to \exp(x) = e^x$ to verify that the graph of $\tanh(x)$ is as given below



Since $x \to \sinh(x)$ is strictly increasing, it has an inverse function, $\sinh^{-1}(x)$. If $y = \sinh^{-1}(x)$, then $\sinh(y) = x$ and so using the chain rule and the theorem about the existence of the derivative of the inverse function, $y' \cosh(y) = 1$. From the identity (8.9) $\cosh(y) = \sqrt{1 + \sinh^2(y)} = \sqrt{1 + x^2}$. Therefore,

$$y' = \frac{1}{\sqrt{1 + x^2}}$$

which gives the formula

$$\frac{1}{\sqrt{1 + x^2}} = \left(\sinh^{-1}\right)'(x). \tag{8.10}$$

The derivative of the hyperbolic tangent is also easy to find. This yields after a short computation

$$(\tanh x)' = 1 - \tanh^2 x. \tag{8.11}$$

Another notation for the inverse hyperbolic functions which is sometimes used is arcsinh or arccosh or arctanh by analogy with the inverse trig. functions.

8.7 Exercises

1. Verify (8.11).
2. Simplify the following.
   (a) $\sin(\arctan(1))$
   (b) $\cos\left(\arctan\sqrt{3}\right)$
   (c) $\tan\left(\arcsin\left(\sqrt{3}/2\right)\right)$
   (d) $\sec\left(\arcsin\left(\sqrt{2}/2\right)\right)$
   (e) $\tan(\arcsin(1/2))$
   (f) $\cos\left(\arctan\left(\sqrt{3}\right) - \arcsin\left(\sqrt{2}/2\right)\right)$
   (g) $\sin(\arctan(x))$
   (h) $\cos(\arcsin(x))$
   (i) $\cos(2\arcsin(x))$
   (j) $\sec(2\arctan(x))$
   (k) $\tan\left(2\arcsin(x) + \arctan\left(x^2\right)\right)$
   (l) $\sec(\arcsin(x) + 2\arccos(x))$
3. Find the derivatives and give the domains of the following functions.
   (a) $\arcsin\left(x^2 + 3x\right)$
   (b) $\arctan(3x + 5)$
   (c) $\operatorname{arcsec}(2x + 1)$
   (d) $\arccos(\sin(x))$
   (e) $\arctan(\sinh(x))$
   (f) $\sinh(\tan(3x))$
   (g) $\cosh(\csc(3x))$

(h) $\tanh(\sec(3x))$

4. For a and b positive constants, find $\frac{d}{dx}\left(\frac{\arctan\left(\frac{bx}{a}\right)}{ab}\right)$.

5. Find $\frac{d}{dx}\left(\arctan\frac{2}{\sqrt{(9x^2-4)}}+\operatorname{arcsec}\left|\frac{3x}{2}\right|\right)$.

6. Use the process of implicit differentiation to find y' in the following examples.

   (a) $\arcsin\left(x^2y\right)+\arctan\left(xy^2\right)=\sinh(y)+3$

   (b) $x^2\tanh(yx)+\arctan\left(x^2\right)=5$

   (c) $y^4\sinh(x)+\arccos\left(x^2+y^2\right)=1$

   (d) $\sinh\left(x\tanh\left(x^2+y^2\right)\right)=2$

7. Find the derivatives of the following functions.

   (a) $\sinh^{-1}\left(x^2+7\right)$

   (b) $\tanh^{-1}(x)$

   (c) $\sin\left(\sinh^{-1}\left(x^2+2\right)\right)$

   (d) $\sin(\tanh(x))$

   (e) $x^2\sinh(\sin(\cos(x)))$

   (f) $\left(\cosh\left(x^3\right)\right)^{\sqrt{6}}$

   (g) $\left(1+x^4\right)^{\sin x}$

8. Simplify for $x>0$, $\arcsin x+\arccos x$.

9. A wonderful identity which was used to compute π for over 200 years[2] is the following.

$$\frac{\pi}{4}=4\arctan\left(\frac{1}{5}\right)-\arctan\left(\frac{1}{239}\right).$$

   Establish this identity by taking the tangent of both sides and using an appropriate formula for the tangent of the difference of two angles. Use De Moivre's theorem to get some help in finding a formula for $\tan(4\theta)$.

10. Find a formula for $\tanh^{-1}$ in terms of ln.

11. Find a formula for $\sinh^{-1}$ in terms of ln.

12. Prove $1-\tanh^2 x=\operatorname{sech}^2 x$.

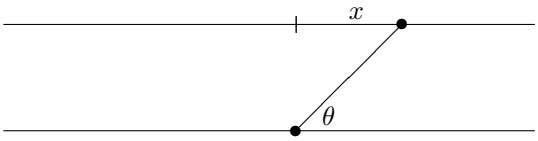
13. Prove $\coth^2 x-1=\operatorname{csch}^2 x$.

14. What about $\cosh^{-1}$? Define it by restricting the domain of cosh to be nonnegative numbers? What is $\cosh^{-1}$ in terms of ln?

[2] John Machin computed π to 100 decimal places in 1706 through the use of this identity. Later in 1873 William Shanks did it to over 700 places using this identity. The next advance was in 1948, 808 decimal places. After this, computers began to be used and currently π is "known" to millions of decimal places. Many other schemes have been used besides this identity for computing π.

15. Show $\arcsin x = \arctan\left(\frac{x}{\sqrt{1-x^2}}\right)$. It is possible to start with the arctan function and obtain all the other trig functions in terms of this one. If you knew the function, arctan explain how to define sin and cos. This is interesting because there is a simple way to define arctan directly as a function of a real variable[15]. Approaches like these avoid all reference to plane geometry.

16. A divided highway is separated by a median which is 1/10 of a mile wide. Two cars pass each other, one going west and the other east. The west bound car is travelling at 70 miles per hour while the east bound car is travelling at 60 miles per hour. The situation is described in the following picture.



How fast is the distance between the two cars changing when x ,shown in the picture, equals 1/10 mile? How fast is the angle, θ, shown in the picture, changing at this instant?

8.8 L'Hôpital's Rule

There is an interesting rule which is often useful for evaluating difficult limits called L'Hôpital's[3] rule. The best versions of this rule are based on the Cauchy Mean value theorem, Theorem 6.8.2 on Page 147.

Theorem 8.8.1 *Let $[a,b] \subseteq [-\infty,\infty]$ and suppose f, g are functions which satisfy,*

$$\lim_{x\to b-} f(x) = \lim_{x\to b-} g(x) = 0, \tag{8.12}$$

and f' and g' exist on (a,b) with $g'(x) \neq 0$ on (a,b). Suppose also that

$$\lim_{x\to b-} \frac{f'(x)}{g'(x)} = L. \tag{8.13}$$

Then

$$\lim_{x\to b-} \frac{f(x)}{g(x)} = L. \tag{8.14}$$

Proof: By the definition of limit and (8.13) there exists $c < b$ such that if $t > c$, then

$$\left|\frac{f'(t)}{g'(t)} - L\right| < \frac{\varepsilon}{2}.$$

Now pick x, y such that $c < x < y < b$. By the Cauchy mean value theorem, there exists $t \in (x,y)$ such that

$$g'(t)(f(x) - f(y)) = f'(t)(g(x) - g(y)).$$

[3]L'Hôpital published the first calculus book in 1696. This rule, named after him, appeared in this book. The rule was actually due to Bernoulli who had been L'Hôpital's teacher. L'Hôpital did not claim the rule as his own but Bernoulli accused him of plagarism. Nevertheless, this rule has become known as L'Hôpital's rule ever since. The version of the rule presented here is superior to what was discovered by Bernoulli and depends on the Cauchy mean value theorem which was found over 100 years after the time of L'Hôpital.

Since $g'(s) \neq 0$ for all $s \in (a,b)$ it follows $g(x) - g(y) \neq 0$. Therefore,

$$\frac{f'(t)}{g'(t)} = \frac{f(x) - f(y)}{g(x) - g(y)}$$

and so, since $t > c$,

$$\left| \frac{f(x) - f(y)}{g(x) - g(y)} - L \right| < \frac{\varepsilon}{2}.$$

Now letting $y \to b-$,

$$\left| \frac{f(x)}{g(x)} - L \right| \leq \frac{\varepsilon}{2} < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this shows (8.14).

The following corollary is proved in the same way.

Corollary 8.8.2 *Let $[a,b] \subseteq [-\infty, \infty]$ and suppose f, g are functions which satisfy,*

$$\lim_{x \to a+} f(x) = \lim_{x \to a+} g(x) = 0, \tag{8.15}$$

and f' and g' exist on (a,b) with $g'(x) \neq 0$ on (a,b). Suppose also that

$$\lim_{x \to a+} \frac{f'(x)}{g'(x)} = L. \tag{8.16}$$

Then

$$\lim_{x \to a+} \frac{f(x)}{g(x)} = L. \tag{8.17}$$

Here is a simple example which illustrates the use of this rule.

Example 8.8.3 *Find* $\lim_{x \to 0} \frac{5x + \sin 3x}{\tan 7x}$.

The conditions of L'Hôpital's rule are satisfied because the numerator and denominator both converge to 0 and the derivative of the denominator is nonzero for x close to 0. Therefore, if the limit of the quotient of the derivatives exists, it will equal the limit of the original function. Thus,

$$\lim_{x \to 0} \frac{5x + \sin 3x}{\tan 7x} = \lim_{x \to 0} \frac{5 + 3\cos 3x}{7 \sec^2(7x)} = \frac{8}{7}.$$

Sometimes you have to use L'Hôpital's rule more than once.

Example 8.8.4 *Find* $\lim_{x \to 0} \frac{\sin x - x}{x^3}$.

Note that $\lim_{x \to 0} (\sin x - x) = 0$ and $\lim_{x \to 0} x^3 = 0$. Also, the derivative of the denominator is nonzero for x close to 0. Therefore, if $\lim_{x \to 0} \frac{\cos x - 1}{3x^2}$ exists and equals L, it will follow from L'Hôpital's rule that the original limit exists and equals L. However, $\lim_{x \to 0} (\cos x - 1) = 0$ and $\lim_{x \to 0} 3x^2 = 0$ so L'Hôpital's rule can be applied again to consider $\lim_{x \to 0} \frac{-\sin x}{6x}$. From L'Hôpital's rule, if this limit exists and equals L, it will follow that $\lim_{x \to 0} \frac{\cos x - 1}{3x^2} = L$ and consequently $\lim_{x \to 0} \frac{\sin x - x}{x^3} = L$. But from Lemma 7.1.1 on Page 153, $\lim_{x \to 0} \frac{-\sin x}{6x} = \frac{-1}{6}$. Therefore, by L'Hôpital's rule, $\lim_{x \to 0} \frac{\sin x - x}{x^3} = \frac{-1}{6}$.

Warning 8.8.5 ***Be sure to check the assumptions of L'Hôpital's rule before using it.***

Example 8.8.6 *Find* $\lim_{x\to 0+}\frac{\cos 2x}{x}$.

The numerator becomes close to 1 and the denominator gets close to 0. Therefore, the assumptions of L'Hôpital's rule do not hold and so it does not apply. In fact there is no limit unless you define the limit to equal $+\infty$. Now lets try to use the conclusion of L'Hôpital's rule even though the conditions for using this rule are not verified. Take the derivative of the numerator and the denominator which yields $\frac{-2\sin 2x}{1}$, an expression whose limit as $x\to 0+$ equals 0. This is a good illustration of the above warning.

Some people get the unfortunate idea that one can find limits by doing experiments with a calculator. If the limit is taken as x gets close to 0, these people think one can find the limit by evaluating the function at values of x which are closer and closer to 0. Theoretically, this should work although you have no way of knowing how small you need to take x to get a good estimate of the limit. In practice, the procedure may fail miserably.

Example 8.8.7 *Find* $\lim_{x\to 0}\frac{\ln\left|1+x^{10}\right|}{x^{10}}$.

This limit equals $\lim_{y\to 0}\frac{\ln|1+y|}{y}=\lim_{y\to 0}\frac{\left(\frac{1}{1+y}\right)}{1}=1$ where L'Hôpital's rule has been used. This is an amusing example. You should plug .001 in to the function, $\frac{\ln\left|1+x^{10}\right|}{x^{10}}$ and see what your calculator or computer gives you. If it is like mine, it will give the answer, 0 and will keep on returning the answer of 0 for smaller numbers than .001. This illustrates the folly of trying to compute limits through calculator or computer experiments.

There is another form of L'Hôpital's rule in which $\lim_{x\to b-}f(x)=\pm\infty$ and $\lim_{x\to b-}g(x)=\pm\infty$.

Theorem 8.8.8 *Let* $[a,b]\subseteq[-\infty,\infty]$ *and suppose* f,g *are functions which satisfy,*

$$\lim_{x\to b-}f(x)=\pm\infty \text{ and } \lim_{x\to b-}g(x)=\pm\infty, \tag{8.18}$$

and f' *and* g' *exist on* (a,b) *with* $g'(x)\neq 0$ *on* (a,b). *Suppose also*

$$\lim_{x\to b-}\frac{f'(x)}{g'(x)}=L. \tag{8.19}$$

Then

$$\lim_{x\to b-}\frac{f(x)}{g(x)}=L. \tag{8.20}$$

Proof: By the definition of limit and (8.19) there exists $c<b$ such that if $t>c$, then

$$\left|\frac{f'(t)}{g'(t)}-L\right|<\frac{\varepsilon}{2}.$$

Now pick x,y such that $c<x<y<b$. By the Cauchy mean value theorem, there exists $t\in(x,y)$ such that

$$g'(t)(f(x)-f(y))=f'(t)(g(x)-g(y)).$$

Since $g'(s)\neq 0$ on (a,b), it follows $g(x)-g(y)\neq 0$. Therefore,

$$\frac{f'(t)}{g'(t)}=\frac{f(x)-f(y)}{g(x)-g(y)}$$

and so, since $t>c$,

$$\left|\frac{f(x)-f(y)}{g(x)-g(y)}-L\right|<\frac{\varepsilon}{2}.$$

Now this implies

$$\left|\frac{f(y)}{g(y)}\frac{\left(\frac{f(x)}{f(y)}-1\right)}{\left(\frac{g(x)}{g(y)}-1\right)}-L\right|<\frac{\varepsilon}{2}$$

where for all y large enough, both $\frac{f(x)}{f(y)}-1$ and $\frac{g(x)}{g(y)}-1$ are not equal to zero. Continuing to rewrite the above inequality yields

$$\left|\frac{f(y)}{g(y)}-L\frac{\left(\frac{g(x)}{g(y)}-1\right)}{\left(\frac{f(x)}{f(y)}-1\right)}\right|<\frac{\varepsilon}{2}\left|\frac{\left(\frac{g(x)}{g(y)}-1\right)}{\left(\frac{f(x)}{f(y)}-1\right)}\right|.$$

Therefore, for y large enough,

$$\left|\frac{f(y)}{g(y)}-L\right|\leq\left|L-L\frac{\left(\frac{g(x)}{g(y)}-1\right)}{\left(\frac{f(x)}{f(y)}-1\right)}\right|+\frac{\varepsilon}{2}\left|\frac{\left(\frac{g(x)}{g(y)}-1\right)}{\left(\frac{f(x)}{f(y)}-1\right)}\right|<\varepsilon$$

due to the assumption (8.18) which implies

$$\lim_{y\to b-}\frac{\left(\frac{g(x)}{g(y)}-1\right)}{\left(\frac{f(x)}{f(y)}-1\right)}=1.$$

Therefore, whenever y is large enough,

$$\left|\frac{f(y)}{g(y)}-L\right|<\varepsilon$$

and this is what is meant by (8.20). This proves the theorem.

As before, there is no essential difference between the proof in the case where $x\to b-$ and the proof when $x\to a+$. This observation is stated as the next corollary.

Corollary 8.8.9 *Let $[a,b]\subseteq[-\infty,\infty]$ and suppose f,g are functions which satisfy,*

$$\lim_{x\to a+}f(x)=\pm\infty \text{ and } \lim_{x\to a+}g(x)=\pm\infty, \tag{8.21}$$

and f' and g' exist on (a,b) with $g'(x)\neq 0$ on (a,b). Suppose also that

$$\lim_{x\to a+}\frac{f'(x)}{g'(x)}=L. \tag{8.22}$$

Then

$$\lim_{x\to a+}\frac{f(x)}{g(x)}=L. \tag{8.23}$$

Theorems 8.8.1 8.8.8 and Corollaries 8.8.2 and 8.8.9 will be referred to as L'Hôpital's rule from now on. Theorem 8.8.1 and Corollary 8.8.2 involve the notion of indeterminate forms of the form $\frac{0}{0}$. Please do not think any meaning is being assigned to the nonsense expression $\frac{0}{0}$. It is just a symbol to help remember the sort of thing described by Theorem 8.8.1 and Corollary 8.8.2. Theorem 8.8.8 and Corollary 8.8.9 deal with indeterminate forms which are of the form $\frac{\pm\infty}{\infty}$. Again, this is just a symbol which is helpful in remembering the sort of thing being considered. There are other indeterminate forms which can be reduced to these forms just discussed. Don't ever try to assign meaning to such symbols.

Example 8.8.10 *Find* $\lim_{y\to\infty}\left(1+\frac{x}{y}\right)^{y}$.

It is good to first see why this is called an indeterminate form. One might think that as $y\to\infty$, it follows $x/y\to 0$ and so $1+\frac{x}{y}\to 1$. Now 1 raised to anything is 1 and so it would seem this limit should equal 1. On the other hand, if $x>0$, $1+\frac{x}{y}>1$ and a number raised to higher and higher powers should approach ∞. It really isn't clear what this limit should be. It is an indeterminate form which can be described as 1^{∞}. By definition,

$$\left(1+\frac{x}{y}\right)^{y}=\exp\left(y\ln\left(1+\frac{x}{y}\right)\right).$$

Now using L'Hôpital's rule,

$$\begin{aligned}\lim_{y\to\infty} y\ln\left(1+\frac{x}{y}\right) &= \lim_{y\to\infty}\frac{\ln\left(1+\frac{x}{y}\right)}{1/y}\\ &= \lim_{y\to\infty}\frac{\frac{1}{1+(x/y)}\left(-x/y^{2}\right)}{\left(-1/y^{2}\right)}\\ &= \lim_{y\to\infty}\frac{x}{1+(x/y)}=x\end{aligned}$$

Therefore,

$$\lim_{y\to\infty} y\ln\left(1+\frac{x}{y}\right)=x$$

Since exp is continuous, it follows

$$\lim_{y\to\infty}\left(1+\frac{x}{y}\right)^{y}=\lim_{y\to\infty}\exp\left(y\ln\left(1+\frac{x}{y}\right)\right)=e^{x}.$$

8.8.1 Interest Compounded Continuously

Suppose 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 $100\left(1+r\right)=100+100r$. In this the second term is the interest and the first is called the principal. Now you have $100\left(1+r\right)$ in the bank. This becomes the new principal. How much will you have at the end of the second month? By analogy to what was just done it would equal

$$100\left(1+r\right)+100\left(1+r\right)r=100\left(1+r\right)^{2}.$$

In general, the amount you would have at the end of n months is $100\left(1+r\right)^{n}$.

When a bank says they offer 6% compounded monthly, this means r, the rate per payment period equals .06/12. Consider the problem of a rate of r per year and compounding the interest n times a year and letting n increase without bound. This is what is meant by compounding continuously. The interest rate per payment period is then r/n and the number of payment periods after time t years is approximately tn. From the above the amount in the account after t years is

$$P\left(1+\frac{r}{n}\right)^{nt} \tag{8.24}$$

Recall from Example 8.8.10 that $\lim_{y\to\infty}\left(1+\frac{x}{y}\right)^{y}=e^{x}$. The expression in (8.24) can be written as

$$P\left[\left(1+\frac{r}{n}\right)^{n}\right]^{t}$$

and so, taking the limit as $n \to \infty$, you get

$$Pe^{rt} = A.$$

This shows how to compound interest continuously.

Example 8.8.11 *Suppose you have \$100 and you put it in a savings account which pays 6% compounded continuously. How much will you have at the end of 4 years?*

From the above discussion, this would be $100e^{(.06)4} = 127.12$. Thus, in 4 years, you would gain interest of about \$27.

8.9 Exercises

1. Find the limits.

   (a) $\lim_{x\to 0} \frac{3x-4\sin 3x}{\tan 3x}$

   (b) $\lim_{x\to \frac{\pi}{2}-} (\tan x)^{x-(\pi/2)}$

   (c) $\lim_{x\to 1} \frac{\arctan(4x-4)}{\arcsin(4x-4)}$

   (d) $\lim_{x\to 0} \frac{\arctan 3x-3x}{x^3}$

   (e) $\lim_{x\to 0+} \frac{9^{\sec x-1}-1}{3^{\sec x-1}-1}$

   (f) $\lim_{x\to 0} \frac{3x+\sin 4x}{\tan 2x}$

   (g) $\lim_{x\to \pi/2} \frac{\ln(\sin x)}{x-(\pi/2)}$

   (h) $\lim_{x\to 0} \frac{\cosh 2x-1}{x^2}$

   (i) $\lim_{x\to 0} \frac{-\arctan x+x}{x^3}$

   (j) $\lim_{x\to 0} \frac{x^8 \sin \frac{1}{x}}{\sin 3x}$

   (k) $\lim_{x\to \infty} (1+5^x)^{\frac{2}{x}}$

   (l) $\lim_{x\to 0} \frac{-2x+3\sin x}{x}$

   (m) $\lim_{x\to 1} \frac{\ln(\cos(x-1))}{(x-1)^2}$

   (n) $\lim_{x\to 0+} \sin^{\frac{1}{x}} x$

   (o) $\lim_{x\to 0} (\csc 5x - \cot 5x)$

   (p) $\lim_{x\to 0+} \frac{3^{\sin x}-1}{2^{\sin x}-1}$

   (q) $\lim_{x\to 0+} (4x)^{x^2}$

   (r) $\lim_{x\to \infty} \frac{x^{10}}{(1.01)^x}$

   (s) $\lim_{x\to 0} (\cos 4x)^{(1/x^2)}$

2. Find the following limits.

   (a) $\lim_{x\to 0+} \frac{1-\sqrt{\cos 2x}}{\sin^4\left(4\sqrt{x}\right)}$.

   (b) $\lim_{x\to 0} \frac{2^{x^2}-2^{5x}}{\sin\left(\frac{x^2}{5}\right)-\sin(3x)}$.

(c) $\lim_{n\to\infty} n\left(\sqrt[n]{7}-1\right)$.

(d) $\lim_{x\to\infty}\left(\frac{3x+2}{5x-9}\right)^{x^2}$.

(e) $\lim_{x\to\infty}\left(\frac{3x+2}{5x-9}\right)^{1/x}$.

(f) $\lim_{n\to\infty}\left(\cos\frac{2x}{\sqrt{n}}\right)^{n}$.

(g) $\lim_{n\to\infty}\left(\cos\frac{2x}{\sqrt{5n}}\right)^{n}$.

(h) $\lim_{x\to 3}\frac{x^x-27}{x-3}$.

(i) $\lim_{n\to\infty}\cos\left(\pi\frac{\sqrt{4n^2+13n}}{n}\right)$.

(j) $\lim_{x\to\infty}\left(\sqrt[3]{x^3+7x^2}-\sqrt{x^2-11x}\right)$.

(k) $\lim_{x\to\infty}\left(\sqrt[5]{x^5+7x^4}-\sqrt[3]{x^3-11x^2}\right)$.

(l) $\lim_{x\to\infty}\left(\frac{5x^2+7}{2x^2-11}\right)^{\frac{x}{1-x}}$.

(m) $\lim_{x\to\infty}\left(\frac{5x^2+7}{2x^2-11}\right)^{\frac{x\ln x}{1-x}}$.

(n) $\lim_{x\to 0+}\frac{\ln\left(e^{2x^2}+7\sqrt{x}\right)}{\sinh\left(\sqrt{x}\right)}$.

(o) $\lim_{x\to 0+}\frac{\sqrt[7]{x}-\sqrt[5]{x}}{\sqrt[9]{x}-\sqrt[11]{x}}$.

3. Find the following limits.

(a) $\lim_{x\to 0+}(1+3x)^{\cot 2x}$

(b) $\lim_{x\to 0}\frac{\sin x-x}{x^2}=0$

(c) $\lim_{x\to 0}\frac{\sin x-x}{x^3}$

(d) $\lim_{x\to 0}\frac{\tan(\sin x)-\sin(\tan x)}{x^7}$

(e) $\lim_{x\to 0}\frac{\tan(\sin 2x)-\sin(\tan 2x)}{x^7}$

(f) $\lim_{x\to 0}\frac{\sin\left(x^2\right)-\sin^2(x)}{x^4}$

(g) $\lim_{x\to 0}\frac{e^{-\left(1/x^2\right)}}{x}$

(h) $\lim_{x\to 0}\left(\frac{1}{x}-\cot(x)\right)$

(i) $\lim_{x\to 0}\frac{\cos(\sin x)-1}{x^2}$

(j) $\lim_{x\to\infty}\left(x^2\left(4x^4+7\right)^{1/2}-2x^4\right)$

(k) $\lim_{x\to 0}\frac{\cos(x)-\cos(4x)}{\tan(x^2)}$

(l) $\lim_{x\to 0}\frac{\arctan(3x)}{x}$

(m) $\lim_{x\to\infty}\left[\left(x^9+5x^6\right)^{1/3}-x^3\right]$

4. Suppose you want to have \$2000 saved at the end of 5 years. How much money should you place into an account which pays 7% per year compounded continuously?

5. Using a good calculator, find $e^{.06}-\left(1+\frac{.06}{360}\right)^{360}$. Explain why this gives a measure of the difference between compounding continuously and compounding daily.

8.10 Related Rates

Sometimes some variables are related by a formula and it is known how fast all are changing but one. The related rates problem asks for how fast the remaining variable is changing.

Example 8.10.1 *A cube of ice is melting such that* $\frac{dV}{dt} = -4cm^3/\sec$ *where* V *is the volume. How fast are the sides changing when they are equal to 5 centimeters in length?*

The volume is $V = x^3$ where x is the length of a side of the cube. Therefore, the chain rule implies

$$-4 = \frac{dV}{dt} = 3x^2 \frac{dx}{dt}$$

and the problem is to find $\frac{dx}{dt}$ when $x = 5$. Therefore,

$$\frac{dx}{dt} = \frac{-4}{3\,(25)} = \frac{-4}{75}\text{cm/second}$$

at this time.

Note there is no way of knowing the volume or the sides as a functions of t.

Example 8.10.2 *One car travels north at 70 miles per hour and the other travels east at 60 miles per hour toward an intersection. How fast is the distance between the two cars changing when this distance equals five miles and the car heading north is at a distance of three miles from the intersection?*

Let l denote the distance between the cars. Thus if x is the distance from the intersection of the car traveling east and y is the distance from the intersection of the car traveling north, $l^2 = x^2 + y^2$. When $y = 3$, it follows that $x = 4$. Therefore, at the instant described

$$\begin{aligned} 2ll' &= 2xx' + 2yy' \\ 10l' &= 8\,(-60) + 6\,(-70) \end{aligned}$$

and so $l' = -90$ miles per hour at this instant.

8.11 Exercises

1. One car travels north at 70 miles per hour toward an intersection and the other travels east at 60 miles per hour away from the intersection. How fast is the distance between the two cars changing when this distance equals five miles and the car heading north is at a distance of three miles from the intersection?

2. A trash compactor compacts some trash which is in the shape of a box having a square base and a height equal to twice the length of a side of the base. Suppose each side of the base is changing at the rate of -3 inches per second. How fast is the volume changing when the side of the base equals 10 inches?

3. An isosceles triangle has two sides of equal length. Imagine such a triangle in which the two legs have length 8 inches and denote the included angle by θ and the area by A. Suppose $\frac{dA}{dt} = \sqrt{3}$ square inches per minute. How fast is θ changing when $\theta = \pi/6$ radians?

4. A point having coordinates (x, y) moves over the ellipse, $\frac{x^2}{4} + \frac{y^2}{9} = 1$. If $\frac{dy}{dt} = 2$, find $\frac{dx}{dt}$ at the point $(2, 3)$.

5. A spectator at a tennis tournament sits 10 feet from the end of the net and on the line determined by the net. He watches the ball go back and forth, and will have a very sore neck when he wakes up the next morning. How fast is the angle between his line of sight and the line determined by the net changing when the ball crosses over the net at a point 12 feet from the end assuming the ball travels at a speed of 60 miles per hour?

6. The surface area of a sphere of radius r equals $4\pi r^2$ and the volume of the ball of radius r equals $(4/3)\pi r^3$. A balloon in the shape of a ball is being inflated at the rate of 6 cubic inches per minute. How fast is the surface area changing when the volume of the ball equals 20π cubic inches?

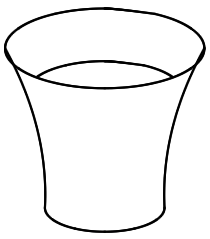
7. A mother cheetah attempts to fix dinner, a Thompson gazelle, for her hungry children. She moves at 100 feet per second while dinner travels at 80 feet per second. How fast is the distance between her and dinner decreasing when she is located at the point $(0, 40)$ feet and dinner is moving in the direction of the positive x axis at the point $(30, 0)$ feet? Is your answer a little surprising? **Hint:** Let (x, y) denote the coordinates of the cheetah and let $(z, 0)$ denote the coordinates of the gazelle. Then if l is the desired distance, $l^2 = (x - z)^2 + y^2$. At the instant described, assuming the cheetah moves toward the gazelle at all times, $y'/x' = -4/3$ (why?) and also $\sqrt{(x')^2 + (y')^2} = 100$.

8. A six foot high man walks at a speed of 5 feet per second away from a light post which is 12 feet high that has the light right on the top. How fast is the end of his shadow moving when he is at a distance of 10 feet from the base of the light pole. How fast is the end of the shadow moving when he is 5 feet from the pole? (Assume he does not walk normally but instead oozes along like a giant amoeba so that his head is always exactly 6 feet above the ground.)

9. The volume of a right circular cone is $\frac{1}{3}\pi r^2 h$. Grain comes off a conveyor belt and falls to the ground making a right circular cone. It is observed that $r'(t) = .5$ feet per minute and $h'(t) = .3$ feet per minute. It is also known that the rate at which the grain falls off the conveyor belt is 100π cubic feet per minute. When the radius of the cone is 10 feet what is the height of the cone?

10. A hemispherical dish of radius 5 inches is sitting on a table. Soup is being poured in at the constant rate of 4 cubic inches per second. How fast is the level of soup rising when the radius of the top surface of the soup equals 3 inches? The volume of soup at depth y will be shown later to equal $V(y) = \pi\left(5y^2 - \frac{y^3}{3}\right)$.

11. A vase of water is sitting on a table.



It will be shown later that if $V(y)$ is the total volume of the vase up to height y, then $\frac{dV}{dy} = A(y)$ where $A(y)$ is the surface area of the top surface of the water at this height. (To see this is very reasonable, note that a little chunk of volume of the vase between heights y and $y + dy$ would be $dV = A(y)\,dy$, area times height.) Also, the

rate at which the water evaporates is proportional to the surface area of the exposed water. Thus $\frac{dV}{dt} = -kA(y)$. Show $\frac{dy}{dt}$ is a constant even though the surface area of the exposed water is constantly changing in a typical vase.

12. A revolving search light at a prison makes one revolution per minute. How fast is the light travelling along the nearest point on a wall 1/4 mile away? Give your answer in miles per hour.

13. A painter is on top of a 13 foot ladder which leans against a house. The base of the ladder is moving away from the house at the rate of 2 feet per second causing the top of the ladder to move down the house. How fast is the painter descending when the base of the ladder is at a distance of 5 feet from the house?

14. A rope fastened to the bow of a row boat has the other end wound around a windlass which is 4 feet above the level of the bow of the boat. The current pulls the boat away at the rate of 2 feet per second. How fast is the rope unwinding when the distance between the bow of the boat and the windlass is 10 feet?

15. A kite 100 feet above the ground is being blown away from the person holding its string in a direction parallel to the ground at the rate of 10 feet per second. At what rate must the string be let out when the length of string already let out is 200 feet?

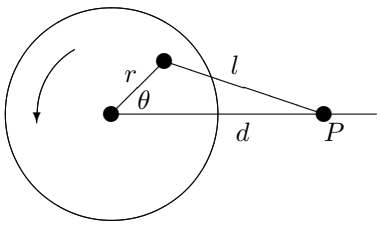
16. A certain volume of an ideal gas satisfies $PV = kT$ where T is the absolute temperature, P is the pressure, V is the volume and k is a constant which depends on the amount of the gas and the sort of gas in the sample. Find a formula for $\frac{dV}{dt}$ in terms of k, P, V and their derivatives.

17. A disposable cup is made in the shape of a right circular cone with height 5 inches and radius 2 inches. Water flows in to this conical cup at the rate of 4 cubic inches per minute. How fast is the water level rising when the water in the cone is three inches deep? The volume of a cone is $\frac{1}{3}\pi r^2 h$.

18. The two equal sides of an isosceles triangle have length x inches and the third leg has length y inches. Suppose $\frac{dx}{dt} = 2$ inches per minute and that the length of the other side changes in such a way that the area of the triangle is always 10 square inches. For θ the angle between the two equal sides, find $\frac{d\theta}{dt}$ when $x = 5$ inches.

19. A object is moving over the ellipse whose equation is $\frac{x^2}{9} + \frac{y^2}{4} = 1$. Near the point $\left(\frac{3\sqrt{2}}{2}, \sqrt{2}\right)$, it is observed that x is changing at the rate of 1 unit per second. How fast is y changing when the object is at this point?

20. Consider the following diagram.



This represents a wheel which is spinning at a constant angular velocity equal to ω. Thus $\frac{d\theta}{dt} = \omega$. The circles represent axels and the lines joining the circles represent rigid bars. The circle on the far right is allowed to slide on the indicated line. Thus,

as the wheel turns, the point, P moves back and forth on the horizontal line shown. Find the velocity of P when $\theta = \pi/4$ in terms of r, l, and ω. **Hint:** Remember the law of sines.

8.12 The Derivative And Optimization

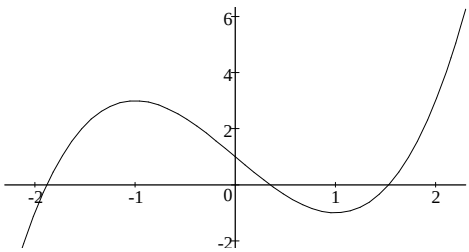
There are existence theorems such as Theorem 5.7.10 on Page 107 which ensure a maximum or minimum of a function exists but in this section the goal is to give ways to find the maximum or minimum values of a function.

Suppose f is continuous on $[a,b]$. The minimum or maximum could occur at either end point or it could occur at a point in the open interval, (a,b). If it occurs at a point of the open interval, (a,b), say at x_0, and if $f'(x_0)$ exists, then from Theorem 6.5.2 on Page 140 $f'(x_0) = 0$. Therefore, the following simple procedure can be used to locate the maximum or minimum of a function, f. Find all points, x, in (a,b) where $f'(x) = 0$ and all points, x, in (a,b) where $f'(x)$ does not exist. Then consider these points along with the end points of the interval. Evaluate f at the end points and at these points where the derivative is zero or does not exist. The largest must be the maximum value of f on the interval, $[a,b]$, and the smallest must be the minimum value of f on the interval, $[a,b]$. Typically, this involves checking only finitely many points.

Sometimes there are no end points. In this case, you do not necessarily know a maximum value or a minimum value of a continuous function even exists. However, if the function is differentiable and if a maximum or minimum exists, it can still be found by looking at the points where the derivative equals zero.

Example 8.12.1 *Find the maximum and minimum values of the function,* $f(x) = x^3 - 3x + 1$ *on the interval* $[-2,2]$.

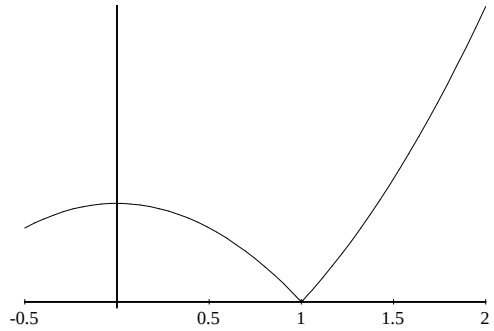
The points where $f'(x) = 3x^2 - 3 = 0$ are $x = 1$ or -1. There are no points where the derivative does not exist. Therefore, evaluate the function at $-1, -2, 2$, and 1. Thus $f(-1) = 3, f(1) = -1, f(-2) = -1$, and $f(2) = 3$. Therefore, the maximum value of the function occurs at the point -1 and 2 and has the value of 3 while the minimum value of the function occurs at -1 and -2 and equals -1. The following is a graph of this function.



Example 8.12.2 *Find the maximum and minimum values of the function* $f(x) = \left|x^2 - 1\right|$ *on the interval* $[-.5,2]$.

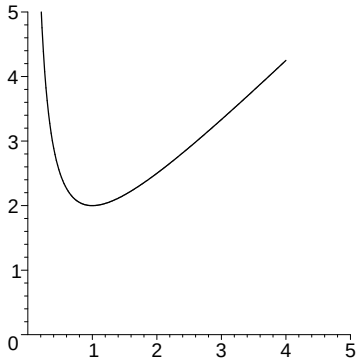
You should verify that this function fails to have a derivative at the point $x = 1$. For $x \in (-.5,1)$ the function equals $1 - x^2$ and so its derivative equals zero at $x = 0$. For $x > 1$, the function equals $x^2 - 1$ and so there is no point larger than 1 where the derivative equals zero. Therefore, the points to look at are the end points, $-.5, 2$, the points where the derivative fails to exist, 1 and the point where the derivative equals zero, $x = 0$. Now $f(-.5) = .75, f(0) = 1, f(1) = 0$, and $f(2) = 3$. It follows the function achieves its

maximum at the end point, $x = 2$ and its minimum at the point, $x = 1$ where the derivative fails to exist. The following is a graph of this function.



Example 8.12.3 *Find the minimum value of the function* $f(x) = x + \frac{1}{x}$ *for* $x \in (0, \infty)$.

The graph of this function is given below.



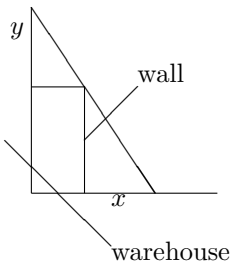
From the graph, it seems there should exist a minimum value at the bottom of the graph. To find it, take the derivative of f and set it equal to zero and then solve for the value of x. Thus

$$1 - \frac{1}{x^2} = 0$$

and so $x = 1$. The solution to the equation, $x = -1$ is of no interest because it is not greater than zero. Therefore, the minimum value of the function on the interval, $(0, \infty)$ equals $f(1) = 2$ as suggested by the graph.

Example 8.12.4 *An eight foot high wall stands one foot from a warehouse. What is the length of the shortest ladder which extends from the ground to the warehouse.*

A diagram of this situation is the following picture.



In this picture, the slanted line represents the ladder and x and y are as shown. By

similar triangles, $y/1 = 8/x$. Therefore, $xy = 8$. From the Pythagorean theorem the length of the ladder equals $\sqrt{1+y^2} + \sqrt{x^2+64}$. Now using the relation between x and y, the function of a single variable, x, to minimize is

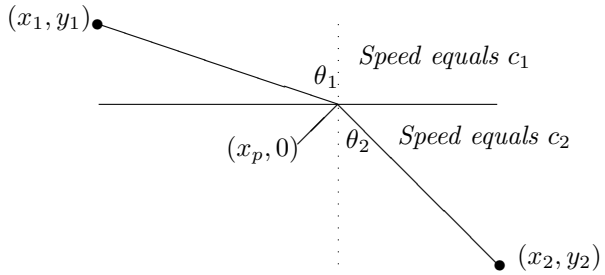
$$f(x) = \sqrt{1 + \frac{64}{x^2}} + \sqrt{x^2+64},$$

Clearly $x > 0$ so there are no endpoints to worry about. (Why?) Also the function is differentiable and so it suffices to consider only points where the derivative equals zero. This is a little messy but finally

$$f'(x) = \frac{1}{\sqrt{(x^2+64)}} \frac{-64+x^3}{x^2}$$

and the value where this equals zero is $x = 4$. It follows the shortest ladder is of length $\sqrt{1+\frac{64}{4^2}} + \sqrt{4^2+64} = 5\sqrt{5}$ feet.

Example 8.12.5 *Fermat's principle says that light travels on a path which will minimize the total time. Consider the following picture of light passing from (x_1, y_1) to (x_2, y_2) as shown. The angle θ_1 is called the angle of incidence while the angle, θ_2 is called the angle of refraction. The picture indicates a situation in which $c_1 > c_2$.*



Then define by x the quantity $x_p - x_1$. What is the relation between θ_1 and θ_2?

The time it takes for the light to go from (x_1, y_1) to the point $(x_p, 0)$ equals $\sqrt{x^2+y_1^2}/c_1$ and the time it takes to go from $(x_p, 0)$ to (x_2, y_2) is $\sqrt{(x_2-x_1-x)^2+y_2^2}/c_2$. Therefore, the total time is

$$T = \frac{\sqrt{x^2+y_1^2}}{c_1} + \frac{\sqrt{(x_2-x_1-x)^2+y_2^2}}{c_2}$$

Thus T is minimized if

$$\begin{aligned}
\frac{dT}{dx} &= \frac{d}{dx}\left(\frac{\sqrt{x^2+y_1^2}}{c_1} + \frac{\sqrt{(x_2-x_1-x)^2+y_2^2}}{c_2}\right) \\
&= \frac{x}{c_1\sqrt{x^2+y_1^2}} - \frac{(x_2-x_1-x)}{c_2\sqrt{(x_2-x_1-x)^2+y_2^2}} \\
&= \frac{\sin(\theta_1)}{c_1} - \frac{\sin(\theta_2)}{c_2} = 0
\end{aligned}$$

at this point. Therefore, this yields the desired relation between θ_1 and θ_2. This is called Snell's law.

8.13 Exercises

1. Find the maximum and minimum values for the following functions defined on the given intervals.

   (a) $x^3 - 3x^2 + x - 7$, $[0, 4]$

   (b) $\ln\left(x^2 - x + 2\right)$, $[0, 2]$

   (c) $x^3 + 3x$, $[-1, 10]$

   (d) $\frac{x^2+1+3x^3}{3x^2+5}$, $[-1, 1]$

   (e) $\sin\left(x^3 - x\right)$, $[-1, 1]$

   (f) $x^2 - x\tan x$, $[-1, 1]$

   (g) $1 - 2x^2 + x^4$, $[-2, 2]$

   (h) $\ln\left(2 - 2x^2 + x^4\right)$, $[-1, 2]$

   (i) $x^2 + 4x - 8$, $[-4, 2]$

   (j) $x^2 - 3x + 6$, $[-2, 4]$

   (k) $-x^2 + 3x$, $[-4, 2]$

   (l) $x + \frac{1}{x}$, $(0, \infty)$

2. A cylindrical can is to be constructed to hold 30 cubic inches. The top and bottom of the can are constructed of a material costing one cent per square inch and the sides are constructed of a material costing 2 cents per square inch. Find the minimum cost for such a can.

3. Two positive numbers sum to 8. Find the numbers if their product is to be as large as possible.

4. The ordered pair, (x, y) is on the ellipse, $x^2 + 4y^2 = 4$. Form the rectangle which has (x, y) as one end of a diagonal and $(0, 0)$ at the other end. Find the rectangle of this sort which has the largest possible area.

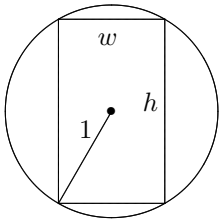
5. A rectangle is inscribed in a circle of radius r. Find the formula for the rectangle of this sort which has the largest possible area.

6. A point is picked on the ellipse, $x^2 + 4y^2 = 4$ which is in the first quadrant. Then a line tangent to this point is drawn which intersects the x axis at a point, x_1 and the y axis at the point y_1. The area of the triangle formed by the y axis, the x axis, and the line just drawn is thus $\frac{x_1 y_1}{2}$. Out of all possible triangles formed in this way, find the one with smallest area.

7. Find maximum and minimum values if they exist for the function, $f(x) = \frac{\ln x}{x}$ for $x > 0$.

8. Describe how you would find the maximum value of the function, $f(x) = \frac{\ln x}{2+\sin x}$ for $x \in (0, 6)$ if it exists. **Hint:** You might want to use a calculator to graph this and get an idea what is going on.

9. A rectangular beam of height h and width w is to be sawed from a circular log of radius 1 foot. Find the dimensions of the strongest such beam assuming the strength is of the form kh^2w. Here k is some constant which depends on the type of wood used.



10. A farmer has 600 feet of fence with which to enclose a rectangular piece of land that borders a river. If he can use the river as one side, what is the largest area that he can enclose.

11. An open box is to be made by cutting out little squares at the corners of a rectangular piece of cardboard which is 20 inches wide and 40 inches long. and then folding up the rectangular tabs which result. What is the largest possible volume which can be obtained?

12. A feeding trough is to be made from a rectangular piece of metal which is 3 feet wide and 12 feet long by folding up two rectangular pieces of dimension one foot by 12 feet. What is the best angle for this fold?

13. Find the dimensions of the right circular cone which has the smallest area given the volume is 30π cubic inches. The volume of the right circular cone is $(1/3)\,\pi r^2 h$ and the area of the cone is $\pi r\sqrt{h^2+r^2}$.

14. A wire of length 10 inches is cut into two pieces, one of length x and the other of length $10-x$. One piece is bent into the shape of a square and the other piece is bent into the shape of a circle. Find the two lengths such that the sum of the areas of the circle and the square is as large as possible. What are the lengths if the sum of the two areas is to be as small as possible.

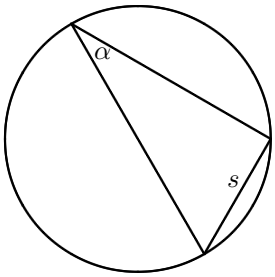
15. A hiker begins to walk to a cabin in a dense forest. He is walking on a road which runs from East to West and the cabin is located exactly one mile north of a point two miles down the road. He walks 5 miles per hour on the road but only 3 miles per hour in the woods. Find the path which will minimize the time it takes for him to get to the cabin.

16. A refinery is on a straight shore line. Oil needs to flow from a mooring place for oil tankers to this refinery. Suppose the mooring place is two miles off shore from a point on the shore 8 miles away from the refinery and that it costs five times as much to lay pipe under water than above the ground. Describe the most economical route for a pipeline from the mooring place to the refinery.

17. Two hallways, one 5 feet wide and the other 6 feet wide meet. It is desired to carry a ladder horizontally around the corner. What is the longest ladder which can be carried in this way? **Hint:** Consider a line through the inside corner which extends to the opposite walls. The shortest such line will be the length of the longest ladder. You might also consider Example 7.1.3 on Page 155.

18. A triangle is inscribed in a circle in such a way that one side of the triangle is always the same length, s. Show that out of all such triangles the maximum area is obtained when the triangle is an isosceles triangle. **Hint:** From theorems in plane geometry, the angle opposite the side having fixed length is a constant, no matter how you draw

the triangle. (See Problem 10 on Page 72 to see why this is if the geometry interests you.) Use the law of sines and this fact. In the following picture, α is a constant.



19. A window is to be constructed for the wall of a church which is to consist of a rectangle of height b surmounted by a half circle of radius a. Suppose the total perimeter of the window is to be no more than $4\pi + 8$ feet. Find the shape and dimensions of the window which will admit the most light.

20. You know $\lim_{x\to\infty} \ln x = \infty$. Show that if $\alpha > 0$, then $\lim_{x\to\infty} \frac{\ln x}{x^\alpha} = 0$.

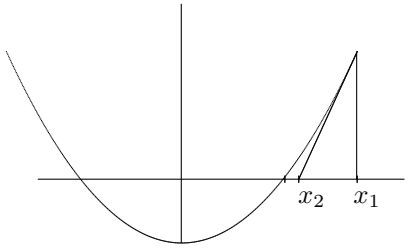
21. Suppose p and q are two positive numbers larger than 1 which satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Now let a and b be two positive numbers and consider $f(t) = \frac{1}{p}(at)^p + \frac{1}{q}\left(\frac{b}{t}\right)^q$ for $t > 0$. Show the minimum value of f is ab. Prove the important inequality, $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$.

22. Using Problem 21 establish the following magnificent inequality which is a case of Holder's inequality. For $\frac{1}{p} + \frac{1}{q} = 1$, and a_i, b_i positive numbers,

$$\sum_{i=1}^{n} a_i b_i \leq \left(\sum_{i=1}^{n} a_i^p\right)^{1/p} \left(\sum_{i=1}^{n} b_i^q\right)^{1/q}.$$

8.14 The Newton Raphson Method

The Newton Raphson method is a way to get approximations of solutions to various equations. For example, suppose you want to find $\sqrt{2}$. The existence of $\sqrt{2}$ is not difficult to establish by considering the continuous function, $f(x) = x^2 - 2$ which is negative at $x = 0$ and positive at $x = 2$. Therefore, by the intermediate value theorem, there exists $x \in (0, 2)$ such that $f(x) = 0$ and this x must equal $\sqrt{2}$. The problem consists of how to find this number, not just to prove it exists. The following picture illustrates the procedure of the Newton Raphson method.



In this picture, a first approximation, denoted in the picture as x_1 is chosen and then the tangent line to the curve $y = f(x)$ at the point $(x_1, f(x_1))$ is obtained. The equation of this tangent line is

$$y - f(x_1) = f'(x_1)(x - x_1).$$

Then extend this tangent line to find where it intersects the x axis. In other words, set $y = 0$ and solve for x. This value of x is denoted by x_2. Thus

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

This second point, x_2 is the second approximation and the same process is done for x_2 that was done for x_1 in order to get the third approximation, x_3. Thus

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}.$$

Continuing this way, yields a sequence of points, $\{x_n\}$ given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{8.25}$$

which hopefully has the property that $\lim_{n\to\infty} x_n = x$ where $f(x) = 0$. You can see from the above picture that this must work out in the case of $f(x) = x^2 - 2$.

Now carry out the computations in the above case for $x_1 = 2$ and $f(x) = x^2 - 2$. From (8.25),

$$x_2 = 2 - \frac{2}{4} = 1.5.$$

Then

$$x_3 = 1.5 - \frac{(1.5)^2 - 2}{2(1.5)} \leq 1.417,$$

$$x_4 = 1.417 - \frac{(1.417)^2 - 2}{2(1.417)} = 1.414\,216\,302\,046\,577,$$

What is the true value of $\sqrt{2}$? To several decimal places this is $\sqrt{2} = 1.414\,213\,562\,373\,095$, showing that the Newton Raphson method has yielded a very good approximation after only a few iterations, even starting with an initial approximation, 2, which was not very good.

This method does not always work. For example, suppose you wanted to find the solution to $f(x) = 0$ where $f(x) = x^{1/3}$. You should check that the sequence of iterates which results does not converge. This is because, starting with x_1 the above procedure yields $x_2 = -2x_1$ and so as the iteration continues, the sequence oscillates between positive and negative values as its absolute value gets larger and larger.

However, if $f(x_0) = 0$ and $f''(x) > 0$ for x near x_0, you can draw a picture to show that the method will yield a sequence which converges to x_0 provided the first approximation, x_1 is taken sufficiently close to x_0. Similarly, if $f''(x) < 0$ for x near x_0, then the method produces a sequence which converges to x_0 provided x_1 is close enough to x_0.

8.15 Exercises

1. By drawing representative pictures, show convergence of the Newton Raphson method in the cases described above where $f''(x) > 0$ near x_0 or $f''(x) < 0$ near x_0.

2. Draw some graphs to illustrate the Newton Raphson method does not yield a convergent sequence in the case where $f(x) = x^{1/3}$.

3. Use the Newton Raphson method to approximate the first positive solution of $x - \tan x = 0$. **Hint:** You may need to use a calculator to deal with $\tan x$.

4. Use the Newton Raphson method to compute an approximation to $\sqrt{3}$ which is within 10^{-6} of the true value. Explain how you know you are this close.

5. Using the Newton Raphson method and an appropriate picture, discuss the convergence of the recursively defined sequence $x_{n+1} = \left((p-1)\, x_n + c x_n^{1-p}\right)/p$ where $x_1, c > 0$ and $p > 1$.

Antiderivatives And Differential Equations

9.0.1 Outcomes

1. Understand and describe the concept of an initial value problem.
2. Understand and use the methods of substitution, integration by parts, trig. substitutions and partial fractions to find solutions to initial value problems.
3. Find areas between curves.
4. Find volumes using the method of cross sections and shells.
5. Find lengths of curves.
6. Find areas of surfaces of revolution.
7. Understand and use methods for solving linear and separable differential equations and understand some examples of these.
8. Understand the concept of fluid pressure of an incompressible fluid and the force it produces.
9. Be able to do simple problems involving work.

A differential equation is an equation which involves an unknown function and its derivatives. Differential equations are the unifying idea in this chapter. Many interesting problems may be solved by formulating them as solutions of a suitable differential equation with initial condition called an initial value problem.

9.1 Initial Value Problems

The initial value problem is to find a function, $y(x)$ for $x \in [a,b]$ such that

$$y'(x) = f(x, y(x)),\ y(a) = y_0.$$

Various assumptions are made on $f(x,y)$. At this time it is assumed that f does not depend on y and f is a given continuous defined $[a,b]$. Thus the initial value problem of interest here is one of the form

$$y'(x) = f(x), y(a) = y_0. \tag{9.1}$$

As an example of an application of an initial value problem, recall the discussion which led to the derivative on Page 131. There $r(t)$ was the x coordinate of a point moving on the x axis at time t and it was shown there that the velocity of this object was $r'(t)$. When this is positive, the object is moving to the right on the x axis and when it is negative, the object is moving to the left. Suppose $r'(t)$ was known, say $r'(t) = f(t)$ along with $r(0) = x_0$ and you wanted to find $r(t)$. Then you are really asking for the solution to an initial value problem of the form

$$r'(t) = f(t),\ r(0) = x_0.$$

Theorem 9.1.1 *There is at most one solution to the initial value problem, (9.1) which is continuous on* $[a,b]$.

Proof: Suppose both $A(x)$ and $B(x)$ are solutions to this initial value problem for $x \in (a,b)$. Then letting $H(x) \equiv A(x) - B(x)$, it follows that $H(a) = 0$ and $H'(x) = A'(x) - B'(x) = f(x) - f(x) = 0$. Therefore, from Corollary 6.8.4 on Page 147, it follows $H(x)$ equals a constant on (a,b). By continuity of H, this constant must equal $H(a) = 0$.

The main difficulty in solving these initial value problems like (9.1) is in finding a function whose derivative equals the given function. This is in general a very hard problem, although techniques for doing this are presented later which will cover many cases of interest. The functions whose derivatives equal a given function, $f(x)$, are called antiderivatives and there is a special notation used to denote them.

Definition 9.1.2 *Let f be a function. $\int f(x)\,dx$ denotes the set of antiderivatives of f. Thus $F \in \int f(x)\,dx$ means $F'(x) = f(x)$. It is customary to refer to $f(x)$ as the integrand. This symbol is also called the indefinite integral and sometimes is referred to as an integral although this last usage is not correct.*

The reason this last usage is not correct is that the integral of a function is a single number not a whole set of functions. Nevertheless, you can't escape the fact that it is common usage to call that symbol an integral.

Lemma 9.1.3 *Suppose $F,\ G \in \int f(x)\ dx$ for $x \in (a,b)$. Then there exists a constant, C such that for all $x \in (a,b)$, $F(x) = G(x) + C$.*

Proof:

$$F'(x) - G'(x) = f(x) - f(x) = 0$$

for all $x \in (a,b)$. Consequently, by Corollary 6.8.4 on Page 147, $F(x) - G(x) = C$. This proves the lemma.

There is another simple lemma about antiderivatives.

Lemma 9.1.4 *If a and b are nonzero real numbers, and if $\int f(x)\ dx$ and $\int g(x)\ dx$ are nonempty, then*

$$\int (af(x) + bg(x))\ dx = a\int f(x)\ dx + b\int g(x)\ dx$$

Proof: The symbols on the two sides of the equation denote sets of functions. It is necessary to verify the two sets of functions are the same. Suppose then that $F \in \int f(x)\ dx$ and $G \in \int g(x)\ dx$. Then $aF + bG$ is a typical function of the right side of the equation. Taking the derivative of this function, yields $af(x) + bg(x)$ and so this shows the set of functions on the right side is a subset of the set of functions on the left.

Now take $H \in \int (af(x) + bg(x))\, dx$ and pick $F \in \int f(x)\, dx$. Then $aF \in a\int f(x)\, dx$ and

$$(H(x) - aF(x))' = af(x) + bg(x) - af(x) = bg(x)$$

showing that

$$H - aF \in \int bg(x)\, dx = b\int g(x)\, dx$$

because $b \neq 0$. Therefore

$$H \in aF + b\int g(x)\, dx \subseteq a\int f(x)\, dx + b\int g(x)\, dx.$$

This has shown the two sets of functions are the same and proves the lemma.

From Lemma 9.1.3 it follows that if $F(x) \in \int f(x)\, dx$, then every other function in $\int f(x)\, dx$ is of the form $F(x) + C$ for a suitable constant, C. Thus it is customary to write

$$\int f(x)\, dx = F(x) + C$$

where it is understood that C is an arbitrary constant, called a constant of integration.

From the formulas for derivatives presented earlier, the following table of antiderivatives follows.

$f(x)$	$\int f(x)\, dx$
$x^n, n \neq -1$	$\frac{x^n}{n+1} + C$
x^{-1}	$\ln\lvert x\rvert + C$
$\cos(x)$	$\sin(x) + C$
$\sin(x)$	$-\cos(x) + C$
$\sec^2(x)$	$\tan(x) + C$
e^x	$e^x + C$
$\cosh(x)$	$\sinh(x) + C$
$\sinh(x)$	$\cosh(x) + C$
$\frac{1}{\sqrt{1-x^2}}$	$\arcsin(x) + C$
$\frac{1}{\sqrt{1+x^2}}$	$\operatorname{arcsinh} x + C$

The above table is a good starting point for other antiderivatives. For example,

Proposition 9.1.5 *Let $\sum_{k=0}^n a_k x^k$ be a polynomial. Then*

$$\int \sum_{k=0}^n a_k x^k dx = \sum_{k=0}^n a_k \frac{x^{k+1}}{k+1} + C$$

Proof: This follows from the above table and Lemma 9.1.4.

Example 9.1.6 *Suppose the velocity of an object moving on the x axis is given by the function $\cos(t)$ and when $t = 0$, the object is at 0. Find the position of the object.*

As explained above, you need to solve

$$r'(t) = \cos(t),\ r(0) = 0.$$

To do this, note that from the differential equation, $r(t) = \sin(t) + C$ and it only remains to find the constant, C such that $r(0) = 0$. Thus $0 = r(0) = \sin(0) + C = C$ and so $r(t) = \sin(t)$.

Example 9.1.7 *Consider the same problem as in Example 9.1.6 but this time let* $r(1) = 5$. *What is the answer in this case?*

From the differential equation it is still the case that $r(t) = \sin(t) + C$ but now C needs to be chosen such that $r(1) = 5$. Thus $5 = r(1) = \sin(1) + C$ and so $C = 5 - \sin(1)$. Therefore, the solution to the initial value problem,

$$r'(t) = \cos(t),\ r(1) = 5$$

is $r(t) = \sin(t) + (5 - \sin(1))$.

They are all like this. You find an antiderivative. The answer you want is this antiderivative added to an appropriate constant chosen to satisfy the given initial condition. The problem is in finding the antiderivative. In general, this is a very hard problem but there are techniques for solving it in some cases. The next section is one such techniques.

9.2 The Method Of Substitution

The method of substitution is based on the following formula which is merely a restatement of the chain rule.

$$\int f(g(x))\, g'(x)\, dx = F(x) + C, \tag{9.2}$$

where $F'(y) = f(y)$. Here are some examples of the method of substitution.

Example 9.2.1 *Find* $\int \sin(x)\cos(x)\sqrt{3 + 2^{-1}\sin^2(x)}\, dx$

Note it is of the form given in (9.2) with $g(x) = 2^{-1}\sin^2(x)$ and $F(u) = \frac{2}{3}\left(\sqrt{(3+u)}\right)^3$. Therefore,

$$\int \sin(x)\cos(x)\sqrt{3 + 2^{-1}\sin^2(x)}\, dx = \frac{1}{12}\left(\sqrt{\left(12 + 2\sin^2 x\right)}\right)^3 + C.$$

Example 9.2.2 *Find* $\int \left(1 + x^2\right)^6 2x\, dx$.

This is a special case of (9.2) when $g(x) = 1 + x^2$ and $F(u) = u^7/7$. Therefore, the answer is

$$\int \left(1 + x^2\right)^6 2x\, dx = \left(1 + x^2\right)^7/7 + C.$$

Example 9.2.3 *Find* $\int \left(1 + x^2\right)^6 x\, dx$

This equals

$$\frac{1}{2}\int \left(1 + x^2\right)^6 2x\, dx = \frac{1}{2}\frac{\left(1 + x^2\right)^7}{7} + C = \frac{\left(1 + x^2\right)^7}{14} + C.$$

Actually, it is not necessary to recall (9.2) and massage things to get them in that form. There is a trick based on the Leibniz notation for the derivative which is very useful and illustrated in the following example.

Example 9.2.4 *Find* $\int \cos(2x)\sin^2(2x)\, dx$.

Let $u = \sin(2x)$. Then $\frac{du}{dx} = 2\cos(2x)$. Now formally

$$\frac{du}{2} = \cos(2x)\,dx.$$

Thus

$$\begin{aligned}\int \cos(2x)\sin^2(2x)\,dx &= \frac{1}{2}\int u^2\,du = \frac{u^3}{6} + C \\ &= \frac{(\sin(2x))^3}{6} + C\end{aligned}$$

The expression $(1/2)\,du$ replaced the expression $\cos(2x)\,dx$ which occurs in the original problem and the resulting problem in terms of u was much easier. This was solved and finally the original variable was replaced. When using this method, it is a good idea to check your answer to be sure you have not made a mistake. Thus in this example, the chain rule implies $\left(\frac{(\sin(2x))^3}{6}\right)' = \cos(2x)\sin^2(2x)$ which verifies the answer is right. Here is another example.

Example 9.2.5 *Find* $\int \sqrt[3]{2x+7}x\,dx$.

In this example $u = 2x + 7$ so that $du = 2dx$. Then

$$\begin{aligned}\int \sqrt[3]{2x+7}x\,dx &= \int \sqrt[3]{u}\overbrace{\frac{u-7}{2}}^{x}\overbrace{\frac{1}{2}\,du}^{dx} \\ &= \int \left(\frac{1}{4}u^{4/3} - \frac{7}{4}u^{1/3}\right)du \\ &= \frac{3}{28}u^{7/3} - \frac{21}{16}u^{4/3} + C \\ &= \frac{3}{28}(2x+7)^{7/3} - \frac{21}{16}(2x+7)^{4/3} + C\end{aligned}$$

Example 9.2.6 *Find* $\int x3^{x^2}\,dx$

Let $u = 3^{x^2}$ so that $\frac{du}{dx} = 2x\ln(3)\,3^{x^2}$ and $\frac{du}{2\ln(3)} = x3^{x^2}dx$. Thus

$$\begin{aligned}\int x3^{x^2}\,dx &= \frac{1}{2\ln(3)}\int du = \frac{1}{2\ln(3)}\left[u + C\right] \\ &= \frac{1}{2\ln(3)}3^{x^2} + \left(\frac{1}{2\ln(3)}\right)C\end{aligned}$$

Since the constant is an arbitrary constant, this is written as

$$\frac{1}{2\ln(3)}3^{x^2} + C.$$

Example 9.2.7 *Find* $\int \cos^2(x)\,dx$

Recall that $\cos(2x) = \cos^2(x) - \sin^2(x)$ and $1 = \cos^2(x) + \sin^2(x)$. Then subtracting and solving for $\cos^2(x)$,

$$\cos^2(x) = \frac{1 + \cos(2x)}{2}.$$

Therefore,

$$\int \cos^2(x)\,dx = \int \frac{1+\cos(2x)}{2}\,dx$$

Now letting $u = 2x$, $du = 2dx$ and so

$$\begin{aligned}\int \cos^2(x)\,dx &= \int \frac{1+\cos(u)}{4}\,du \\ &= \frac{1}{4}u + \frac{1}{4}\sin u + C \\ &= \frac{1}{4}(2x + \sin(2x)) + C.\end{aligned}$$

Also

$$\int \sin^2(x)\,dx = -\frac{1}{2}\cos x \sin x + \frac{1}{2}x + C$$

which is left as an exercise.

Example 9.2.8 *Find* $\int \tan(x)\,dx$

Let $u = \cos x$ so that $du = -\sin(x)\,dx$. Then writing the antiderivative in terms of u, this becomes $\int \frac{-1}{u}\,du$. At this point, recall that $(\ln|u|)' = 1/u$. Thus this antiderivative is $-\ln|u| + C = \ln\left|u^{-1}\right| + C$ and so $\int \tan(x)\,dx = \ln|\sec x| + C$.

This illustrates a general procedure.

Procedure 9.2.9 $\int \frac{f'(x)}{f(x)}dx = \ln|f(x)| + C$.

This follows from the chain rule.

Example 9.2.10 *Find* $\int \sec(x)\,dx$.

This is usually done by a trick. You write as $\int \frac{\sec(x)(\sec(x)+\tan(x))}{(\sec(x)+\tan(x))}dx$ and note that the numerator of the integrand is the derivative of the denominator. Thus $\int \sec(x)\,dx = \ln|\sec(x) + \tan(x)| + C$.

Example 9.2.11 *Find* $\int \csc(x)\,dx$.

This is done like the antiderivatives for the secant. $\frac{d}{dx}\csc(x) = -\csc(x)\cot(x)$ and $\frac{d}{dx}\cot(x) = -\csc^2(x)$. Write the integral as $-\int \frac{-\csc(x)(\cot(x)+\csc(x))}{(\cot(x)+\csc(x))}dx = -\ln|\cot(x) + \csc(x)| + C$.

9.3 Exercises

1. Find the indicated antiderivatives.

   (a) $\int \frac{x}{\sqrt{2x-3}}\,dx$

   (b) $\int x\left(3x^2+6\right)^5\,dx$

   (c) $\int x\sin\left(x^2\right)\,dx$

   (d) $\int \sin^3(2x)\cos(2x)$

   (e) $\int \frac{1}{\sqrt{1+4x^2}}\,dx$ **Hint:** Remember the $\sinh^{-1}$ function and its derivative.

2. Solve the initial value problems.

   (a) $\frac{dy}{dx} = \frac{x}{\sqrt{2x-3}}, y(0) = 1$

   (b) $\frac{dy}{dx} = 5x\left(3x^2+6\right)^5, y(0) = 3$

   (c) $\frac{dy}{dx} = 3x^2 \sin\left(2x^3\right), y(1) = 1$

   (d) $y'(x) = \frac{1}{\sqrt{1+3x^2}}.y(1) = 1$

   (e) $y'(x) = \sec(x), y(0) = 3$

   (f) $y'(x) = x\csc\left(x^2\right), y(1) = 1$

3. An object moves on the x axis having velocity equal to $\frac{3t^2}{7+t^4}$. Find the position of the object given that at $t = 1$, it is at the point 2.

4. An object moves on the x axis having velocity equal to $t\sin\left(2t^2\right)$. Find the position of the object given that at $t = 1$, it is at the point 1.

5. An object moves on the x axis having velocity equal to $\sec(t)$. Find the position of the object given that at $t = 1$, it is at the point -2.

6. Find the indicated antiderivatives.

   (a) $\int \sec(3x)\, dx$

   (b) $\int \sec^2(3x)\tan(3x)\, dx$

   (c) $\int \frac{1}{3+5x^2}\, dx$

   (d) $\int \frac{1}{\sqrt{5-4x^2}}\, dx$

   (e) $\int \frac{3}{x\sqrt{4x^2-5}}\, dx$

7. Find the indicated antiderivatives.

   (a) $\int x\cosh\left(x^2+1\right)\, dx$

   (b) $\int x^3 5^{x^4}\, dx$

   (c) $\int \sin(x)\, 7^{\cos(x)}\, dx$

   (d) $\int x\sin\left(x^2\right)\, dx$

   (e) $\int x^5\sqrt{2x^2+1}\, dx$ **Hint:** Let $u = 2x^2+1$.

8. Find $\int \sin^2(x)\, dx$. **Hint:** Derive and use $\sin^2(x) = \frac{1-\cos(2x)}{2}$.

9. Find the indicated antiderivatives.

   (a) $\int \frac{\ln x}{x}\, dx$

   (b) $\int \frac{x^3}{3+x^4}\, dx$

   (c) $\int \frac{1}{x^2+2x+2}\, dx$ **Hint:** Complete the square in the denominator and then let $u = x+1$.

   (d) $\int \frac{1}{\sqrt{4-x^2}}\, dx$

   (e) $\int \frac{1}{x\sqrt{x^2-9}}\, dx$ **Hint:** Let $x = 3u$.

   (f) $\int \frac{\ln\left(x^2\right)}{x}\, dx$

(g) Find $\int \frac{x^3}{\sqrt{(6x^2+5)}}\,dx$

(h) Find $\int x\sqrt[3]{(6x+4)}\,dx$

10. Find the indicated antiderivatives.

(a) $\int x\sqrt{2x+4}\,dx$

(b) $\int x\sqrt{3x+2}\,dx$

(c) $\int \frac{1}{\sqrt{36-25x^2}}\,dx$

(d) $\int \frac{1}{x}\sqrt{3x+5}\,dx$

(e) $\int \frac{1}{\sqrt{9-4x^2}}\,dx$

(f) $\int \frac{1}{\sqrt{(1+4x^2)}}\,dx$

(g) $\int \frac{x}{\sqrt{(3x-1)}}\,dx$

(h) $\int \frac{1}{x^2}\sqrt[3]{6x+4}\,dx$

(i) $\int \frac{x}{\sqrt{5x+1}}\,dx$

(j) $\int \frac{1}{x\sqrt{9x^2-4}}\,dx$

(k) $\int \frac{1}{\sqrt{9+4x^2}}\,dx$

11. Find $\int \frac{1}{x^{1/3}+x^{1/2}}dx$. **Hint:** Try letting $x=u^6$ and use long division.

9.4 Integration By Parts

Another technique for finding antiderivatives is called integration by parts and is based on the product rule. Recall the product rule. If u' and v' exist, then

$$(uv)'(x) = u'(x)\,v(x) + u(x)\,v'(x). \tag{9.3}$$

Therefore,

$$(uv)'(x) - u'(x)\,v(x) = u(x)\,v'(x)$$

Proposition 9.4.1 *Let u and v be differentiable functions for which $\int u(x)\,v'(x)\,dx$ and $\int u'(x)\,v(x)\,dx$ are nonempty. Then*

$$uv - \int u'(x)\,v(x)\,dx = \int u(x)\,v'(x)\,dx. \tag{9.4}$$

Proof: Let $F \in \int u'(x)\,v(x)\,dx$. Then

$$(uv-F)' = (uv)' - F' = (uv)' - u'v = uv'$$

by the chain rule. Therefore every function from the left in (9.4) is a function found in the right side of (9.4). Now let $G \in \int u(x)\,v'(x)\,dx$. Then $(uv-G)' = -uv' + (uv)' = u'v$ by the chain rule. It follows that $uv - G \in \int u'(x)\,v(x)\,dx$ and so $G \in uv - \int u'(x)\,v(x)\,dx$. Thus every function from the right in (9.4) is a function from the left. This proves the proposition.

Example 9.4.2 *Find $\int x\sin(x)\,dx$*

Let $u(x) = x$ and $v'(x) = \sin(x)$. Then applying (9.4),

$$\begin{aligned}\int x\sin(x)\,dx &= (-\cos(x))\,x - \int(-\cos(x))\,dx \\ &= -x\cos(x) + \sin(x) + C.\end{aligned}$$

Example 9.4.3 *Find* $\int x\ln(x)\,dx$

Let $u(x) = \ln(x)$ and $v'(x) = x$. Then from (9.4),

$$\begin{aligned}\int x\ln(x)\,dx &= \frac{x^2}{2}\ln(x) - \int\frac{x^2}{2}\left(\frac{1}{x}\right) \\ &= \frac{x^2}{2}\ln(x) - \int\frac{x}{2} \\ &= \frac{x^2}{2}\ln(x) - \frac{1}{4}x^2 + C\end{aligned}$$

Example 9.4.4 *Find* $\int \arctan(x)\,dx$

Let $u(x) = \arctan(x)$ and $v'(x) = 1$. Then from (9.4),

$$\begin{aligned}\int \arctan(x)\,dx &= x\arctan(x) - \int x\left(\frac{1}{1+x^2}\right)dx \\ &= x\arctan(x) - \frac{1}{2}\int\frac{2x}{1+x^2}\,dx \\ &= x\arctan(x) - \frac{1}{2}\ln\left(1+x^2\right) + C.\end{aligned}$$

Sometimes you want to find antiderivatives for something like $\int fgdx$ where $f^{(m)} = 0$ for some positive integer, m. For example, $\int x^5\sin x\,dx$. If you do integration by parts repeatedly, what do you get? Let $G_1' = g, G_2' = G_1, G_3' = G_2$ etc. Then the first application of integration by parts yields $fG_1 - \int G_1 f'dx$. The next application of integration by parts yields $fG_1 - G_2f' + \int G_2f''dx$. Yet another application of integration by parts yields $fG_1 - G_2f' + G_3f'' - \int G_3f'''dx$. Eventually the process will stop because a high enough derivative of f equals zero. This justifies the following procedure for finding antiderivatives in this case.

Procedure 9.4.5 *Suppose* $f^{(m)} = 0$ *for some* m *a positive integer and let* $G_k' = G_{k-1}$ *for all* k *and* $G_0 = g$. *Then*

$$\int fg\,dx = fG_1 - f'G_2 + f''G_3 - f'''G_4 + \cdots$$

Just keep writing these terms, alternating signs until the process yields a zero. Then add on an arbitrary constant of integration and stop.

Example 9.4.6 *Find* $\int x^5\sin x\,dx$.

From the above procedure and letting $f(x) = x^5$, this equals

$$x^5(-\cos(x)) - 5x^4(-\sin x) + 20x^3(\cos x) - 60x^2(\sin x) + 120x(-\cos x) - 120(-\sin x) + C.$$

9.5 Exercises

1. Find the following antiderivatives.

   (a) $\int x^3 e^{-3x}\,dx$
   (b) $\int x^4 \cos x\,dx$
   (c) $\int x^5 e^x\,dx$
   (d) $\int x^6 \sin(2x)\,dx$
   (e) $\int x^3 \cos\left(x^2\right)\,dx$

2. Find the following antiderivatives.

   (a) $\int xe^{-3x}\,dx$
   (b) $\int \frac{1}{x(\ln(|x|))^2}\,dx$
   (c) $\int x\sqrt{2-x}\,dx$
   (d) $\int (\ln|x|)^2\,dx$ **Hint:** Let $u(x) = (\ln|x|)^2$ and $v'(x) = 1$.
   (e) $\int x^3 \cos\left(x^2\right)\,dx$

3. Show that $\int \sec^3(x)\,dx = \frac{1}{2}\tan(x)\sec(x) + \frac{1}{2}\ln|\sec x + \tan x| + C$.

4. Consider the following argument. Integrate by parts, letting $u(x) = x$ and $v'(x) = \frac{1}{x^2}$ to get

$$\begin{aligned}\int \frac{1}{x}\,dx &= \int x\left(\frac{1}{x^2}\right)dx = \left(-\frac{1}{x}\right)x + \int \frac{1}{x}\,dx \\ &= -1 + \int \frac{1}{x}\,dx.\end{aligned}$$

   Now subtracting $\int \frac{1}{x}\,dx$ from both sides, $0 = -1$. Is there anything wrong here? If so, what?

5. Find the following antiderivatives.

   (a) $\int x^3 \arctan(x)\,dx$
   (b) $\int x^3 \ln(x)\,dx$
   (c) $\int x^2 \sin(x)\,dx$
   (d) $\int x^2 \cos(x)\,dx$
   (e) $\int x \arcsin(x)\,dx$
   (f) $\int \cos(2x)\sin(3x)\,dx$
   (g) $\int x^3 e^{x^2}\,dx$
   (h) $\int x^3 \cos\left(x^2\right)\,dx$

6. Find the antiderivatives

   (a) $\int x^2 \sin x\,dx$
   (b) $\int x^2 \sin x\,dx$
   (c) $\int x^3 7^x\,dx$

(d) $\int x^2 \ln x \, dx$

(e) $\int (x+2)^2 e^x \, dx$

(f) $\int x^3 2^x \, dx$

(g) $\int \sec^3(2x) \tan(2x) \, dx$

(h) $\int x^2 7^x \, dx$

7. Solve the initial value problem $y'(x) = f(x), y(1) = 1$ where $f(x)$ is each of the integrands in Problem 6.

8. Solve the initial value problem $y'(x) = f(x), y(1) = 2$ where $f(x)$ is each of the integrands in Problem 5.

9. Try doing $\int \sin^2 x \, dx$ the obvious way. If you don't make any mistakes, the process will go in circles. Now do it by taking $\int \sin^2 x \, dx = x \sin^2 x - 2 \int x \sin x \cos x \, dx = x \sin^2 x - \int x \sin(2x) \, dx$.

10. An object moves on the x axis having velocity equal to $t \sin t$. Find the position of the object given that at $t = 1$, it is at the point 2.

11. An object moves on the x axis having velocity equal to $\sec^3(t)$. Find the position of the object given that at $t = 1$, it is at the point 2. **Hint:** You might want to use Problem 3.

9.6 Trig. Substitutions

Certain antiderivatives are easily obtained by making an auspicious substitution involving a trig. function. The technique will be illustrated by presenting examples.

Example 9.6.1 *Find* $\int \frac{1}{(x^2+2x+2)^2} \, dx$.

Complete the square as before and write

$$\int \frac{1}{(x^2+2x+2)^2} \, dx = \int \frac{1}{\left((x+1)^2+1\right)^2} \, dx$$

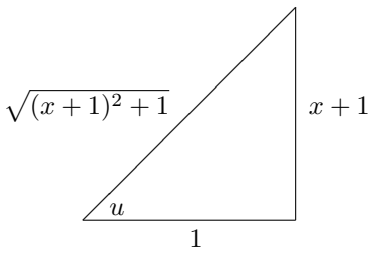
Use the following substitution next.

$$x + 1 = \tan u \tag{9.5}$$

so $dx = \left(\sec^2 u\right) du$. Therefore, this last indefinite integral becomes

$$\begin{aligned} \int \frac{\sec^2 u}{\left(\tan^2 u + 1\right)^2} \, du &= \int \left(\cos^2 u\right) du \\ &= \int \frac{1+\cos 2u}{2} \, du \\ &= \frac{u}{2} + \frac{\sin 2u}{4} + C \\ &= \frac{u}{2} + \frac{2 \sin u \cos u}{4} + C \end{aligned}$$

Next write this in terms of x using the following device based on the following picture.



In this picture which is descriptive of (9.5), $\sin u = \frac{x+1}{\sqrt{(x+1)^2+1}}$ and $\cos u = \frac{1}{\sqrt{(x+1)^2+1}}$. Therefore, putting in this information to change back to the x variable,

$$\int \frac{1}{\left(x^2+2x+2\right)^2}\,dx$$

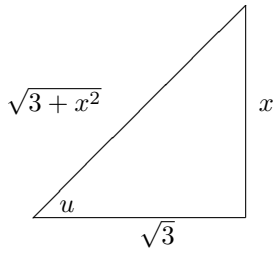
$$\begin{aligned} &= \frac{1}{2}\arctan(x+1) + \frac{1}{2}\frac{x+1}{\sqrt{(x+1)^2+1}}\frac{1}{\sqrt{(x+1)^2+1}} + C \\ &= \frac{1}{2}\arctan(x+1) + \frac{1}{2}\frac{x+1}{(x+1)^2+1} + C. \end{aligned}$$

Example 9.6.2 *Find* $\int \frac{1}{\sqrt{x^2+3}}\,dx$.

Let $x = \sqrt{3}\tan u$ so $dx = \sqrt{3}\left(\sec^2 u\right)du$. Making the substitution, consider

$$\begin{aligned} &\int \frac{1}{\sqrt{3}\sqrt{\tan^2 u+1}}\sqrt{3}\left(\sec^2 u\right)\,du \\ &= \int (\sec u)\,du = \ln|\sec u + \tan u| + C \end{aligned}$$

Now the following diagram is descriptive of the above transformation.



Using the above diagram, $\sec u = \frac{\sqrt{3+x^2}}{\sqrt{3}}$ and $\tan u = \frac{x}{\sqrt{3}}$. Therefore, restoring the x variable,

$$\begin{aligned} \int \frac{1}{\sqrt{x^2+3}}\,dx &= \ln\left|\frac{\sqrt{3+x^2}}{\sqrt{3}} + \frac{x}{\sqrt{3}}\right| + C \\ &= \ln\left|\sqrt{3+x^2} + x\right| + C. \end{aligned}$$

Note the constant C changed in going from the top to the bottom line. It is $C - \ln\sqrt{3}$ but it is customary to simply write this as C because C is arbitrary.

Example 9.6.3 *Find* $\int \left(4x^2+3\right)^{1/2}\,dx$.

Let $2x = \sqrt{3}\tan u$ so $2dx = \sqrt{3}\sec^2(u)\,du$. Then making the substitution,

$$\begin{aligned} & \sqrt{3}\int \left(\tan^2 u + 1\right)^{1/2} \frac{\sqrt{3}}{2}\sec^2(u)\,du \\ = \quad & \frac{3}{2}\int \sec^3(u)\,du. \end{aligned} \tag{9.6}$$

Now use integration by parts to obtain

$$\int \sec^3(u)\,du = \int \sec^2(u)\sec(u)\,du =$$

$$\begin{aligned} &= \tan(u)\sec(u) - \int \tan^2(u)\sec(u)\,du \\ &= \tan(u)\sec(u) - \int \left(\sec^2(u)-1\right)\sec(u)\,du \\ &= \tan(u)\sec(u) + \int \sec(u)\,du - \int \sec^3(u)\,du \\ &= \tan(u)\sec(u) + \ln|\sec(u)+\tan(u)| - \int \sec^3(u)\,du \end{aligned}$$

Therefore,

$$2\int \sec^3(u)\,du = \tan(u)\sec(u) + \ln|\sec(u)+\tan(u)| + C$$

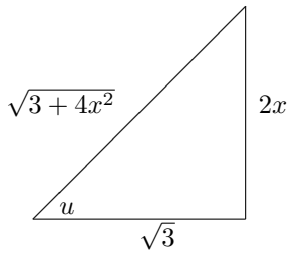
and so

$$\int \sec^3(u)\,du = \frac{1}{2}\left[\tan(u)\sec(u) + \ln|\sec(u)+\tan(u)|\right] + C. \tag{9.7}$$

Now it follows from (9.6) that in terms of u the set of antiderivatives is given by

$$\frac{3}{4}\left[\tan(u)\sec(u) + \ln|\sec(u)+\tan(u)|\right] + C$$

Use the following diagram to change back to the variable, x.



From the diagram, $\tan(u) = \frac{2x}{\sqrt{3}}$ and $\sec(u) = \frac{\sqrt{3+4x^2}}{\sqrt{3}}$. Therefore,

$$\int \left(4x^2+3\right)^{1/2}\,dx =$$

$$\begin{aligned} & \frac{3}{4}\left[\frac{2x}{\sqrt{3}}\frac{\sqrt{3+4x^2}}{\sqrt{3}}+\ln\left|\frac{\sqrt{3+4x^2}}{\sqrt{3}}+\frac{2x}{\sqrt{3}}\right|\right]+C \\ = & \frac{3}{4}\left[\frac{2x}{\sqrt{3}}\frac{\sqrt{3+4x^2}}{\sqrt{3}}+\ln\left|\frac{\sqrt{3+4x^2}}{\sqrt{3}}+\frac{2x}{\sqrt{3}}\right|\right]+C \\ = & \frac{1}{2}x\sqrt{(3+4x^2)}+\frac{3}{4}\ln\left|\sqrt{3+4x^2}+2x\right|+C \end{aligned}$$

Note that these examples involved something of the form $\left(a^2+(bx)^2\right)$ and the trig substitution,

$$bx = a\tan u$$

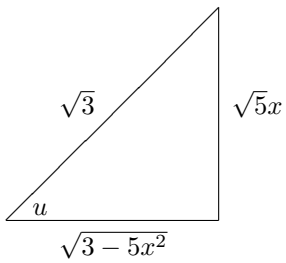
was the right one to use. This is the auspicious substitution which often simplifies these sorts of problems.

Example 9.6.4 *Find* $\int \sqrt{3-5x^2}\,dx$

In this example, let $\sqrt{5}x = \sqrt{3}\sin(u)$ so $\sqrt{5}dx = \sqrt{3}\cos(u)\,du$. The reason this might be a good idea is that it will get rid of the square root sign as shown below. Making the substitution,

$$\begin{aligned} \int \sqrt{3-5x^2}\,dx &= \sqrt{3}\int \sqrt{1-\sin^2(u)}\frac{\sqrt{3}}{\sqrt{5}}\cos(u)\;du \\ &= \frac{3}{\sqrt{5}}\int \cos^2(u)\;du \\ &= \frac{3}{\sqrt{5}}\int \frac{1+\cos 2u}{2}\,du \\ &= \frac{3}{\sqrt{5}}\left(\frac{u}{2}+\frac{\sin 2u}{4}\right)+C \\ &= \frac{3}{2\sqrt{5}}u+\frac{3}{2\sqrt{5}}\sin u\cos u+C \end{aligned}$$

The appropriate diagram is the following.



From the diagram, $\sin(u) = \frac{\sqrt{5}x}{\sqrt{3}}$ and $\cos(u) = \frac{\sqrt{3-5x^2}}{\sqrt{3}}$. Therefore, changing back to x,

$$\int \sqrt{3-5x^2}\,dx =$$

$$\begin{aligned} & \frac{3}{2\sqrt{5}}\arcsin\left(\frac{\sqrt{5}x}{\sqrt{3}}\right)+\frac{3}{2\sqrt{5}}\frac{\sqrt{5}x}{\sqrt{3}}\frac{\sqrt{3-5x^2}}{\sqrt{3}}+C \\ = & \frac{3}{10}\sqrt{5}\arcsin\left(\frac{1}{3}\sqrt{15}x\right)+\frac{1}{2}x\sqrt{(3-5x^2)}+C \end{aligned}$$

Example 9.6.5 *Find* $\int \sqrt{5x^2-3}\,dx$

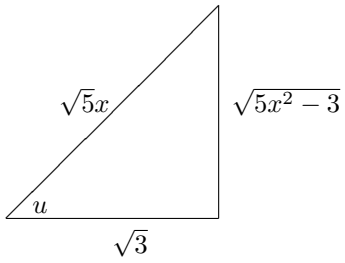
In this example, let $\sqrt{5}x=\sqrt{3}\sec(u)$ so $\sqrt{5}dx=\sqrt{3}\sec(u)\tan(u)\,du$. Then changing the variable, consider

$$\begin{aligned} & \sqrt{3}\int\sqrt{\sec^2(u)-1}\frac{\sqrt{3}}{\sqrt{5}}\sec(u)\tan(u)\,du \\ = & \frac{3}{\sqrt{5}}\int\tan^2(u)\sec(u)\,du \\ = & \frac{3}{\sqrt{5}}\left[\int\sec^3(u)\,du-\int\sec(u)\,du\right]. \end{aligned}$$

Now from (9.7), this equals

$$\begin{aligned} & \frac{3}{\sqrt{5}}\left[\frac{1}{2}\left[\tan(u)\sec(u)+\ln|\sec(u)+\tan(u)|\right]-\ln|\tan(u)+\sec(u)|\right]+C \\ = & \frac{3}{2\sqrt{5}}\tan(u)\sec(u)-\frac{3}{2\sqrt{5}}\ln|\sec(u)+\tan(u)|+C. \end{aligned}$$

Now it is necessary to change back to x. The diagram is as follows.



Therefore, $\tan(u)=\frac{\sqrt{5x^2-3}}{\sqrt{3}}$ and $\sec(u)=\frac{\sqrt{5}x}{\sqrt{3}}$ and so

$$\int\sqrt{5x^2-3}\,dx=$$

$$\begin{aligned} = & \frac{3}{2\sqrt{5}}\frac{\sqrt{5x^2-3}}{\sqrt{3}}\frac{\sqrt{5}x}{\sqrt{3}}-\frac{3}{2\sqrt{5}}\ln\left|\frac{\sqrt{5}x}{\sqrt{3}}+\frac{\sqrt{5x^2-3}}{\sqrt{3}}\right|+C \\ = & \frac{1}{2}\left(\sqrt{5x^2-3}\right)x-\frac{3}{10}\sqrt{5}\ln\left|\sqrt{5}x+\sqrt{(-3+5x^2)}\right|+C \end{aligned}$$

To summarize, here is a short table of auspicious substitutions corresponding to certain expressions.

Expression	$a^2+b^2x^2$	$a^2-b^2x^2$	$a^2x^2-b^2$
Trig. substitution	$bx=a\tan(u)$	$bx=a\sin(u)$	$ax=b\sec(u)$

Of course there are no "magic bullets" but these substitutions will often simplify an expression enough to allow you to find an antiderivative. These substitutions are often especially useful when the expression is enclosed in a square root.

9.7 Exercises

1. Find the antiderivatives.
   (a) $\int \frac{x}{\sqrt{4-x^2}}\,dx$
   (b) $\int \frac{3}{\sqrt{36-25x^2}}\,dx$
   (c) $\int \frac{3}{\sqrt{16-25x^2}}\,dx$
   (d) $\int \frac{1}{\sqrt{4-9x^2}}\,dx$
   (e) $\int \frac{1}{\sqrt{36-x^2}}\,dx$
   (f) $\int \left(\sqrt{9-16x^2}\right)^3\,dx$
   (g) $\int \left(\sqrt{16-x^2}\right)^5\,dx$
   (h) $\int \sqrt{25-36x^2}\,dx$
   (i) $\int \left(\sqrt{4-9x^2}\right)^3\,dx$
   (j) $\int \sqrt{1-9x^2}\,dx$

2. Find the antiderivatives.
   (a) $\int \sqrt{36x^2-25}\,dx$
   (b) $\int \sqrt{x^2-4}\,dx$
   (c) $\int \left(\sqrt{16x^2-9}\right)^3\,dx$
   (d) $\int \sqrt{25x^2-16}\,dx$

3. Find the antiderivatives.
   (a) $\int \frac{1}{26+x^2-2x}\,dx$ **Hint:** Complete the square.
   (b) $\int \sqrt{x^2+9}\,dx$
   (c) $\int \sqrt{4x^2+25}\,dx$
   (d) $\int x\sqrt{4x^4+9}\,dx$
   (e) $\int x^3\sqrt{4x^4+9}\,dx$
   (f) $\int \frac{1}{\left(25+36(2x-3)^2\right)^2}\,dx$
   (g) $\int \frac{1}{\left(16+25(x-3)^2\right)^2}\,dx$
   (h) $\int \frac{1}{261+25x^2-150x}\,dx$ **Hint:** Complete the square.
   (i) $\int \left(\sqrt{25x^2+9}\right)^3\,dx$
   (j) $\int \frac{1}{25+16x^2}\,dx$

4. Find the antiderivatives. **Hint:** Complete the square.

(a) $\int \sqrt{4x^2+16x+15}\,dx$
(b) $\int \sqrt{x^2+6x}\,dx$
(c) $\int \frac{3}{\sqrt{-32-9x^2-36x}}\,dx$
(d) $\int \frac{3}{\sqrt{-5-x^2-6x}}\,dx$
(e) $\int \frac{1}{\sqrt{9-16x^2-32x}}\,dx$
(f) $\int \sqrt{4x^2+16x+7}\,dx$

9.8 Partial Fractions

The main technique for finding antiderivatives in the case $f(x) = \frac{p(x)}{q(x)}$ for p and q polynomials is the technique of partial fractions. Before presenting this technique, a few more examples are presented.

Example 9.8.1 *Find* $\int \frac{1}{x^2+2x+2}\,dx$.

To do this complete the square in the denominator to write

$$\int \frac{1}{x^2+2x+2}\,dx = \int \frac{1}{(x+1)^2+1}\,dx$$

Now change the variable, letting $u = x+1$ so that $du = dx$. Then the last indefinite integral reduces to

$$\int \frac{1}{u^2+1}\,du = \arctan u + C$$

and so

$$\int \frac{1}{x^2+2x+2}\,dx = \arctan(x+1) + C.$$

Example 9.8.2 *Find* $\int \frac{1}{3x+5}\,dx$.

Let $u = 3x+5$ so $du = 3dx$ and changing the variable,

$$\frac{1}{3}\int \frac{1}{u}\,du = \frac{1}{3}\ln|u| + C.$$

Therefore,

$$\int \frac{1}{3x+5}\,dx = \frac{1}{3}\ln|3x+5| + C.$$

Example 9.8.3 *Find* $\int \frac{3x+2}{x^2+x+1}\,dx$.

First complete the square in the denominator.

$$\begin{aligned}\int \frac{3x+2}{x^2+x+1}\,dx &= \int \frac{3x+2}{x^2+x+\frac{1}{4}+\frac{3}{4}}\,dx \\ &= \int \frac{3x+2}{\left(x+\frac{1}{2}\right)^2+\frac{3}{4}}\,dx.\end{aligned}$$

Now let

$$\left(x+\frac{1}{2}\right)^2 = \frac{3}{4}u^2$$

so that $x+\frac{1}{2}=\frac{\sqrt{3}}{2}u$. Therefore, $dx=\frac{\sqrt{3}}{2}du$ and changing the variable,

$$\frac{4}{3}\int \frac{3\left(\frac{\sqrt{3}}{2}u-\frac{1}{2}\right)+2}{u^2+1}\frac{\sqrt{3}}{2}\,du$$

$$\begin{aligned}
&= \frac{\sqrt{3}}{2}\left(2\sqrt{3}\int \frac{u}{u^2+1}\,du-\frac{2}{3}\int \frac{1}{u^2+1}\,du\right)\\
&= \frac{\sqrt{3}}{2}\left(\sqrt{3}\int \frac{2u}{u^2+1}\,du-\frac{2}{3}\int \frac{1}{u^2+1}\,du\right)\\
&= \frac{3}{2}\ln\left(u^2+1\right)-\frac{\sqrt{3}}{3}\arctan u+C
\end{aligned}$$

Therefore,

$$\int \frac{3x+2}{x^2+x+1}\,dx=$$

$$\frac{3}{2}\ln\left(\left(\frac{2}{\sqrt{3}}\left(x+\frac{1}{2}\right)\right)^2+1\right)-\frac{\sqrt{3}}{3}\arctan\left(\frac{2}{\sqrt{3}}\left(x+\frac{1}{2}\right)\right)+C.$$

The following simple but important Lemma is needed to continue.

Lemma 9.8.4 *Let $f(x)$ and $g(x)$ be polynomials. Then there exists a polynomial, $q(x)$ such that*

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

where the degree of $r(x)<$ degree of $g(x)$ or $r(x)=0$.

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 discussed earlier. 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 $r(x)<$ 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

$$\begin{aligned}
r(x) &= b_m x^m+\cdots+b_1x+b_0\\
g(x) &= a_n x^n+\cdots+a_1x+a_0
\end{aligned}$$

where $m\geq n$ and b_m and a_n are nonzero. Then letting

$$\begin{aligned}
r_1(x) &= r(x)-\frac{x^{m-n}b_m}{a_n}g(x)\\
&= f(x)-g(x)\,q_1(x)-\frac{x^{m-n}b_m}{a_n}g(x)\\
&= f(x)-g(x)\left(q_1(x)+\frac{x^{m-n}b_m}{a_n}\right),
\end{aligned}$$

it follows this is not zero by the assumption that $f(x)-g(x)\,l(x)$ is never equal to zero for any $l(x)$. Now the degree of $r_1(x)<$ degree of $r(x)$, a contradiction to the construction of $r(x)$. This proves the lemma.

Corollary 9.8.5 *Let $f(x)$ and $g(x)$ be polynomials. Then there exists a polynomial, $r(x)$ such that the degree of $r(x)$ < degree of $g(x)$ and a polynomial, $q(x)$ such that*

$$\frac{f(x)}{g(x)} = q(x) + \frac{r(x)}{g(x)}.$$

Example 9.8.6 *Find* $\int \frac{-x^3+11x^2+24x+14}{(2x+3)(x+5)(x^2+x+1)}\,dx$.

In this problem, first check to see if the degree of the numerator in the integrand is less than the degree of the denominator. In this case, this is so. If it is not so, use long division to write the integrand as the sum of a polynomial with a rational function in which the degree of the numerator is less than the degree of the denominator. See the preceding corollary which guarantees this can be done. Now look for a partial fractions expansion for the integrand which is in the following form.

$$\frac{a}{2x+3} + \frac{b}{x+5} + \frac{cx+d}{x^2+x+1}$$

and try to find constants, a, b, c, and d so that the above rational functions sum to the integrand. The reason $cx + d$ is used in the numerator of the last expression is that $x^2 + x + 1$ cannot be factored using real polynomials. Thus the problem involves finding a, b, c, d, such that

$$\frac{-x^3+11x^2+24x+14}{(2x+3)(x+5)(x^2+x+1)} = \frac{a}{2x+3} + \frac{b}{x+5} + \frac{cx+d}{x^2+x+1}$$

and so

$$-x^3+11x^2+24x+14 = a(x+5)\left(x^2+x+1\right) +$$
$$b(2x+3)\left(x^2+x+1\right) + (cx+d)(2x+3)(x+5). \tag{9.8}$$

Now these are two polynomials which are supposed to be equal. Therefore, they have the same coefficients. Multiplying the right side out and collecting the terms,

$$-x^3+11x^2+24x+14 =$$

$$(2b+2c+a)x^3 + (6a+5b+13c+2d)x^2 + (6a+13d+5b+15c)x + 15d+5a+3b$$

and therefore, it is necessary to solve the equations,

$$\begin{gathered} 2b+2c+a = -1 \\ 6a+5b+13c+2d = 11 \\ 6a+13d+5b+15c = 24 \\ 15d+5a+3b = 14 \end{gathered}$$

The solution is $c = 1, a = 1, b = -2, d = 1$. Therefore,

$$\frac{-x^3+11x^2+24x+14}{(2x+3)(x+5)(x^2+x+1)} = \frac{1}{2x+3} - \frac{2}{x+5} + \frac{1+x}{x^2+x+1}.$$

This may look like a fairly formidable problem. In reality it is not that bad. First let $x = -5$ in (9.8) and obtain a simple equation for finding b. Next let $x = -3/2$ to get a simple equation for a. This reduces the above system to a more manageable size. Anyway, it is now possible to find an antiderivative of the given function.

$$\int \frac{-x^3+11x^2+24x+14}{(2x+3)(x+5)(x^2+x+1)}\,dx =$$

$$\int \frac{1}{2x+3}\,dx - \int \frac{2}{x+5}\,dx + \int \frac{1+x}{x^2+x+1}\,dx.$$

Each of these indefinite integrals can be found using the techniques given above. Thus the antiderivatives are

$$\frac{1}{2}\ln|2x+3| - 2\ln|x+5| +$$

$$\frac{1}{2}\ln\left(x^2+x+1\right) + \frac{1}{3}\sqrt{3}\arctan\left(\frac{\sqrt{3}}{3}(2x+1)\right) + C.$$

This was a long example. Here is an easier one.

Example 9.8.7 *Find* $\int \frac{3x^5+7}{x^2-1}\,dx$.

In this case the degree of the numerator is larger than the degree of the denominator and so long division must first be used. Thus

$$\frac{3x^5+7}{x^2-1} = 3x^3 + 3x + \frac{7+3x}{x^2-1}$$

Now look for a partial fractions expansion of the form

$$\frac{7+3x}{x^2-1} = \frac{a}{(x-1)} + \frac{b}{(x+1)}.$$

Therefore,

$$7+3x = a(x+1) + b(x-1).$$

Letting $x=1$, $a=5$. Then letting $x=-1$, it follows $b=-2$. Therefore,

$$\frac{7+3x}{x^2-1} = \frac{5}{x-1} - \frac{2}{x+1}$$

and so

$$\frac{3x^5+7}{x^2-1} = 3x^3 + 3x + \frac{5}{x-1} - \frac{2}{x+1}.$$

therefore,

$$\int \frac{3x^5+7}{x^2-1}\,dx = \frac{3}{4}x^4 + \frac{3}{2}x^2 + 5\ln(x-1) - 2\ln(x+1) + C.$$

What is done when the factors are repeated?

Example 9.8.8 *Find* $\int \frac{3x+7}{(x+2)^2(x+3)}\,dx$.

First observe that the degree of the numerator is less than the degree of the denominator. In this case the correct form of the partial fraction expansion is

$$\frac{a}{(x+2)} + \frac{b}{(x+2)^2} + \frac{c}{(x+3)}.$$

The reason there are two terms devoted to $(x+2)$ is that this is squared. Computing the constants yields

$$\frac{3x+7}{(x+2)^2(x+3)} = \frac{1}{(x+2)^2} + \frac{2}{x+2} - \frac{2}{x+3}$$

and therefore,

$$\int \frac{3x+7}{(x+2)^2(x+3)}\,dx = -\frac{1}{x+2} + 2\ln(x+2) - 2\ln(x+3) + C.$$

Example 9.8.9 *Find the proper form for the partial fractions expansion of*

$$\frac{x^3+7x+9}{\left(x^2+2x+2\right)^3\left(x+2\right)^2\left(x+1\right)\left(x^2+1\right)}.$$

First check to see if the degree of the numerator is smaller than the degree of the denominator. Since this is the case, look for a partial fractions decomposition in the following form.

$$\frac{ax+b}{\left(x^2+2x+2\right)}+\frac{cx+d}{\left(x^2+2x+2\right)^2}+\frac{ex+f}{\left(x^2+2x+2\right)^3}+$$
$$\frac{A}{\left(x+2\right)}+\frac{B}{\left(x+2\right)^2}+\frac{D}{\left(x+1\right)}+\frac{gx+h}{x^2+1}.$$

These examples illustrate what to do when using the method of partial fractions. You first check to be sure the degree of the numerator is less than the degree of the denominator. If this is not so, do a long division. Then you factor the denominator into a product of factors, some linear of the form $ax+b$ and others quadratic, ax^2+bx+c which cannot be factored further. Next follow the procedure illustrated in the above examples.

Warning: When you use partial fractions, **be sure you look for something which is of the right form**. Otherwise you may be looking for something which is not there.

9.9 Rational Functions Of Trig. Functions

Example 9.9.1 *Find* $\int \frac{\cos\theta}{1+\cos\theta}\,d\theta$.

The integrand is an example of a rational function of cosines and sines. When such a thing occurs there is a substitution which will reduce the integrand to a rational function like those above which can then be integrated using partial fractions. The substitution is $u=\tan\left(\frac{\theta}{2}\right)$. Thus in this example, $du=\left(1+\tan^2\left(\frac{\theta}{2}\right)\right)\frac{1}{2}d\theta$ and so in terms of this new variable, the indefinite integral is

$$\int \frac{2\cos\left(2\arctan u\right)}{\left(1+\cos\left(2\arctan u\right)\right)\left(1+u^2\right)}\,du.$$

You can evaluate $\cos\left(2\arctan u\right)$ exactly. This equals $2\cos^2\left(\arctan u\right)-1$. Setting up a little triangle as above, $\cos\left(\arctan u\right)$ equals $1/\sqrt{1+u^2}$ and so the integrand reduces to

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

therefore, in terms of u the antiderivative equals $-u+2\arctan u$. Now replace u to obtain

$$-\tan\left(\frac{\theta}{2}\right)+2\arctan\left(\tan\left(\frac{\theta}{2}\right)\right)+C.$$

This procedure can be expected to work in general. Suppose you want to find $\int \frac{p(\cos\theta,\sin\theta)}{q(\cos\theta,\sin\theta)}d\theta$ where p and q are polynomials in each argument. Make the substitution $u=\tan\frac{\theta}{2}$. As above this means

$$du=\left(1+\tan^2\left(\frac{\theta}{2}\right)\right)\frac{1}{2}d\theta=\frac{1}{2}\left(1+u^2\right)d\theta.$$

It remains to substitute for $\sin\theta$ and $\cos\theta$. Recall that $\sin\left(\frac{\theta}{2}\right) = \pm\sqrt{\frac{1-\cos\theta}{2}}$ and $\cos\left(\frac{\theta}{2}\right) = \pm\sqrt{\frac{1+\cos\theta}{2}}$. Thus,

$$\tan\left(\frac{\theta}{2}\right) = \frac{\pm\sqrt{1-\cos\theta}}{\sqrt{1+\cos\theta}}$$

and so

$$u^2 = \tan^2\left(\frac{\theta}{2}\right) = \frac{1-\cos\theta}{1+\cos\theta}$$

and solving this for $\cos\theta$ and $\sin\theta$ yields

$$\cos\theta = \frac{1-u^2}{1+u^2},\ \sin\theta = \pm\frac{2u}{1+u^2}.$$

It follows that in terms of u the integral becomes

$$\int \frac{p\left(\frac{1-u^2}{1+u^2}, \pm\frac{2u}{1+u^2}\right)}{q\left(\frac{1-u^2}{1+u^2}, \pm\frac{2u}{1+u^2}\right)} \frac{2du}{1+u^2}$$

which is a rational function of u and so in theory, you might be able to find the integral from the method of partial fractions.

9.10 Exercises

1. Give a condition on a, b, and c such that ax^2+bx+c cannot be factored as a product of two polynomials which have real coefficients.

2. Find the partial fractions expansion of the following rational functions.

   (a) $\frac{2x+7}{(x+1)^2(x+2)}$

   (b) $\frac{5x+1}{(x^2+1)(2x+3)}$

   (c) $\frac{5x+1}{(x^2+1)^2(2x+3)}$

   (d) $\frac{5x^4+10x^2+3+4x^3+6x}{(x+1)(x^2+1)^2}$

3. Find the antiderivatives

   (a) $\int \frac{x^5+4x^4+5x^3+2x^2+2x+7}{(x+1)^2(x+2)}\,dx$

   (b) $\int \frac{5x+1}{(x^2+1)(2x+3)}\,dx$

   (c) $\int \frac{5x+1}{(x^2+1)^2(2x+3)}$

4. Each of $\cot\theta, \tan\theta, \sec\theta$, and $\csc\theta$ is a rational function of $\cos\theta$ and $\sin\theta$. Use the technique of substituting $u = \tan\left(\frac{\theta}{2}\right)$ to find antiderivatives for each of these.

5. Find $\int \frac{\sin\theta}{1+\sin\theta}\,d\theta$. **Hint:** Use the above procedure of letting $u = \tan\left(\frac{\theta}{2}\right)$ and then multiply both the top and the bottom by $(1-\sin\theta)$ to see another way of doing it.

6. Find $\int \frac{\cos\theta+1}{\cos\theta+2}d\theta$ using the substitution $u = \tan\left(\frac{\theta}{2}\right)$.

7. In finding $\int \sec(x)\,dx$, try the substitution $u = \sin(x)$.

8. In finding $\int \csc(x)\,dx$ try the substitution $u = \cos(x)$.

9. Find the antiderivatives.
   (a) $\int \frac{17x-3}{(6x+1)(x-1)}\,dx$
   (b) $\int \frac{50x^4-95x^3-20x^2-3x+7}{(5x+3)(x-2)(2x-1)}\,dx$ **Hint:** Notice the degree of the numerator is larger than the degree of the denominator.
   (c) $\int \frac{8x^2+x-5}{(3x+1)(x-1)(2x-1)}\,dx$
   (d) $\int \frac{3x+2}{(5x+3)(x+1)}\,dx$

10. Find the antiderivatives
    (a) $\int \frac{52x^2+68x+46+15x^3}{(x+1)^2(5x^2+10x+8)}\,dx$
    (b) $\int \frac{9x^2-42x+38}{(3x+2)(3x^2-12x+14)}\,dx$
    (c) $\int \frac{9x^2-6x+19}{(3x+1)(3x^2-6x+5)}\,dx$

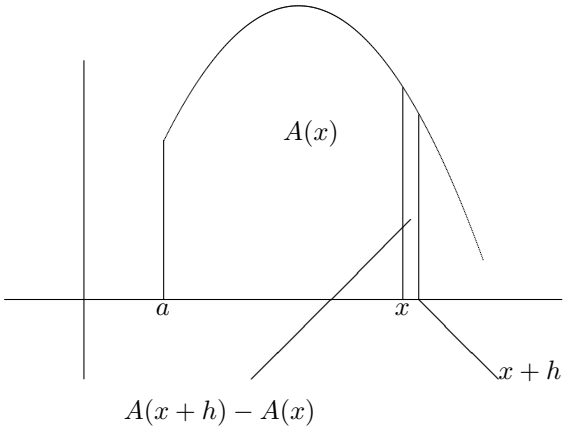
11. Solve the initial value problem, $y' = f(x), y(0) = 1$ for $f(x)$ equal to each of the integrands in Poblem 10.

12. Find the antiderivatives.
    (a) $\int \frac{1}{(3x^2+12x+13)^2}\,dx$
    (b) $\int \frac{1}{(5x^2+10x+7)^2}\,dx$
    (c) $\int \frac{1}{(5x^2-20x+23)^2}\,dx$

13. Solve the initial value problem, $y' = f(x), y(0) = 1$ for $f(x)$ equal to each of the integrands in Poblem 12.

9.11 Areas

Consider the problem of finding the area between the graph of a function of one variable and the x axis as illustrated in the following picture.



The curved line on the top represents the graph of the function $y = f(x)$ and the symbol, $A(x)$ represents the area between this curve and the x axis between the point, a and the

point x as shown. The vertical line from the point, $x+h$ up to the curve and the vertical line from x up to the curve define the area, $A(x+h)-A(x)$ as indicated. You can see that this area is between $hf(x)$ and $hf(x+h)$. This happens because the function is decreasing near x. In general, for continuous functions, f, Theorem 5.7.10 on Page 107 implies there exists x_M , $x_m \in [x, x+h]$ with the properties

$$f(x_M) \equiv \max\{f(x) : x \in [x, x+h]\}$$

and

$$f(x_m) \equiv \min\{f(x) : x \in [x, x+h]\}$$

Then,

$$f(x_m) = \frac{hf(x_m)}{h} \leq \frac{A(x+h)-A(x)}{h} \leq \frac{hf(x_M)}{h} = f(x_M).$$

Therefore, using the squeezing theorem, Theorem 5.9.5, and the continuity of f,

$$A'(x) \equiv \lim_{h\to 0} \frac{A(x+h)-A(x)}{h} = f(x).$$

The consideration of $h<0$ is also straightforward. This discussion implies the following theorem.

Theorem 9.11.1 *Let $a<b$ and let $f:[a,b]\to[0,\infty)$ be continuous. Then letting $A(x)$ denote the area between a, x, the graph of the function, and the x axis,*

$$A'(x) = f(x) \text{ for } x \in (a,b),\ A(a) = 0. \tag{9.9}$$

Also, A is continuous on $[a,b]$.

The problem for A described in the above theorem is called an initial value problem and the equation, $A'(x) = f(x)$ is a differential equation. It is called this because it is an equation for an unknown function, $A(x)$ written in terms of the derivative of this unknown function. The assertion that A should be continuous on $[a,b]$ follows from the fact that it has to be continuous on (a,b) because of the existence of its derivative and the above argument can also be used to obtain one sided derivatives for A at the end points, a and b, which yields continuity on $[a,b]$.

Example 9.11.2 *Let $f(x) = x^2$ for $x \in [1,2]$. Find the area between the graph of the function, the points 1 and 2, and the x axis.*

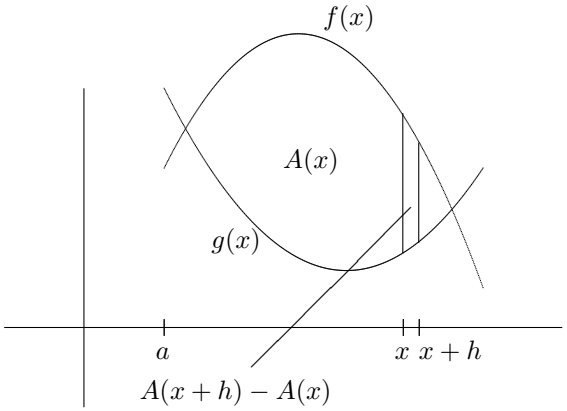
The function, $\frac{x^3}{3} + C$ has the property that its derivative gives x^2. This is true for any C. It only remains to choose C in such a way that the function equals zero at $x=1$. Thus $C = \frac{-1}{3}$. It follows that $A(x) = \frac{x^3}{3} - \frac{1}{3}$. Therefore, the area described equals $A(2) = \frac{8}{3} - \frac{1}{3} = \frac{7}{3}$ square units.

Example 9.11.3 *Find the area between the graph of the function $y = 1/x^2$ and the x axis for x between $1/2$ and 3.*

The function $-\frac{1}{x} + C$ has the property that its derivative equals $1/x^2$. Letting $C=2$, $A(x) = -\frac{1}{x} + 2$ satisfies the appropriate initial value problem and so the area equals $A(3) = \frac{5}{3}$.

9.12 Area Between Graphs

It is a minor generalization to consider the area between the graphs of two functions. Consider the following picture.



You see that sometimes the function, $f(x)$ is on top and sometimes the function, $g(x)$ is on top. It is the length of the vertical line joining the two graphs which is of importance and this length is always $|f(x) - g(x)|$ regardless of which function is larger. By Theorem 5.7.10 on Page 107 there exist x_M, $x_m \in [x, x+h]$ satisfying

$$|f(x_M) - g(x_M)| \equiv \max\{|f(x) - g(x)| : x \in [x, x+h]\}$$

and

$$|f(x_m) - g(x_m)| \equiv \min\{|f(x) - g(x)| : x \in [x, x+h]\}.$$

Then

$$\frac{|f(x_m) - g(x_m)|\, h}{h} \leq \frac{A(x+h) - A(x)}{h} \leq \frac{|f(x_M) - g(x_M)|\, h}{h},$$

and using the squeezing theorem, Theorem 5.9.5 on Page 111, as $h \to 0$

$$A'(x) = |f(x) - g(x)|$$

Also $A(a) = 0$ as before. This yields the following theorem which generalizes the one presented earlier because the x axis is the graph of the function, $y = 0$.

Theorem 9.12.1 *Let $a < b$ and let $f, g : [a,b] \to \mathbb{R}$ be continuous. Then letting $A(t)$ denote the area between the graphs of the two functions for $x \in [a,t]$,*

$$A'(t) = |f(t) - g(t)| \text{ for } t \in (a,b),\ A(a) = 0. \tag{9.10}$$

Also, A is continuous on $[a,b]$.

This theorem provides the justification for the following procedure for finding the area between the graphs of two functions which also applies to more general situations than described in the above theorem.

Procedure 9.12.2 *To find the area between the graphs of the functions $y = f(x)$ and $y = g(x)$ for $x \in [a,b]$, split the interval into non overlapping subintervals, $I_1, I_2, \cdots, I_k$ which have the property that on I_i, $|f(x) - g(x)|$ equals either $f(x) - g(x)$ or $g(x) - f(x)$.*

If $I_i = [p_i, q_i]$, take an antiderivative of $|f(x) - g(x)|$ on $[p_i, q_i], H_i$. (This might be a reasonable problem because on this interval, you won't need to write in absolute value signs.) Then the area between the curves for $x \in I_i$ is $H_i(q_i) - H_i(p_i)$. The desired area is the sum of these.

Proof: Consider the area between the curves for $x \in [p_i, q_i]$. You need $A(q_i)$ where $A' = |f(x) - g(x)|$ and $A(p_i) = 0$. Let H be any antiderivative. Then from Lemma 9.1.3, $A(x) = H(x) + C$ where C is some constant. Thus $C = -H(p_i)$ and so $A(q_i) = H(q_i) - H(p_i)$.

Example 9.12.3 *Let $f(x) = 8 - \frac{x^2}{2}$ and $g(x) = \frac{x^2}{2} - 1$. Find the area between the graphs of the two functions for $x \in [-4, 3]$.*

You should graph the two functions. $|f(x) - g(x)| =$

$$\left|9 - x^2\right| = \begin{cases} x^2 - 9 \text{ if } x \in [-4, -3] \\ 9 - x^2 \text{ if } x \in [-3, 3] \end{cases}$$

It follows that on $(-4, -3)$, an antiderivative is $H(x) = \frac{x^3}{3} - 9x$. Therefore, the area between the curves for $x \in [-4, -3]$ is

$$\left(\frac{(-3)^3}{3} - 9(-3)\right) - \left(\frac{(-4)^3}{3} - 9(-4)\right) = \frac{10}{3}$$

Now consider $x \in [-3, 3]$. On this interval, an antiderivative is $H(x) = 9x - \frac{x^3}{3}$ and so the area between the curves for x in this interval is

$$\left(9(3) - \frac{(3)^3}{3}\right) - \left(9(-3) - \frac{(-3)^3}{3}\right) = 36$$

The total area is the sum of these. Thus the total area is $36 + \frac{10}{3} = \frac{118}{3}$.

A similar procedure holds for finding the area between two functions which are of the form $x = f(y)$ and $x = g(y)$ for $y \in [c, d]$. You just let y play the role of x in the above.

Example 9.12.4 *Find the area between $x = 4 - y^2$ and $x = -3y$.*

First find where the two graphs intersect. $4 - y^2 = -3y$. The solution is $y = -1$ and 4. For y in this interval, you can verify that $4 - y^2 > -3y$ and so $\left|4 - y^2 - (-3y)\right| = 4 - y^2 + 3y$. An antiderivative is $4y - \frac{y^3}{3} + \frac{3y^2}{2}$ and so the desired area is

$$4(4) - \frac{(4)^3}{3} + \frac{3(4)^2}{2} - \left(4(-1) - \frac{(-1)^3}{3} + \frac{3(-1)^2}{2}\right) = \frac{125}{6}.$$

9.13 Exercises

1. Find the area between the graphs of the functions, $y = x^2 + 1$ and $y = 3x + 5$.
2. Find the area between the graphs of the functions, $y = 2x^2$ and $y = 6x + 8$.
3. Find the area between the graphs of $y = 5x + 14$ and $y = x^2$.
4. Find the area between the graphs of the functions $y = x + 1$ and $y = 2x$ for $x \in [0, 3]$.

5. Find the area between $y = |x|$ and the x axis for $x \in [-2, 2]$.

6. Find the area between the graphs of $y = x$ and $y = \sin x$ for $x \in \left[-\frac{\pi}{2}, \pi\right]$.

7. Find the area between the graphs of $x = y^2$ and $y = 2 - x$.

8. Find the area between the x axis and the graph of the function $2x\sqrt{1+x^2}$ for $x \in [0, 2]$. **Hint:** Recall the chain rule for derivatives.

9. Show that the area of a right triangle equals one half the product of two sides which are not the hypotenuse.

10. Let A denote the region between the x axis and the graph of the function, $f(x) = x - x^2$. For $k \in (0, 1)$, the line $y = kx$ divides this region into two pieces. Explain why there exists a number, k such that the area of these two pieces is exactly equal. **Hint:** This will likely involve the intermediate value theorem. Write an equation satisfied by k and then find an approximate value for k. **Hint:** You should draw plenty of pictures to do this last part.

11. Find the area between the graph of $f(x) = 1/x$ for $x \in [1, 2]$ and the x axis in terms of known functions.

12. Find the area between the graph of $f(x) = 1/x^2$ for $x \in [1, 2]$ and the x axis in terms of known functions.

13. Find the area between $y = \sin x$ and $y = \cos x$ for $x \in \left[0, \frac{\pi}{4}\right]$.

14. Find the area between e^x and $\cos x$ for $x \in [0, 2\pi]$.

15. Find the area between the graphs of $y = \sin(2x)$ and $y = \cos(2x)$ for $x \in [0, 2\pi]$.

16. Find the area between the graphs of $y = \sin^2 x$, and the x axis, for $x \in [0, 2\pi]$.

17. Find the area between the graphs of $y = \cos^2 x$ and $y = \sin^2 x$ for $x \in [0, \pi/2]$.

18. Find the area between the graph of $f(x) = x^3/\left(2 + 3x^4\right)$ and the x axis for $x \in [0, 4]$.

19. Find the area between the graph of $f(x) = x^3 \sin\left(x^2\right) + 30$ for $x \in [0, \pi]$.

20. Find the area between the graph of $f(x) = x \sin 2x$ and $x \cos x - 4$ for $x \in [0, \pi]$.

21. Find the area between $y = \frac{3x^3+9x^2+10x+6}{(x+1)^2(x^2+2x+2)}$ and the x axis for $x \in [0, 4]$.

22. Find the area between $y = \frac{5x^2+4+8x+2x^4+3x^3}{(2x+3)(x^2+1)}$ and the x axis for $x \in [0, 2]$.

23. Find the area between $\arctan(x)$ and $\ln x$ for $x \in [0, b]$ where $\arctan(b) - \ln(b) = 0$ and b is the first positive number for which this is so. $\arctan(b) - \ln(b) = 0$. You will need to use Newton's method or graphing on a calculator to find b.

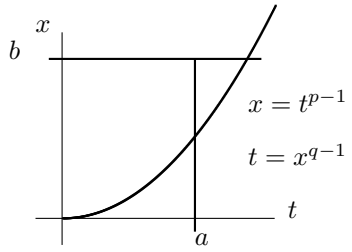
24. Find the area between $y = e^x$ and $y = 2x + 1$ for $x > 0$. In order to do this, you have to find a solution to $e^x = 2x + 1$ and this will require a numerical procedure such as Newton's method or graphing and zooming on your calculator.

25. Find the area between $y = \ln x$ and $y = \sin x$ for $x > 0$. In order to do this, you have to find a solution to $\ln x = \sin x$ and this will require a numerical procedure such as Newton's method or graphing and zooming on your calculator.

26. Let $p > 1$. An inequality which is of major importance is

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

where here q is defined by $1/p + 1/q = 1$. Establish this inequality by adding up areas in the following picture.



In the picture the right side of the inequality represents the sum of all the areas and the left side is the area of the rectangle determined by $(a, 0)$ and $(0, b)$.

9.14 Practice Problems For Antiderivatives

The process of finding antiderivatives has absolutely nothing to do with mathematics. However, it is fun and it is good to become adept at doing it. This can only be accomplished through working lots of problems. Here are lots of practice problems for finding antiderivatives. Some of these are very hard but you don't have to do all of them if you don't want to. However, the more you do, the better you will be at taking antiderivatives. Most of these problems are modifications of problems I found in a Russian calculus book. This book had some which were even harder. I shall give answers to these problems so you can see whether you have it right. Beware that sometimes you may get it right even though it looks different than the answer given. Also, there is no guarantee that my answers are right.

1. Find $\int \frac{\sqrt{2x}+1}{x}\,dx$.

   Answer:

   $\int \frac{\sqrt{2x}+1}{x}\,dx = 2\sqrt{2}\sqrt{x} + \ln x + C$

2. Find $\int \frac{te^t}{(t+1)^2}\,dt$ **Hint:** Write this as $\int \left(\frac{(1+t)e^t}{(t+1)^2} - \frac{e^t}{(1+t)^2}\right) dt$

   $= \int \left(\frac{e^t}{(t+1)} - \frac{e^t}{(1+t)^2}\right) dt.$

   Answer: $\frac{e^t}{t+1} + C$

3. Find $\int \frac{5-2x^2}{5+2x^2}\,dx$. **Hint:** $\frac{5-2x^2}{5+2x^2} = -1 + \frac{10}{5+2x^2}$.

   Answer:

   $\int \frac{5-2x^2}{5+2x^2}\,dx = -x + \sqrt{10}\arctan\frac{1}{5}x\sqrt{10} + C$

4. Find $\int \left(3-x^5\right)^2\,dx$.

   Answer:

   $\int \left(3-x^5\right)^2\,dx = \frac{1}{11}x^{11} - x^6 + 9x + C$

5. Find $\int (2x+3)^{-25}\,dx$. **Hint:** Let $u = 2x+3$

   Answer:

   $\int (2x+3)^{-25}\,dx = -\frac{1}{48(2x+3)^{24}} + C$

6. Find $\int 5^x\,dx$.

   Answer:

   $\int 5^x\,dx = \frac{1}{\ln 5}5^x + C$

7. Find $\int \cosh^2 8x\,dx$. **Hint:** Try integration by parts.

   Answer:

   $\int \cosh^2 8x\,dx = \frac{1}{16}\cosh 8x \sinh 8x + \frac{1}{2}x + C$

8. Find $\int \tanh^2 2x\,dx$.

   Answer:

   $\int \tanh^2 2x\,dx =$

   $-\frac{1}{2}\tanh 2x - \frac{1}{4}\ln(-1+\tanh 2x)$

   $+\frac{1}{4}\ln(1+\tanh 2x) + C$

9. Find $\int \cosh(3x+3)\,dx$.

   Answer:

   $\int \cosh(3x+3)\,dx = \frac{1}{3}\sinh(3x+3) + C$

10. Find $\int 8^{1+x}\,dx$. **Hint:** This equals $8\int e^{x\ln 8}dx$.

    Answer:

    $\int 8^{1+x}\,dx = \frac{8}{\ln 8}8^x + C$

11. Find $\int \frac{\sqrt{(36+x^2)}+\sqrt{(36-x^2)}}{\sqrt{(1296-x^4)}}\,dx$. **Hint:** The integrand equals $\frac{\sqrt{(36+x^2)}+\sqrt{(36-x^2)}}{\sqrt{(36-x^2)(36+x^2)}}$

    $= \frac{1}{\sqrt{36-x^2}} + \frac{1}{\sqrt{36+x^2}}$.

    Answer:

    $\int \frac{\sqrt{36+x^2}+\sqrt{36-x^2}}{\sqrt{1296-x^4}}\,dx = \int \frac{1}{\sqrt{36-x^2}}\,dx +$

    $\int \frac{1}{\sqrt{(36+x^2)}}\,dx = \arcsin\frac{1}{6}x +$

    $\ln\left(x+\sqrt{36+x^2}\right) + C$

12. Find $\int (7x+1)^{30}\,dx$.

    Answer:

    $\int (7x+1)^{30}\,dx = \frac{1}{217}(7x+1)^{31} + C$

13. Find $\int \sqrt{1+\sin 3x}\,dx$. **Hint:** The integrand equals $\frac{\cos(3x)}{\sqrt{1-\sin(3x)}}$.

    Answer:

    $\int \sqrt{1+\sin 3x}\,dx =$

    $\frac{2}{3}(\sin 3x - 1)\frac{\sqrt{(1+\sin 3x)}}{\cos 3x} + C$

14. Find $\int \left(\sqrt{3+2x}\right)^5\,dx$.

    Answer:

    $\int \left(\sqrt{3+2x}\right)^5\,dx = \frac{1}{7}\left(\sqrt{3+2x}\right)^7 + C$

15. Find $\int \frac{1}{49+4x^2}\,dx$.

    Answer:

    $\int \frac{1}{49+4x^2}\,dx = \frac{1}{14}\arctan\frac{2}{7}x + C$

16. Find $\int \frac{1}{\sqrt{4x^2-9}}\,dx$.

    Answer:

    $\int \frac{1}{\sqrt{4x^2-9}}\,dx =$

    $\frac{1}{2}\ln\left(2x+\sqrt{4x^2-9}\right) + C$

17. Find $\int \frac{1}{\sin^2(2x+3)}\,dx$.

    Answer:

    $\int \frac{1}{\sin^2(2x+3)}\,dx =$

    $-\frac{1}{2\sin(2x+3)}\cos(2x+3) + C$

18. Find $\int \frac{1}{1+\cos 6x}\,dx$. **Hint:** You could let $u = 6x$ and then use the technique for rational functions of $\cos x$ and $\sin x$.

    Answer:

    $\int \frac{1}{1+\cos(6x)}\,dx = \frac{1}{6}\tan 3x + C.$

19. Find $\int \frac{1}{1+\sin(3x)}\,dx$.

    Answer:

    $\int \frac{1}{1+\sin 3x}\,dx = -\frac{2}{3\left(1+\tan\frac{3}{2}x\right)} + C$

20. Find $\int x^2\sqrt{4x^3+2}\,dx$. **Hint:** Let $u = 4x^3 + 2$.

    Answer:

    $\int x^2\sqrt{4x^3+2}\,dx =$

    $\frac{1}{18}\left(\sqrt{4x^3+2}\right)^3 + C$

21. Find $\int x^8\left(\sqrt{3x^9+2}\right)^5\,dx$.

    Answer:

    $\int x^8\left(\sqrt{3x^9+2}\right)^5\,dx$

    $= \frac{2}{189}\left(\sqrt{3x^9+2}\right)^7 + C$

22. Find $\int \frac{x}{3+8x^4}\,dx$. **Hint:** Try $u = x^2$.

    Answer:

    $\int \frac{x}{3+8x^4}\,dx =$

    $\frac{1}{24}\sqrt{6}\arctan\left(\frac{2}{3}x^2\sqrt{6}\right) + C$

23. Find $\int \frac{1}{3(2+x)\sqrt{x}}\,dx$. **Hint:** Try $x = u^2$.

    Answer:

    $\int \frac{1}{3(2+x)\sqrt{x}}\,dx =$

    $\frac{1}{9}\sqrt{3}\sqrt{6}\arctan\left(\frac{1}{6}\sqrt{3}\sqrt{x}\sqrt{6}\right) + C$

24. Find $\int \frac{1}{x\sqrt{(25x^2-9)}}\,dx$.

    Answer:

    You could let $3\sec u = 5x$ so

    $(3\sec u\tan u)\,du = 5\,dx$ and then

    $\int \frac{1}{x\sqrt{(25x^2-9)}}\,dx = \frac{1}{3}\int du = \frac{u}{3} + C.$

    Now restoring the original variables, this yields $\frac{1}{3}\operatorname{arcsec}\frac{5}{3}|x| + C$.

25. Find $\int \frac{1}{\sqrt{(x(3x-5))}}\,dx$. **Hint:** You might try completing the square in $x(3x-5)$ and then changing the variable in an appropriate manner.

    Answer:

    $\int \frac{dx}{\sqrt{(x(3x-5))}} =$

    $\frac{1}{3}\sqrt{3}\ln\left(\sqrt{3}\left(x-\frac{5}{6}\right) + \sqrt{(3x^2-5x)}\right) +$

    C

26. Find $\int \frac{1}{x\cos(\ln 4x)}\,dx$. **Hint:** You might try letting $u = \ln(4x)$.

    Answer:

    $\int \frac{1}{x\cos(\ln 4x)}\,dx =$

    $\ln(\sec(\ln 4x) + \tan(\ln 4x)) + C$

27. Find $\int x^7 e^{x^8}\,dx$.

    Answer:

    $\int x^7 e^{x^8}\,dx = \frac{1}{8}e^{x^8} + C$

28. Find $\int \frac{\ln^4 x}{x}\,dx$.

    Answer:

    $\int \frac{\ln^4 x}{x}\,dx = \frac{1}{5}\ln^5 x + C$

29. Find $\int \frac{1}{x(2+\ln 2x)}\,dx$.

    Answer:

    $\int \frac{1}{x(2+\ln 2x)}\,dx =$

    $\ln(2+\ln 2x) + C$

30. Find $\int \frac{\sin 7x+\cos 7x}{\sqrt{\sin 7x-\cos 7x}}\,dx$. **Hint:** Try $u = \sin 7x - \cos 7x$.

    Answer:

    $\int \frac{\sin 7x+\cos 7x}{\sqrt{(\sin 7x-\cos 7x)}}\,dx =$

    $\frac{2}{7}\sqrt{(\sin 7x - \cos 7x)} + C$

31. Find $\int \csc 4x\,dx$.

    Answer:

    $\int \csc 4x\,dx =$

    $\frac{1}{4}\ln|\csc 4x - \cot 4x| + C$

32. Find $\int \frac{\arctan 5x}{1+25x^2}\,dx$. **Hint:** Try $u = \arctan(5x)$.

    Answer:

    $\int \frac{\arctan 5x}{1+25x^2}\,dx = \frac{1}{10}\arctan^2 5x + C$

33. Find $\int \frac{18}{(6+2x)(6-x)}\cos\left(\ln\frac{6+2x}{6-x}\right)\,dx$. **Hint:** It might help to first let $u = \ln\frac{6+2x}{6-x}$ and see if it simplifies.

    Answer:

    $\int \frac{18}{(6+2x)(6-x)}\cos\left(\ln\frac{6+2x}{6-x}\right)\,dx =$

    $\sin\left(\ln\frac{6+2x}{6-x}\right) + C$

34. Find $\int x^{23}\left(2-6x^{12}\right)^{10}\,dx$ **Hint:** Maybe let $u=2-6x^{12}$.

Answer:

$\int x^{23}\left(2-6x^{12}\right)^{10}\,dx=$

$\frac{1}{5184}\left(2-6x^{12}\right)^{12}-\frac{1}{2376}\left(2-6x^{12}\right)^{11}+C$

35. Find $\int \frac{x^5}{\sqrt{6-3x^2}}\,dx$.

Answer:

$\int \frac{x^5}{\sqrt{(6-3x^2)}}\,dx=$

$-\frac{1}{15}x^4\sqrt{(6-3x^2)}-\frac{8}{45}x^2\sqrt{(6-3x^2)}$

$-\frac{32}{45}\sqrt{(6-3x^2)}+C$

36. Find $\int \cos^3(3x)\sin^{\frac{1}{2}}(3x)\,dx$.

Answer:

$\int \cos^3 3x\sin^{\frac{1}{2}}3x\,dx=$

$\int\left(\cos 3x\left(1-\sin^2 3x\right)\sin^{\frac{1}{2}}3x\right)\,dx$

$=-\frac{2}{21}\sin^{\frac{7}{2}}3x+\frac{2}{9}\sin^{\frac{3}{2}}3x+C$

37. Find $\int \frac{1}{e^{2x}+e^x}\,dx$. **Hint:** Try $u=e^x$.

Answer:

$\int \frac{1}{e^{2x}+e^x}\,dx=\frac{-1-xe^x}{e^x}+\ln\left(e^x+1\right)+C$

38. Find $\int \frac{1}{\sqrt{e^x+1}}\,dx$.

Answer:

Let $u^2=e^x$ so $2udu=e^xdx=au^2dx$. In terms of u this is

$2\int \frac{1}{u\sqrt{1+u^2}}\,du$. Now let

$u=\tan\theta$ so $du=\left(\sec^2\theta\right)d\theta$. Then the indefinite integral becomes

$2\int \frac{\sec^2\theta}{\tan\theta\sec(\theta)}\,d\theta=2\int\csc\theta\,d\theta=$

$2\ln|\csc\theta-\cot\theta|+C.$

In terms of u this is

$2\ln\left|\frac{\sqrt{u^2+1}}{u}-\frac{1}{u}\right|+C$ and in terms of x this is

$2\ln\left|\frac{\sqrt{(e^x+1)}}{e^{\frac{1}{2}x}}-e^{-\frac{1}{2}x}\right|+C$.

39. Find $\int \frac{\arctan\sqrt{x}}{\sqrt{x}(1+x)}\,dx$.

Answer:

$\int \frac{\arctan\sqrt{x}}{\sqrt{x}(1+x)}=\arctan^2\sqrt{x}+C$

40. Find $\int\sqrt{\left(\frac{1+x}{1-x}\right)}\,dx$.

Answer:

Multiply the fraction on the top and bottom by $1+x$ to get

$\int \frac{1+x}{\sqrt{1-x^2}}\,dx$. Now let $x=\sin\theta$ so $dx=\cos\theta\,d\theta$. Then this is

$\int \frac{1+\sin\theta}{\sqrt{1-\sin^2\theta}}\cos\theta\,d\theta=$

$\int(1+\sin\theta)\,d\theta=\theta-\cos\theta+C.$

In terms of x this gives

$\arcsin x-\sqrt{1-x^2}+C.$

41. Find $\int\sqrt{\frac{x-2}{x+2}}\,dx$.

Answer:

Multiply the fraction on the top and bottom by $x+2$ to get $\int \frac{\sqrt{x^2-4}}{x+2}\,dx$ Now let $x=2\sec\theta$ so $dx=2\sec\theta\tan\theta\,d\theta$ and in terms of θ the indefinite integral is

$2\int \frac{\sqrt{\sec^2(\theta)-1}}{\sec\theta+1}\sec\theta\tan\theta\,d\theta=$

$2\int \frac{\tan^2\theta\sec(\theta)}{1+\sec\theta}\,d\theta$

$=2\int \frac{\tan^2\theta}{1+\cos\theta}=2\int\sec^2\theta\,d\theta-2\int\sec\theta\,d\theta=$

$2\tan\theta-2\ln(\sec\theta+\tan\theta)+C$. Now in terms of x this is

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

$2\ln\left(\frac{1}{2}x+\frac{1}{2}\sqrt{(x^2-4)}\right)+C.$

42. Find $\int \frac{x^2}{\sqrt{(4+9x^2)}}\,dx$.

Answer:

$\int \frac{x^2}{\sqrt{4+9x^2}}\,dx=$

$\frac{1}{18}x\sqrt{4+9x^2}-$

$\frac{2}{27}\ln\left(3x+\sqrt{4+9x^2}\right)+C$

43. Find $\int \frac{1}{\sqrt{x^2+49}}\,dx$.

Answer:

$\int \frac{1}{\sqrt{x^2+49}}\,dx=\ln\left(x+\sqrt{x^2+49}\right)+C$

44. Find $\int \frac{1}{\sqrt{x^2-36}}\,dx$.

Answer:

$\int \frac{1}{\sqrt{x^2-36}}\,dx=\ln\left(x+\sqrt{x^2-36}\right)+C$

45. Find $\int x \ln(3x)\ dx$.

    Answer:

    $\int x \ln(3x)\ dx = \frac{1}{2}x^2 \ln 3x - \frac{1}{4}x^2 + C$

46. Find $\int x \ln^2(6x)\ dx$.

    Answer:

    $\int x \ln^2 6x\ dx = \frac{1}{2}x^2 \ln^2 6x - \frac{1}{2}x^2 \ln 6x + \frac{1}{4}x^2 + C$

47. Find $\int x^3 e^{x^2}\ dx$.

    Answer:

    $\int x^3 e^{x^2}\ dx = \frac{1}{2}x^2 e^{x^2} - \frac{1}{2}e^{x^2} + C$

48. Find $\int x^2 \sin 8x\ dx$

    Answer:

    $\int x^2 \sin 8x\, dx =$

    $-\frac{1}{8}x^2 \cos 8x + \frac{1}{256}\cos 8x + \frac{1}{32}x \sin 8x + C$

49. Find $\int \arcsin x\, dx$

    Answer:

    $\int \arcsin x\ dx$

    $= x \arcsin x + \sqrt{(1-x^2)} + C$

50. Find $\int \arctan x\, dx$

    Answer:

    $\int \arctan x\, dx$

    $= x \arctan x - \frac{1}{2}\ln\left(1+x^2\right) + C$

51. Find $\int \frac{\arctan 6x}{x^3}\, dx$. **Hint:** Do integration by parts on this one. This will get things started. Then recall partial fractions.

    Answer:

    $\int \frac{\arctan 6x}{x^3}\, dx$

    $= -\frac{1}{2x^2}\arctan 6x - \frac{3}{x} - 18 \arctan 6x + C$

52. Find $\int \sin(4x) \ln(\tan 4x)\ dx$

    Answer:

    Integration by parts gives

    $-\frac{1}{4}\cos 4x \ln(\tan 4x) +$

    $\frac{1}{4}\ln(\csc 4x - \cot 4x) + C$

53. Find $\int e^{2x}\sqrt{e^{4x}+1}\, dx$. **Hint:** Try $u = e^{2x}$. This will yield something which will look a little different than the answer I have given below.

    Answer:

    $\frac{1}{2}\int \sqrt{u^2+1} du$

    $\int e^{2x}\sqrt{(e^{4x}+1)}\, dx$

    $= \frac{1}{4}e^{2x}\sqrt{(e^{4x}+1)} + \frac{1}{4}\operatorname{arcsinh}\left(e^{2x}\right) + C$

54. Find $\int \cos(\ln 7x)\ dx$.

    Answer:

    $\int \cos(\ln 7x)\ dx = x\cos(\ln 7x)$

    $+ \int x \sin(\ln(7x))\left(\frac{1}{x}\right)\ dx$

    $= x\cos(\ln 7x) +$

    $\left[x \sin(\ln 7x) - \int \cos(\ln(7x))\ dx\right]$ and so

    $\int \cos(\ln 7x)\ dx =$

    $\frac{1}{2}\left[x\cos(\ln 7x) + x\sin(\ln 7x)\right] + C$

55. Find $\int \sin(\ln 7x)\,.$

    Answer:

    $\int \sin(\ln 7x)\ dx$

    $= \frac{1}{2}\left[x\sin(\ln 7x) - x\cos(\ln 7x)\right] + C$

56. Find $\int e^{5x}\sin 3x\, dx$.

    Answer:

    $\int e^{5x}\sin 3x\, dx$

    $= -\frac{3}{34}e^{5x}\cos 3x + \frac{5}{34}e^{5x}\sin 3x + C$

57. Find $\int e^{3x}\sin^2 2x\, dx$.

    Answer:

    $\int e^{3x}\sin^2 2x\, dx$

    $= \frac{1}{6}e^{3x} - \frac{3}{25}e^{3x}(\cos 2)\, x - \frac{4}{25}e^{3x}(\sin 2)\, x + C$

58. Find $\int \frac{x+4}{(x+3)^2}\, dx$.

    Answer:

    $\int \frac{x+4}{(x+3)^2}\, dx =$

    $\ln(x+3) - \frac{1}{x+3} + C$

59. Find $\int \frac{7x^2+100x+347}{(x+2)(x+7)^2}\, dx$.

    Answer:

    $\int \frac{7x^2+100x+347}{(x+2)(x+7)^2}\, dx$

    $= 7\ln(x+2) - \frac{2}{x+7} + C$

60. Find $\int \frac{3x^2+7x+8}{(x+2)(x^2+2x+2)}\, dx$.

    Answer:

    $\int \frac{3x^2+7x+8}{(x+2)(x^2+2x+2)}\, dx$

    $= 3\ln(x+2) + \arctan(1+x) + C$

61. Find $\int \frac{4x^2+65x+266}{(x+2)(x^2+16x+66)}\,dx$.

Answer:

$\int \frac{4x^2+65x+266}{(x+2)(x^2+16x+66)}\,dx$

$= 4\ln(2+x)+\frac{1}{2}\sqrt{2}\arctan\frac{1}{4}(2x+16)\sqrt{2}$

62. Find $\int \frac{1}{x^3+8}\,dx$. **Hint:** You need to first factor x^3+8 and then use partial fractions.

Answer:

$\int \frac{1}{x^3+8}\,dx =$

$\frac{1}{12}\ln(x+2)-\frac{1}{24}\ln\left(x^2-2x+4\right)$

$+\frac{1}{12}\sqrt{3}\arctan\frac{1}{6}(2x-2)\sqrt{3}+C$

63. Find $\int \frac{1}{x^4+x^2+1}\,dx$.

Answer:

$\int \frac{1}{x^4+x^2+1}\,dx =$

$\frac{1}{4}\ln\left(x^2+x+1\right)+$

$\frac{1}{6}\sqrt{3}\arctan\frac{1}{3}(2x+1)\sqrt{3}$

$-\frac{1}{4}\ln\left(x^2-x+1\right)+$

$\frac{1}{6}\sqrt{3}\arctan\frac{1}{3}(2x-1)\sqrt{3}+C$

64. Find $\int \frac{1}{x^4+81}\,dx$. **Hint:**

$x^4+81 =$

$\left(x^2-3x\sqrt{2}+9\right)\left(x^2+3x\sqrt{2}+9\right).$

Answer:

Now you can use partial fractions to find

$\int \frac{1}{x^4+81}\,dx =$

$\frac{1}{216}\sqrt{2}\ln\frac{x^2+3x\sqrt{2}+3}{x^2-3x\sqrt{2}+3}+$

$\frac{1}{108}\sqrt{2}\arctan\left(\frac{1}{3}x\sqrt{2}+1\right)$

$+\frac{1}{108}\sqrt{2}\arctan\left(\frac{1}{3}x\sqrt{2}-1\right)+C$

65. Find $\int \frac{1}{x^6+64}\,dx$.

Answer:

First factor the denominator.

$x^6+64 =$

$\left(x^2+4\right)\left(\left(x-\sqrt{3}\right)^2+1\right)\left(\left(x+\sqrt{3}\right)^2+1\right).$

$\frac{1}{x^6+64} = \frac{Ax+B}{x^2+4}+\frac{Cx+D}{\left(x-\sqrt{3}\right)^2+1}+\frac{Ex+F}{\left(x+\sqrt{3}\right)^2+1}$

and so after wading through much affliction, the partial fractions decomposition is

$\frac{16}{3x^2+12}+\frac{-\frac{1}{192}\sqrt{3}x+\frac{1}{48}}{\left(x-\sqrt{3}\right)^2+1}+\frac{\frac{1}{192}\sqrt{3}x+\frac{1}{48}}{\left(x+\sqrt{3}\right)^2+1}$. Therefore, the indefinite integral is

$\int\left(\frac{16}{3x^2+12}+\frac{-\frac{1}{192}\sqrt{3}x+\frac{1}{48}}{\left(x-\sqrt{3}\right)^2+1}+\right.$

$\left.+\frac{\frac{1}{192}\sqrt{3}x+\frac{1}{48}}{\left(x+\sqrt{3}\right)^2+1}\right)dx =$

$\frac{1}{96}\arctan\frac{1}{2}x-\frac{1}{384}\sqrt{3}\ln\left(x^2-2\sqrt{3}x+4\right)$

$+\frac{1}{192}\arctan\left(x-\sqrt{3}\right)+$

$\frac{1}{384}\sqrt{3}\ln\left(x^2+2\sqrt{3}x+4\right)+$

$\frac{1}{192}\arctan\left(x+\sqrt{3}\right)+C$

66. Find $\int \frac{x^2}{(x-3)^{100}}\,dx$. **Hint:** You ought to let $u=x-3$.

Answer:

$\int \frac{x^2}{(x-3)^{100}}\,dx$

$= \frac{1}{11(3-x)^{99}}-\frac{3}{49(3-x)^{98}}+\frac{1}{97(3-x)^{97}}+C$

67. Find $\int 2\frac{4+4x+3x^2+x^3}{(x^2+1)(x^2+2x+5)}\,dx$.

Answer:

$\int 2\frac{4+4x+3x^2+x^3}{(x^2+1)(x^2+2x+5)}\,dx$

$= \frac{1}{2}\ln\left(x^2+1\right)+\arctan x+$

$\frac{1}{2}\ln\left(x^2+2x+5\right)+\arctan\left(\frac{1}{2}+\frac{1}{2}x\right)+C$

68. Find $\int \frac{1-x^7}{x(1+x^7)}\,dx$

Answer:

$\int \frac{1-x^7}{x(1+x^7)}\,dx$

$= \ln x-\frac{2}{7}\ln\left(1+x^7\right)+C$

69. Find $\int \frac{x^2+1}{x^4-2x^2+1}\,dx$

Answer:

$\int \frac{x^2+1}{x^4-2x^2+1}\,dx$

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

70. Find $\int \frac{x^5}{x^8+1}\,dx$

Answer:

Let $u=x^2$. $\int \frac{x^5}{x^8+1}\,dx$

$= \int \frac{u^2}{2u^4+2}\,du$. Now use partial fractions.

$\int \frac{x^5}{x^8+1}\,dx = \frac{1}{16}\sqrt{2}\ln\frac{x^4-x^2\sqrt{2}+1}{x^4+x^2\sqrt{2}+1}$

$+\frac{1}{8}\sqrt{2}\arctan\left(x^2\sqrt{2}+1\right)$

$+\frac{1}{8}\sqrt{2}\arctan\left(x^2\sqrt{2}-1\right)+C$

71. Find $\int \frac{x^2+4}{x^6+64}\,dx$.

Hint: $x^6+64 = \left(x^2+4\right)\left(x^4-4x^2+16\right)$.

Answer:

$\int \frac{x^2+4}{x^6+64}\,dx =$

$-\frac{1}{96}\sqrt{3}\ln\left(x^2-2\sqrt{3}x+4\right)+$

$\frac{1}{16}\arctan\left(x-\sqrt{3}\right)+$

$\frac{1}{96}\sqrt{3}\ln\left(x^2+2\sqrt{3}x+4\right)$

$+\frac{1}{16}\arctan\left(x+\sqrt{3}\right)+C$

72. Find $\int \frac{dx}{x\left(2+3\sqrt{x}+\sqrt[3]{x}\right)}$.

Answer:

$\int \frac{dx}{x\left(2+3\sqrt{x}+\sqrt[3]{x}\right)} = \int \frac{6}{u(2+3u^3+u^2)}\,du =$

$3\ln u-\frac{6}{7}\ln(1+u)-\frac{15}{14}\ln\left(3u^2-2u+2\right)$

$-\frac{3}{35}\sqrt{5}\arctan\frac{1}{10}(6u-2)\sqrt{5}+C$

$= 3\ln x^{1/6}-\frac{6}{7}\ln\left(1+x^{1/6}\right)-$

$\frac{15}{14}\ln\left(3x^{1/3}-2x^{1/6}+2\right)-$

$\frac{3}{35}\sqrt{5}\arctan\frac{1}{10}\left(6x^{1/6}-2\right)\sqrt{5}+C$

73. Find $\int \frac{\sqrt{x+3}-\sqrt{x-3}}{\sqrt{x+3}+\sqrt{x-3}}\,dx$.

Hint: Show the integrand equals $\frac{1}{3}x-\frac{1}{3}\sqrt{x^2-9}$.

Answer:

$\int \frac{\sqrt{x+3}-\sqrt{x-3}}{\sqrt{x+3}+\sqrt{x-3}}\,dx = \int\left(\frac{1}{3}x-\frac{1}{3}\sqrt{x^2-9}\right)dx$

$= \frac{1}{6}x^2-\frac{1}{6}x\sqrt{x^2-9}+\frac{3}{2}\ln\left(x+\sqrt{x^2-9}\right)+C$

74. Find $\int \frac{x^2}{\sqrt{x^2+6x+13}}\,dx$. **Hint:** Complete the square in the denominator.

Answer:

$\int \frac{x^2}{\sqrt{x^2+6x+13}}\,dx =$

$\frac{1}{2}x\sqrt{x^2+6x+13}-\frac{9}{2}\sqrt{x^2+6x+13}+$

$7\ln\left(x+3+\sqrt{x^2+6x+13}\right)+C$

75. Find $\int \frac{\sqrt{x^2+4x+5}}{x}\,dx$.

Answer:

$\int \frac{\sqrt{x^2+4x+5}}{x}\,dx =$

$\sqrt{x^2+4x+5}+2\operatorname{arcsinh}(x+2)$

$-\sqrt{5}\operatorname{arctanh}\frac{1}{5}(5+2x)\frac{\sqrt{5}}{\sqrt{x^2+4x+5}}+C$

You might try letting $u = \sinh^{-1}(x+2)$.

76. Find $\int \frac{1}{x^3\sqrt{x^2+9}}\,dx$.

Answer:

$\int \frac{1}{x^3\sqrt{x^2+9}}\,dx =$

$-\frac{1}{18x^2}\sqrt{x^2+9}+$

$\frac{1}{54}\operatorname{arctanh}\frac{3}{\sqrt{x^2+9}}+C$

You might try letting $u = \frac{1}{\sqrt{x^2+9}}$.

77. Find $\int \frac{dx}{x^4\sqrt{x^2-9}}$.

Answer:

$\int \frac{dx}{x^4\sqrt{x^2-9}} =$

$\frac{1}{27x^3}\sqrt{(x^2-9)}+\frac{2}{243x}\sqrt{(x^2-9)}+C$

78. Find $\int \sin^6(3x)\,dx$.

Answer:

$\int \sin^6(3x)\,dx =$

$-\frac{1}{2}\sin^5 x\cos x-\frac{5}{8}\sin^3 x\cos x$

$-\frac{15}{16}\cos x\sin x+\frac{5}{16}x+C$

79. Find $\int \frac{\sin^3(x)}{\cos^4(x)}\,dx$.

Answer:

$\int \frac{\sin^3(x)}{\cos^4(x)}\,dx =$

$\frac{1}{3}\frac{\sin^4 x}{\cos^3 x}-\frac{1}{3}\frac{\sin^4 x}{\cos x}-$

$\frac{1}{3}\sin^2 x\cos x-\frac{2}{3}\cos x+C$

80. Find $\int \tan^5(2x)\,dx$.

Answer:

$\int \tan^5(2x)\,dx =$

$-\frac{1}{4}\tan^2 2x+\frac{1}{8}\tan^4 2x$

$+\frac{1}{4}\ln\left(2+2\tan^2 2x\right)+C$

81. Find $\int \frac{dx}{\sqrt{\tan(2x)}}$.

Answer:

$\int \frac{dx}{\sqrt{\tan(2x)}} =$

$2\frac{\tan^{\frac{1}{2}}x}{1+2\tan^2 x}+2\frac{\tan^{\frac{5}{2}}x}{1+2\tan^2 x}$

$-2\tan^{\frac{1}{2}}x+\frac{1}{4}\sqrt{2}\arctan 2\sqrt{2}\frac{\tan^{\frac{1}{2}}x}{1-2\tan x}$

$+\frac{1}{4}\sqrt{2}\ln\frac{2\tan x+2\sqrt{2}\tan^{\frac{1}{2}}x+1}{\sqrt{(1+2\tan^2 x)}}+C$

You might try the substitution, $u = \tan(x)$.

82. Find $\int \frac{dx}{\cos x+3\sin x+4}$.

Answer:

$\int \frac{dx}{\cos x+3\sin x+4} =$

$\frac{1}{3}\sqrt{6}\arctan\frac{1}{12}\left(6\tan\frac{1}{2}x+6\right)\sqrt{6}+C$

Try the substitution, $u=\tan\left(\frac{x}{2}\right)$.

9.15 Volumes

9.15.1 Volumes Using Cross Sections

Imagine a building having height h and consider the intersection of the building with a plane which is parallel to the ground at height y. Let the area of this intersection, called the cross section at height y, equal $A(y)$. Then the volume of the building between the two parallel planes at height y and height $y+\Delta y$ would be approximately $\Delta y\,(A(y))$. Written in terms of differentials, $dV = A(y)\,dy$ and so the total volume of the building between 0 and y, $V(y)$, would satisfy the initial value problem,

$$\frac{dV}{dy} = A(y),\ V(0) = 0.$$

The volume of the building would be $V(h)$. This approach is completely general and can be used to find the volume of many different kinds of solids.

Example 9.15.1 *Consider a pyramid which sits on a square base of length 500 feet and suppose the pyramid has height 300 feet. Find the volume of the pyramid in cubic feet.*

At height, y, the length of one of the sides, x, would satisfy $\frac{x}{300-y} = \frac{500}{300} = \frac{5}{3}$and so $x = \frac{5}{3}(300-y)$. Therefore,

$$\frac{dV}{dy} = \left(\frac{5}{3}(300-y)\right)^2$$

and so $V(y) = -\frac{1}{5}\left(500-\frac{5}{3}y\right)^3 + C$. Since $V(0) = 0$, the constant must equal $\frac{1}{5}(500)^3$. Therefore, the total volume of the pyramid would equal

$$-\frac{1}{5}\left(500-\frac{5}{3}300\right)^3 + \frac{1}{5}(500)^3 = 25,\,000,\,000 \text{ cubic feet.}$$

Example 9.15.2 *The base of a solid is a circle of radius 10 meters in the xy plane. When this solid is cut with a plane which is perpendicular to the xy plane and the x axis, the result is a square. Find the volume of the resulting solid.*

Reasoning as above for the building and letting $V(x)$ denote the volume between -10 and x, yields

$$\frac{dV}{dx} = A(x),\ V(-10) = 0$$

as an appropriate differential equation for this volume. Here $A(x)$ is the area of the surface resulting from the intersection of the plane with the solid. The length of one side of this surface is $2\sqrt{100-x^2}$ because the equation of the circle which bounds the base is $x^2+y^2 = 100$. Therefore, $A(x) = 4\left(100-x^2\right)$ and so $V(x) = 400x - \frac{4}{3}x^3 + C$. To find C, use $V(-10) = 0$ and so $C = -400(-10) + \frac{4}{3}(-10)^3 = \frac{8000}{3}$. Therefore, the total volume is

$$V(10) = 400(10) - \frac{4}{3}(10)^3 + \frac{8000}{3} = \frac{16\,000}{3}$$

Example 9.15.3 *Find the volume of a sphere of radius R.*

The sphere is obtained by revolving a disk of radius R about the y axis. Thus the radius of the cross section at height y would be $\sqrt{R^2-y^2}$ and so the area of this cross section is $\pi\left(R^2-y^2\right)$. Therefore,

$$\frac{dV}{dy} = \pi\left(R^2-y^2\right),\ V(-R) = 0$$

and so $V(y) = \pi\left(R^2 y - \frac{1}{3}y^3\right) + C$. Now since $V(-R) = 0$, $C = \pi\left(-\frac{R^3}{3} + R^3\right)$ and so

$$V(R) = \pi\left(R^3 - \frac{1}{3}(R)^3\right) + \pi\left(-\frac{R^3}{3} + R^3\right) = \frac{4}{3}\pi R^3$$

Example 9.15.4 *The graph of the function, $f(x) = \sqrt{x}$ is revolved about the x axis. Find the volume of the resulting shape between $x = 0$ and $x = 8$.*

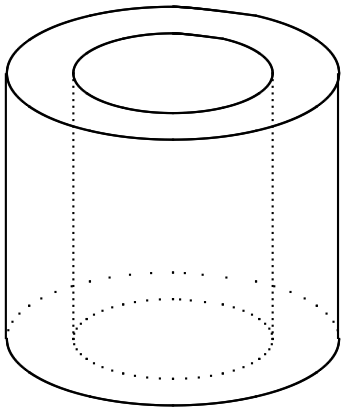
The cross sections perpendicular to the x axis for this shape are circles and the cross section at x has radius $\sqrt{x}$. Therefore, $A(x) = \pi x$ and the appropriate differential equation and initial value is

$$\frac{dV}{dx} = \pi x,\ V(0) = 0$$

and the answer is $V(8)$. From the differential equation and initial condition, $V(x) = \pi\frac{x^2}{2}$ and so $V(8) = \pi\frac{64}{2} = 32\pi$.

9.15.2 Volumes Using Shells

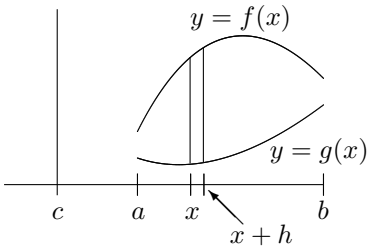
There is another way to find some volumes without using cross sections. This method involves the notion of shells. Consider the following picture of a circular shell.



In this picture the radius of the inner circle will be r and the radius of the outer circle will be $r + \Delta r$ while the height of the shell is H. Therefore, the volume of the shell would be the difference in the volumes of the two cylinders or

$$\pi(r + \Delta r)^2 H - \pi r^2 h = 2\pi H r(\Delta r) + \pi H(\Delta r)^2. \tag{9.11}$$

Now consider the problem of revolving the region between $y = f(x)$ and $y = g(x)$ for $x \in [a, b]$ about the line $x = c$ for $c < a$. The following picture is descriptive of the situation.



Let $V(x)$ denote the volume of the solid which results from revolving the region between the graphs of f and g above the interval, $[a,x]$ about the line $x=c$. Thus $V(x+h)-V(x)$ equals the volume which results from revolving the region between the graphs of f and g which is also between the two vertical lines shown in the above picture. This results in a solid which is very nearly a circular shell like the one shown in the previous picture and the approximation gets better as let h decreases to zero. Therefore,

$$\begin{aligned} V'(x) &= \lim_{h\to 0}\frac{V(x+h)-V(x)}{h} \\ &= \lim_{h\to 0}\frac{2\pi\left|f(x)-g(x)\right|(x-c)\,h+\pi\left|f(x)-g(x)\right|h^2}{h} \\ &= 2\pi\left|f(x)-g(x)\right|(x-c). \end{aligned} \tag{9.12}$$

Also, $V(a)=0$ and this means that to find the volume of revolution it suffices to solve the initial value problem,

$$\frac{dV}{dx}=2\pi\left|f(x)-g(x)\right|(x-c),\ V(a)=0 \tag{9.13}$$

and the volume of revolution will equal $V(b)$. Note that in the above formula, it is not necessary to worry about which is larger, $f(x)$ or $g(x)$ because it is expressed in terms of the absolute value of their difference. However, in doing the computations necessary to solve a given problem, you typically will have to worry about which is larger.

Example 9.15.5 *Find the volume of the solid formed by revolving the region between* $y=\sin(x)$ *and the* x *axis for* $x\in[0,\pi]$ *about the* y *axis.*

In this example, $c=0$ and since $\sin(x)\geq 0$ for $x\in[0,\pi]$, the initial value problem is

$$\frac{dV}{dx}=2\pi x\sin(x),\ V(0)=0.$$

Then using integration by parts,

$$V(x)=2\pi(\sin x-x\cos x)+C$$

and from the initial condition, $C=0$. Therefore, the volume is $V(\pi)=2\pi^2$.

Example 9.15.6 *Find the volume of the solid formed by revolving the region between* $y=\sin x, y=\cos x$ *and the* x *axis for* $x\in[0,\pi/4]$ *about the line* $x=-4$

In this example, $\cos x>\sin x$ for $x\in[0,\pi/4]$ and so the initial value problem is

$$\frac{dV}{dx}=2\pi(x+4)(\cos x-\sin x),\ V(0)=0.$$

Using integration by parts

$$V \in \int 2\pi (x+4)(\cos x - \sin x)\, dx = 2(5\cos x + x\sin x + 3\sin x + x\cos x)\pi + C$$

and the initial condition gives $C = -10\pi$. Therefore, the volume is

$$\begin{aligned} V(\pi/4) &= 2(5\cos(\pi/4) + (\pi/4)\sin(\pi/4) + 3\sin(\pi/4) + (\pi/4)\cos(\pi/4))\pi - 10\pi \\ &= 2\left(4\sqrt{2} + \frac{1}{4}\pi\sqrt{2}\right)\pi - 10\pi \end{aligned}$$

Example 9.15.7 *Find the volume of a sphere of radius R using the method of shells.*

In this case the volume of the sphere is obtained by revolving the region between $y = \sqrt{R^2 - x^2}, y = -\sqrt{R^2 - x^2}$ and the x axis for $x \in [0, R]$ about the y axis. Thus this volume satisfies the initial value problem

$$\frac{dV}{dx} = 2\pi\left(2\sqrt{R^2 - x^2}\right)x,\ V(0) = 0.$$

Thus, using the method of substitution,

$$\int 2\pi\left(2\sqrt{R^2 - x^2}\right)x\, dx = -\frac{4}{3}\left(\sqrt{R^2 - x^2}\right)^3\pi + C$$

and from the initial condition, $C = \frac{4}{3}R^3\pi$. Therefore,

$$V(x) = -\frac{4}{3}\left(\sqrt{R^2 - x^2}\right)^3\pi + \frac{4}{3}R^3\pi$$

and so the volume of the sphere is $V(R) = \frac{4}{3}R^3\pi$, the same formula obtained earlier using the method of cross sections to set up the differential equation.

9.16 Exercises

1. The equation of an ellipse is $\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$. Sketch the graph of this in the case where $a = 2$ and $b = 3$. Now find the volume of the solid obtained by revolving this ellipse about the y axis. What is the volume if it is revolved about the x axis? What is the volume of the solid obtained by revolving it about the line $x = -2a$?

2. A sphere of radius R has a hole drilled through a diameter which is centered at the diameter and of radius $r < R$. Find the volume of what is left after the hole has been drilled. What is the volume of the material which was taken out?

3. Show the volume of a right circular cone is $(1/3) \times$ area of the base $\times$ height.

4. Let R be a region in the xy plane of area A and consider the cone formed by fixing a point in space h units above R and taking the union of all lines starting at this point which end in R. Show that under these general conditions, the volume of the cone is $(1/3) \times A \times h$. **Hint:** The cross sections at height y look just like R but shrunk. Argue that the area at height y, denoted by $A(y)$ is simply $A(y) = A\frac{(h-y)^2}{h^2}$.

5. A circle of radius r in the xy plane is the base of a solid which has the property that cross sections perpendicular to the x axis are equilateral triangles. Find the volume of this solid.

6. The region between $y = x^2$ and $y = x^3$ for $x \in [0, 1]$ is revolved about the y axis. Find the volume of the resulting solid using the method of cross sections. Now find it using the method of shells.

7. A square having each side equal to r in the xy plane is the base of a solid which has the property that cross sections perpendicular to the x axis are equilateral triangles. Find the volume of this solid.

8. The bounded region between $y = x^2$ and $y = x$ is revolved about the x axis. Find the volume of the solid which results.

9. The region between $y = \ln x$, and the x axis for $x \in [1, 3]$ is revolved about the y axis. What is the volume of the resulting solid?

10. The region between $y = x^2$ for $x \in [0, 1]$ is revolved about the line, $x = -4$. Find the volume of the solid which results.

11. The region between $y = \arctan(x)$, and the x axis for $x \in [0, 2]$ is revolved about the y axis. Find the volume of the resulting solid.

12. The region between $y = \arctan(x)$, and the x axis for $x \in [0, 2]$ is revolved about the line $x = -1$. Find the volume of the resulting solid.

13. The region between $y = \sin(x)$, and the x axis for $x \in [0, \pi]$ is revolved about the y axis. Find the volume of the resulting solid.

14. The region between $y = \sin(x)$, and the x axis for $x \in [0, \pi]$ is revolved about the line $x = -1$. Find the volume of the resulting solid.

15. The region between $y = \sin(x)$, and the x axis for $x \in [0, \pi]$ is revolved about the line $x = 5$. Find the volume of the resulting solid.

16. The region between $y = \sin(x)$, and the x axis for $x \in [0, \pi]$ is revolved about the line $y = 2$. Find the volume of the resulting solid.

17. The region between $y = 1 + \sin x$ and the x axis for $x \in [0, 2\pi]$ is revolved about the y axis. Find the volume of the solid which results.

18. The region between $y = 2 + \sin 3x$ and the x axis for $x \in [0, \pi/3]$ is revolved about the line $x = -1$. Find the volume of the solid which results.

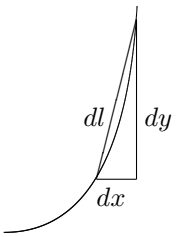
19. The region between $y = x^3 - x$ and the x axis is revolved about the line $x = -1$. Find the volume of the solid which results.

20. The region between $y = \sin 2x$ and the x axis for $x \in [0, \pi]$ is revolved about the line $x = -1$. Find the volume of the solid which results.

9.17 Lengths And Areas Of Surfaces Of Revolution

9.17.1 Lengths

The same techniques can be used to compute lengths of the graph of a function, $y = f(x)$. Consider the following picture.



which depicts a small right triangle attached as shown to the graph of a function, $y = f(x)$ for $x \in [a,b]$. If the triangle is small enough, this shows the length of the curve joined by the hypotenuse of the right triangle is essentially equal to the length of the hypotenuse. Thus, $(dl)^2 = (dx)^2 + (dy)^2$ and dividing by $(dx)^2$ yields

$$\frac{dl}{dx} = \sqrt{1 + \left(\frac{dy}{dx}\right)^2} = \sqrt{1 + f'(x)^2}, l(a) = 0$$

as an initial value problem for the function, $l(x)$ which gives the length of this curve on $[a,x]$.

This definition gives the right answer for the length of a straight line. To see this, consider a straight line through the points (a,b) and (c,d) where $a < c$. Then the right answer is given by the Pythagorean theorem or distance formula and is $\sqrt{(a-c)^2 + (b-c)^2}$. What is obtained from the above initial value problem? The equation of the line is $f(x) = b + \left(\frac{d-b}{c-a}\right)(x-a)$ and so $f'(x) = \left(\frac{d-b}{c-a}\right)$. Therefore, letting l denote the arc length function,

$$\frac{dl}{dx} = \sqrt{1 + \left(\frac{d-b}{c-a}\right)^2}, l(a) = 0.$$

Thus $l(x) = \sqrt{1 + \left(\frac{d-b}{c-a}\right)^2}(x-a)$ and in particular, $l(c)$, the length of the line is given by

$$\sqrt{1 + \left(\frac{d-b}{c-a}\right)^2}(c-a) = \sqrt{(a-c)^2 + (b-c)^2}$$

as hoped. Thus this differential equation gives the right answer in the familiar cases but it also can be used to find lengths for more general curves than straight lines. Here is another familiar example.

Example 9.17.1 *Find the length of the part of the circle having radius r which is between the points $\left(0, \frac{\sqrt{2}}{2}r\right)$ and $\left(\frac{\sqrt{2}}{2}r, \frac{\sqrt{2}}{2}r\right)$.*

Here the function is $f(x) = \sqrt{r^2 - x^2}$ and so $f'(x) = -x/\sqrt{r^2 - x^2}$. Therefore, our differential equation is

$$\frac{dl}{dx} = \sqrt{1 + \frac{x^2}{(r^2 - x^2)}} = \sqrt{\frac{r^2}{r^2 - x^2}} = \frac{r}{\sqrt{r^2 - x^2}}.$$

Therefore, l is an antiderivative of this last function. Using a trig substitution, $x = r\sin\theta$, it follows $dx = r\cos\theta d\theta$ and so

$$\begin{aligned}\int \frac{r}{\sqrt{r^2-x^2}}\,dx &= \int \frac{1}{\sqrt{1-\sin^2\theta}} r\cos\theta\ d\theta \\ &= r\int d\theta = r\theta + C\end{aligned}$$

Hence changing back to the variable x it follows $l(x) = r\arcsin\left(\frac{x}{r}\right) + C$. It only remains to find the constant. Plugging in $x = 0$ this gives $0 = C$ and $l(x) = r\arcsin\left(\frac{x}{r}\right)$ so in particular, the length of the desired arc is

$$l\left(\frac{\sqrt{2}}{2}r\right) = r\arcsin\left(\frac{\sqrt{2}}{2}\right) = r\frac{\pi}{4}.$$

Note this gives the length of one eighth of the circle and so from this the length of the whole circle should be $2r\pi$. Here is another example

Example 9.17.2 *Find the length of the graph of $y = x^2$ between $x = 0$ and $x = 1$.*

Here $f'(x) = 2x$ and so the initial value problem to be solved is

$$\frac{dl}{dx} = \sqrt{1+4x^2},\ l(0) = 0.$$

Thus, getting the exact answer depends on finding

$$\int \sqrt{1+4x^2}\,dx.$$

Use the trig. substitution, $2x = \tan u$ so $dx = \frac{1}{2}\left(\sec^2 u\right)du$. Therefore, making this substitution and using (9.7) on Page 213,

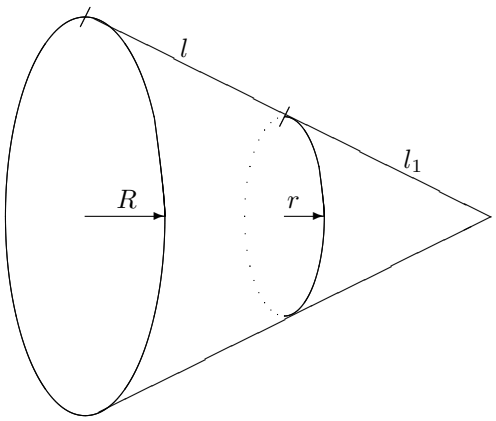
$$\begin{aligned}\int \sqrt{1+4x^2}\,dx &= \frac{1}{2}\int \left(\sec^3 u\right)\,du \\ &= \frac{1}{4}(\tan u)(\sec u) + \frac{1}{4}\ln|\sec u + \tan u| + C \\ &= \frac{1}{4}(2x)\left(\sqrt{1+4x^2}\right) + \frac{1}{4}\ln\left|2x + \sqrt{1+4x^2}\right| + C\end{aligned}$$

and since $l(0) = 0$ it must be the case that $C = 0$ and so the desired length is

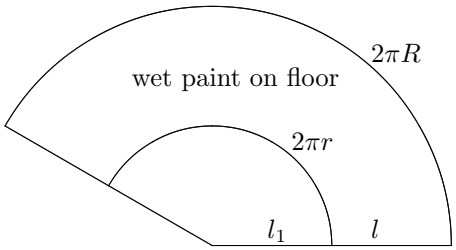
$$l(1) = \frac{1}{2}\sqrt{5} + \frac{1}{4}\ln\left|2+\sqrt{5}\right|$$

9.17.2 Surfaces Of Revolution

The problem of finding the surface area of a solid of revolution is closely related to that of finding the length of a graph. First consider the following picture of the frustum of a cone in which it is desired to find the lateral surface area. In this picture, the frustum of the cone is the left part which has an l next to it and the lateral surface area is this part of the area of the cone.



To do this, imagine painting the sides and rolling the shape on the floor for exactly one revolution. The wet paint would make the following shape.



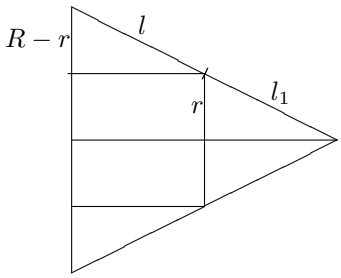
What would be the area of this wet paint? Its area would be the difference between the areas of the two sectors shown, one having radius l_1 and the other having radius $l+l_1$. Both of these have the same central angle equal to

$$\frac{2\pi R}{2\pi\left(l+l_1\right)}2\pi = \frac{2\pi R}{l+l_1}.$$

Therefore, by Theorem 3.8.2 on Page 69, this area is

$$\left(l+l_1\right)^2\frac{\pi R}{\left(l+l_1\right)} - l_1^2\frac{\pi R}{\left(l+l_1\right)} = \pi R l\frac{l+2l_1}{l+l_1}$$

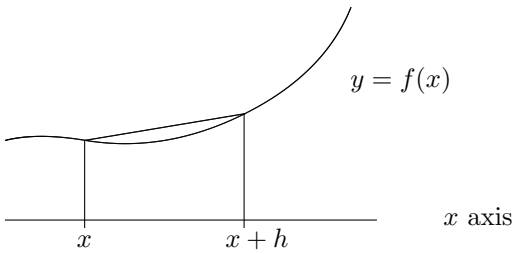
The view from the side is



and so by similar triangles, $l_1 = lr/(R-r)$. Therefore, substituting this into the above, the area of this frustum is

$$\pi Rl\frac{l+2\left(\frac{lr}{R-r}\right)}{l+\left(\frac{lr}{R-r}\right)} = \pi l(R+r) = 2\pi l\left(\frac{R+r}{2}\right).$$

Now consider a function, f, defined on an interval, $[a,b]$ and suppose it is desired to find the area of the surface which results when the graph of this function is revolved about the x axis. Consider the following picture of a piece of this graph.



Let $A(x)$ denote the area which results from revolving the graph of the function restricted to $[a,x]$. Then from the above formula for the area of a frustum,

$$\frac{A(x+h)-A(x)}{h} \approx 2\pi\frac{1}{h}\sqrt{h^2+(f(x+h)-f(x))^2}\left(\frac{f(x+h)+f(x)}{2}\right)$$

where $\approx$ denotes that these are close to being equal and the approximation gets increasingly good as $h \to 0$. Therefore, rewriting this a little yields

$$\frac{A(x+h)-A(x)}{h} \approx 2\pi\sqrt{1+\left(\frac{f(x+h)-f(x)}{h}\right)^2}\left(\frac{f(x+h)+f(x)}{2}\right)$$

Therefore, taking the limit as $h \to 0$, and using $A(a)=0$, this yields the following initial value problem for A which can be used to find the area of a surface of revolution.

$$A'(x) = 2\pi f(x)\sqrt{1+f'(x)^2},\ A(a)=0.$$

Example 9.17.3 *Find the surface area of the surface obtained by revolving the function, $y=r$ for $x\in[a,b]$ about the x axis. Of course this is just the cylinder of radius r and height $b-a$ so this area should equal $2\pi r(b-a)$. (Imagine painting it and rolling it on the floor and then taking the area of the rectangle which results.)*

Using the above initial value problem, solve

$$A'(x) = 2\pi r\sqrt{1+0^2},\ A(a) = 0.$$

The solution is $A(x) = 2\pi r(x-a)$. Therefore, $A(b) = 2\pi r(b-a)$ as expected.

Example 9.17.4 *Find the surface area of a sphere of radius r.*

Here the function involved is $f(x) = \sqrt{r^2-x^2}$ for $x \in [-r, r]$ and it is to be revolved about the x axis. In this case

$$f'(x) = \frac{-x}{\sqrt{r^2-x^2}}$$

and so the initial value problem is of the form

$$A'(x) = 2\pi \overbrace{\sqrt{r^2-x^2}\sqrt{1+\frac{x^2}{r^2-x^2}}}^{=r},\ A(-r) = 0$$

Thus, simplifying the above yields $A'(x) = 2\pi r$ and so $A(x) = 2\pi rx + C$ and since $A(-r) = 0$, it follows that $C = 2\pi r^2$. Therefore, the surface area equals $A(r) = 2\pi r^2 + 2\pi r^2 = 4\pi r^2$.

9.18 Exercises

1. Find the length of the graph of $y = \ln(\cos x)$ for $x \in [0, \pi/4]$.
2. The curve defined by $y = \ln(\cos x)$ for $x \in [0, 2]$ is revolved about the x axis. Find the area of the resulting surface of revolution.
3. Find the length of the graph of $y = x^{1/2} - \frac{x^{3/2}}{3}$ for $x \in [0, 3]$.
4. The graph of the function, $y = x^3$ is revolved about the x axis for $x \in [0, 1]$. Find the area of the resulting surface of revolution.
5. The graph of the function, $y = x^3$ is revolved about the y axis for $x \in [0, 1]$. Find the area of the resulting surface of revolution. **Hint:** Consider x as a function of y.
6. The graph of the function, $y = \ln x$ is revolved about the y axis for $x \in [1, 2]$. Find the area of the resulting surface of revolution. **Hint:** Consider x as a function of y.
7. The graph of the function, $y = \ln x$ is revolved about the x axis for $x \in [1, 2]$. Find the area of the resulting surface of revolution. If you can't do the integral, set it up.
8. Find the length of $y = \cosh(x)$ for $x \in [0, 1]$.
9. Find the length of $y = \ln|\sec x + \tan x|$ for $x \in \left[0, \frac{\pi}{4}\right]$.
10. Find the length of $y = 2x^2 - \frac{1}{16}\ln x$ for $x \in [1, 2]$.
11. The curve defined by $y = 2x^2 - \frac{1}{16}\ln x$ for $x \in [1, 2]$ is revolved about the x axis. Find the area of the resulting surface of revolution.
12. Find the length of $y = x^2 - \frac{1}{8}\ln x$ for $x \in [1, 2]$.
13. The curve defined by $y = x^2 - \frac{1}{8}\ln x$ for $x \in [1, 2]$ is revolved about the x axis. Find the area of the resulting surface of revolution.

14. The curve defined by $y = \cosh(x)$ for $x \in [0,1]$ is revolved about the x axis. Find the area of the resulting surface of revolution.

15. The curve defined by $y = \cosh(x)$ for $x \in [0,1]$ is revolved about the line $y = -3$. Find the area of the resulting surface of revolution.

16. For a a positive real number, find the length of $y = \frac{ax^2}{2} - \frac{1}{4a}\ln x$ for $x \in [1,2]$. Of course your answer should depend on a.

17. The graph of the function, $y = x^2$ for $x \in [0,1]$ is revolved about the x axis. Find the area of the surface of revolution.

18. The graph of the function, $y = \sqrt{x}$ for $x \in [0,1]$ is revolved about the y axis. Find the area of the surface of revolution. **Hint:** Switch x and y and then use the previous problem.

19. The graph of the function, $y = x^{1/2} - \frac{x^{3/2}}{3}$ is revolved about the x axis. Find the area of the surface of revolution if $x \in [0,2]$.

20. The graph of the function, $y = \sinh x$ for $x \in [0,1]$ is revolved about the x axis. Find the area of the surface of revolution.

21. The ellipse, $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ is revolved about the x axis. Find the area of the surface of revolution.

22. Find the length of the graph of $y = \frac{2}{3}(x-1)^{3/2}$ for $x \in [0,1]$.

23. The curve defined by $y = \frac{2}{3}(x-1)^{3/2}$ for $x \in [0,2]$ is revolved about the x axis. Find the area of the resulting surface of revolution.

24. Suppose $f'(x) = \sqrt{\sec^2 x - 1}$ and $f(0) = 0$. Find the length of the graph of $y = f(x)$ for $x \in [0,1]$.

25. The curve defined by $y = f(x)$ for $x \in [0,\pi]$ is revolved about the x axis where $f'(x) = \sqrt{(2+\sin x)^2 - 1}$, $f(0) = 1$. Find the area of the resulting surface of revolution.

9.19 Other Differential Equations

9.19.1 The Equation $y' + a(t)\,y = b(t)$

The homogeneous first order constant coefficient linear differential equation is a differential equation of the form

$$y' + ay = 0. \tag{9.14}$$

It is arguably the most important differential equation in existence. Generalizations of it include the entire subject of linear differential equations and even many of the most important partial differential equations occurring in applications.

Here is how to find the solutions to this equation. Multiply both sides of the equation by e^{at}. Then use the product and chain rules to verify that

$$e^{at}(y' + ay) = \frac{d}{dt}\left(e^{at}y\right) = 0.$$

Therefore, since the derivative of the function $t \to e^{at}y(t)$ equals zero, it follows this function must equal some constant, C. Consequently, $ye^{at} = C$ and so $y(t) = Ce^{-at}$. This shows

that if there is a solution of the equation, $y' + ay = 0$, then it must be of the form Ce^{-at} for some constant, C. You should verify that every function of the form, $y(t) = Ce^{-at}$ is a solution of the above differential equation, showing this yields all solutions. This proves the following theorem.

Theorem 9.19.1 *The solutions to the equation, $y' + ay = 0$ consist of all functions of the form, Ce^{-at} where C is some constant.*

Exercise 9.19.2 *Radioactive substances decay in the following way. The rate of decay is proportional to the amount present. In other words, letting $A(t)$ denote the amount of the radioactive substance at time t, $A(t)$ satisfies the following initial value problem.*

$$A'(t) = -k^2 A(t), \ A(0) = A_0$$

where A_0 is the initial amount of the substance. What is the solution to the initial value problem?

Write the differential equation as $A'(t) + k^2 A(t) = 0$. From Theorem 9.19.1 the solution is

$$A(t) = Ce^{-k^2 t}$$

and it only remains to find C. Letting $t = 0$, it follows $A_0 = A(0) = C$. Thus $A(t) = A_0 \exp\left(-k^2 t\right)$.

Now consider a slightly harder equation.

$$y' + a(t)\, y = b(t).$$

In the easier case, you multiplied both sides by e^{at}. In this case, you multiply both sides by $e^{A(t)}$ where $A'(t) = a(t)$. In other words, you find an antiderivative of $a(t)$ and multiply both sides of the equation by e raised to that function. (It will be shown later in Theorem 10.3.4 on Page 267 that such an A always exists provided a is continuous.)Thus

$$e^{A(t)}\left(y' + a(t)\, y\right) = e^{A(t)} b(t).$$

Now you notice that this becomes

$$\frac{d}{dt}\left(e^{A(t)} y\right) = e^{A(t)} b(t). \tag{9.15}$$

This follows from the chain rule.

$$\frac{d}{dt}\left(e^{A(t)} y\right) = A'(t)\, e^{A(t)} y + e^{A(t)} y' = e^{A(t)}\left(y' + a(t)\, y\right).$$

Then from (9.15),

$$e^{A(t)} y \in \int e^{A(t)} b(t)\, dt.$$

Therefore, to find the solution, you find a function in $\int e^{A(t)} b(t)\, dt$, say $F(t)$, and

$$e^{A(t)} y = F(t) + C$$

for some constant, C, so the solution is given by $y = e^{-A(t)} F(t) + e^{-A(t)} C$. This proves the following theorem.

Theorem 9.19.3 *The solutions to the equation, $y' + a(t)\, y = b(t)$ consist of all functions of the form*

$$y = e^{-A(t)} F(t) + e^{-A(t)} C$$

where $F(t) \in \int e^{A(t)} b(t)\, dt$ and C is a constant.

Example 9.19.4 *Find the solution to the initial value problem $y' + 2ty = \sin(t)\, e^{-t^2}$, $y(0) = 3$.*

Multiply both sides by e^{t^2} because $t^2 \in \int t dt$. Then $\frac{d}{dt}\left(e^{t^2} y\right) = \sin(t)$ and so $e^{t^2} y = -\cos(t) + C$. Hence the solution is of the form $y(t) = -\cos(t)\, e^{-t^2} + Ce^{-t^2}$. It only remains to choose C in such a way that the initial condition is satisfied. From the initial condition, $3 = y(0) = -1 + C$ and so $C = 4$. Therefore, the solution is $y = -\cos(t)\, e^{-t^2} + 4e^{-t^2}$. Now at this point, you should check and see if it works. It needs to solve both the initial condition and the differential equation.

Finally, here is a uniqueness theorem.

Theorem 9.19.5 *If $a(t)$ is a continuous function, there is at most one solution to the initial value problem, $y' + a(t)\, y = b(t)$, $y(r) = y_0$.*

Proof: If there were two solutions, y_1, and y_2, then letting $w = y_1 - y_2$, it follows $w' + a(t)\, w = 0$ and $w(r) = 0$. Then multiplying both sides of the differential equation by $e^{A(t)}$ where $A'(t) = a(t)$, (It will be shown later in Theorem 10.3.4 on Page 267 that such an A always exists provided a is continuous.) it follows

$$\left(e^{A(t)} w\right)' = 0$$

and so $e^{A(t)} w(t) = C$ for some constant, C. However, $w(r) = 0$ and so this constant can only be 0. Hence $w = 0$ and so $y_1 = y_2$.

9.19.2 Separable Differential Equations

Definition 9.19.6 *Separable differential equations are those which can be written in the form*

$$\frac{dy}{dx} = \frac{f(x)}{g(y)}.$$

The reason these are called separable is that if you formally cross multiply,

$$g(y)\, dy = f(x)\, dx$$

and the variables are "separated". The x variables are on one side and the y variables are on the other.

Proposition 9.19.7 *If $G'(y) = g(y)$ and $F'(x) = f(x)$, then if the equation, $F(x) - G(y) = c$ specifies y as a differentiable function of x, then $x \to y(x)$ solves the separable differential equation*

$$\frac{dy}{dx} = \frac{f(x)}{g(y)}. \tag{9.16}$$

Proof: Differentiate both sides of $F(x) - G(y) = c$ with respect to x. Using the chain rule,

$$F'(x) - G'(y)\, \frac{dy}{dx} = 0.$$

Therefore, since $F'(x) = f(x)$ and $G'(y) = g(y)$, $f(x) = g(y)\, \frac{dy}{dx}$ which is equivalent to (9.16).

Example 9.19.8 *Find the solution to the initial value problem,*

$$y' = \frac{x}{y^2},\ y(0) = 1.$$

This is a separable equation and in fact, $y^2 dy = x dx$ so the solution to the differential equation is of the form

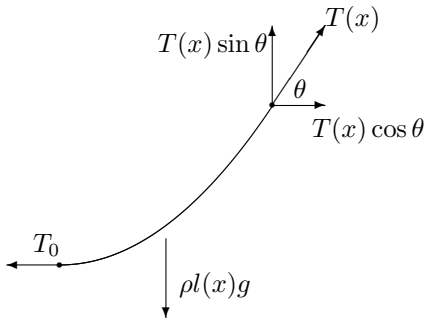
$$\frac{y^3}{3} - \frac{x^2}{2} = C \tag{9.17}$$

and it only remains to find the constant, C. To do this, you use the initial condition. Letting $x = 0$, it follows $\frac{1}{3} = C$ and so

$$\frac{y^3}{3} - \frac{x^2}{2} = \frac{1}{3}$$

Example 9.19.9 *What is the equation of a hanging chain?*

Consider the following picture of a portion of this chain.



In this picture, ρ denotes the density of the chain which is assumed to be constant and g is the acceleration due to gravity. $T(x)$ and T_0 represent the magnitude of the tension in the chain at x and at 0 respectively, as shown. Let the bottom of the chain be at the origin as shown. If this chain does not move, then all these forces acting on it must balance. In particular,

$$T(x)\sin\theta = l(x)\rho g,\ T(x)\cos\theta = T_0.$$

Therefore, dividing these yields

$$\frac{\sin\theta}{\cos\theta} = l(x)\overbrace{\rho g/T_0}^{\equiv c}.$$

Now letting $y(x)$ denote the y coordinate of the hanging chain corresponding to x,

$$\frac{\sin\theta}{\cos\theta} = \tan\theta = y'(x).$$

Therefore, this yields

$$y'(x) = c\,l(x).$$

Now differentiating both sides of the differential equation,

$$y''(x) = cl'(x) = c\sqrt{1 + y'(x)^2}$$

and so

$$\frac{y''(x)}{\sqrt{1 + y'(x)^2}} = c.$$

Let $z(x) = y'(x)$ so the above differential equation becomes

$$\frac{z'(x)}{\sqrt{1+z^2}} = c.$$

Therefore, $\int \frac{z'(x)}{\sqrt{1+z^2}}\,dx = cx + d$. Change the variable in the antiderivative letting $u = z(x)$ and this yields

$$\begin{aligned}\int \frac{z'(x)}{\sqrt{1+z^2}}\,dx &= \int \frac{du}{\sqrt{1+u^2}} = \sinh^{-1}(u) + C \\ &= \sinh^{-1}(z(x)) + C.\end{aligned}$$

by (8.10) on Page 180. Therefore, combining the constants of integration,

$$\sinh^{-1}(y'(x)) = cx + d$$

and so

$$y'(x) = \sinh(cx + d).$$

Therefore,

$$y(x) = \frac{1}{c}\cosh(cx + d) + k$$

where d and k are some constants and $c = \rho g/T_0$. Curves of this sort are called catenaries. Note these curves result from an assumption the only forces acting on the chain are as shown.

9.20 Exercises

1. For x sufficiently large, let $f(x) = (\ln x)^{\ln x}$ Find $f'(x)$. How big does x need to be in order for this to make sense. **Hint:** You should have $\ln x > 0$.

2. In the hanging chain problem the picture and the derivation involved an assumption that at its lowest point, the chain was horizontal. Imagine lifting the end higher and higher and you will see this might not be the case in general. Can you modify the above derivation for the hanging chain to show that even in this case the chain will be in the form of a catenary?

3. Find the solution to the initial value problem,

$$y' = 1 + y^2,\ y(0) = 0.$$

4. Verify that for any constant, C, the function, $y(t) = Ce^{-at}$ solves the differential equation, $y' + ay = 0$.

5. In Example 9.19.2 the half life is the time it takes for half of the original amount of the substance to decay. Find a formula for the half life assuming you know k^2.

6. There are ten grams of a radioactive substance which is allowed to decay for five years. At the end of the five years there are 9.5 grams of the substance left. Find the half life of the substance. **Hint:** Use the given information to find k^2 and then use Problem 5.

7. The giant arch in St. Louis is in the form of an inverted catenary. Why?

8. Sometimes banks compound interest continuously. One way to think of this is to let the amount in the account satisfy the initial value problem,

$$A'(t) = rA(t),\ A(0) = A_0$$

where here $A(t)$ is the amount at time t measured in years, r is the interest rate per year, and A_0 is the initial amount. Find $A(t)$ explicitly. If \$100 is placed in an account which is compounded continuously at 6% per year, how many years will it be before there is \$200 in the account? **Hint:** In this case, $r = .06$.

9. An object falling through the air experiences a force of gravity and air resistance. Later, it will be shown that this implies the velocity satisfies a differential equation of the form

$$v' = \frac{g}{m} - kv$$

where k is a positive constant and $\frac{g}{m}$ is also a positive constant. Find the solutions to this differential equation and determine what happens to v as t gets large. If you do it right you will find the terminal velocity.

10. Solve the initial value problem

$$A'(t) = rA(t) + 1 + \sin(t), A(0) = 1$$

and describe its behaviour as $t \to \infty$. Assume r is a positive constant.

11. A population is growing at the rate of 11% per year. This means it satisfies the differential equation, $A' = .11A$. Find the time it takes for the population to double.

12. A substance is decaying at the rate of .0 1 per year. Find the half life of the substance.

13. The half live of a substance is 40 years. Find the rate of decay of the substance.

14. A sample of 4 grams of a radioactive substance has decayed to 3 grams in 5 days. Find the half life of the substance. Give your answer in terms of logarithms.

15. \$1000 is deposited in an account that earns interest at the rate of 5% per year compounded continuously. How much will be in the account after 10 years?

16. A sample of one ounce of water from a water supply is cultured in a Petri dish. After four hours, there are 3000.0 bacteria present and after six hours there are 7000 bacteria present. How many were present in the original sample?

17. Carbon 14 is a radioactive isotope of carbon and it is produced at a more or less constant rate in the earth's atmosphere by radiation from the sun. It also decays at a rate proportional to the amount present. Show this implies that, assuming this has been going on for billions of years, it is reasonable to assume the amount of Carbon 14 in the atmosphere is essentially constant over time. **Hint:** By assumption, if A is the amount of Carbon 14 in the atmosphere, $\frac{dA}{dt} = -k^2A + r$ where k^2 is the constant of decay described above and r is the constant rate of production. Now show $A'(t) = Ce^{-k^2t}$. Conclude $A(t) = D - \left(C/k^2\right)e^{-k^2t}$. What happens to this over a long period of time?

18. The method of carbon dating is based on the result of Problem 17. When an animal or plant is alive it absorbs carbon from the atmosphere and so when it is living it has a known percentage of carbon 14. When it dies, it quits absorbing carbon and the carbon 14 begins to decay. After some time, t, the amount of carbon 14 can be

measured and on this basis, the time since the death of the animal or plant can be estimated. Given the half life of carbon 14 is 5730 years and the amount of carbon 14 in a mummy is .3 what it was at the time of death, how long has it been since the mummy was alive? (To see how to come up with this figure for the half life, see Problem 6. You do experiments and take measurements over a smaller period of time.)

19. The half life of carbon 14 is known to be 5730 years. A certain tree stump is known to be 4000 years old. What percentage of the original carbon 14 should it contain?

20. One model for population growth is to assume the rate of growth is proportional to both the population and the difference between some maximum sustainable population and the population. The reason for this is that when the population gets large enough, there begin to be insufficient resources. Thus $\frac{dA}{dt} = kA(M - A)$, where k and M are positive constants. Show this is a separable differential equation and its solutions are of the form

$$A(t) = \frac{M}{1 + CMe^{-kMt}}$$

where C is a constant. Given three measurements of population at three equally spaced times, show how to predict the maximum sustainable population[1].

21. Homogeneous differential equations are those which can be written in the form $y' = f\left(\frac{y}{x}\right)$. For example, $y' = \frac{x^2}{y^2+x^2}$. There is a trick to solving such equations. You define a new variable, $v = y/x$ and then write the differential equation in terms of v rather than y. Show that $xv' + v = y'$ and so the differential equation reduces to $xv' = f(v) - v$, a separable differential equation. Use the technique to solve $y' = 1 + 2\frac{y}{x}$, $y(1) = 1$.

22. Solve the following initial value problems involving homogeneous differential equations.

    (a) $y' = 1 + \left(\frac{y}{x}\right) + \left(\frac{y}{x}\right)^2$, $y(1) = 1$.

    (b) $y' = \tan\left(\frac{y}{x}\right) + \frac{y}{x}$, $y(1) = 1$.

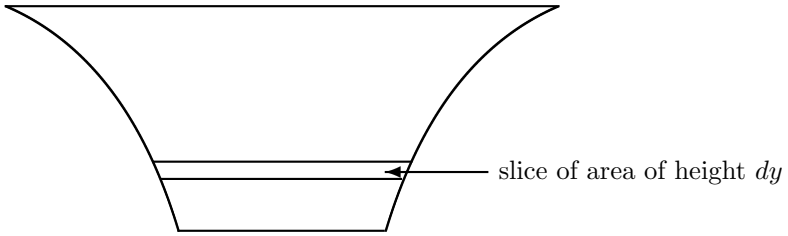
    (c) $y' = \frac{x^3+xy^2+y^3}{x^2y+xy^2}$, $y(1) = 1$.

9.21 Force On A Dam And Work

9.21.1 Force On A Dam

Imagine you are a fish swimming in a lake behind a dam and you are interested in the total force acting on the dam. The following picture is what you would see.

[1]This has been done with the earth's population and the maximum sustainable population has been exceeded. Therefore, the model is far too simplistic for human population growth. However, it would work somewhat better for predicting the growth of things like bacteria.



The reason you would be interested in that long thin slice of area having essentially the same depth, say at y feet is because the pressure in the water at that depth is constant and equals $62.5y$ pounds per square foot[2]. Therefore, the total force the water exerts on the long thin slice is

$$dF = 62.5yL(y)\,dy$$

where $L(y)$ denotes the length of the slice. Therefore, the total force on the dam up to depth y is obtained as a solution to the initial value problem,

$$\frac{dF}{dy} = 62.5yL(y),\ F(0) = 0.$$

Example 9.21.1 *Suppose the width of a dam at depth y feet equals $L(y) = 1000 - y$ and its depth is 500 feet. Find the total force in pounds exerted on the dam.*

From the above, this is obtained as the solution to the initial value problem

$$\frac{dF}{dy} = 62.5y(1000 - y),\ F(0) = 0$$

which is $F(y) = -20.83y^3 + 31250y^2$. The total force on the dam would be

$$F(500) = -20.83(500)^3 + 31250(500)^2 = 5,208,750,000.0$$

pounds. In tons this is $2,604,375$. That is a lot of force.

9.21.2 Work

Now suppose you are pumping water from a tank of depth d to a height of H feet above the top of the water in the tank. Suppose also that at depth y below the surface, the area of a cross section having constant depth is $A(y)$. The total weight of a slice of water having thickness dy at this depth is $62.5A(y)\,dy$ and the pump needs to lift this weight a distance of $y + H$ feet. Therefore, the work done is $dW = (y + H)\,62.5A(y)\,dy$. An initial value problem for the work done to pump the water down to a depth of y feet would be

$$\frac{dW}{dy} = (y + H)\,62.5A(y),\ W(0) = 0.$$

The reason for the initial condition is that the pump has done no work to pump no water. If the weight of the fluid per cubic foot were different than 62.5 you would do the same things but replace the number.

[2]Later on a nice result on hydrostatic pressure will be presented which will verify this assertion. Here 62.5 is the weight in pounds of a cubic foot of water. If you like, think of a column of water of height y having base area equal to 1 square foot. Then the total force acting on this base area would be $62.5 \times y$ pounds.

Example 9.21.2 *A spherical storage tank sitting on the ground having radius 30 feet is half filled with a fluid which weighs 50 pounds per cubic foot. How much work is done to pump this fluid to a height of 100 feet?*

Letting r denote the radius of a cross section y feet below the level of the fluid, $r^2+y^2 = 900$. Therefore,

$$r = \sqrt{900-y^2}.$$

It follows the area of the cross section at depth y is $\pi\left(900-y^2\right)$. Here $H = 70$ and so the initial value problem to solve is

$$\frac{dW}{dy} = (y+70)\,50\pi\left(900-y^2\right),\ W(0) = 0.$$

Therefore, $W(y) = 50\pi\left(-\frac{1}{4}y^4 - \frac{70}{3}y^3 + 450y^2 + 63\,000y\right)$ and the total work in foot pounds equals

$$W(30) = 50\pi\left(-\frac{1}{4}(30)^4 - \frac{70}{3}(30)^3 + 450(30)^2 + 63\,000(30)\right) = 73\,,125,\ 000\pi$$

In general, the work done by a constant force in a straight line equals the product of the force times the distance over which it acts. This is an over simplification and it will be made more correct later. If the force is varying with respect to position, then you have to use calculus to compute the work. For now, consider the following examples.

Example 9.21.3 *A 500 pound safe is lifted 10 feet. How much work is done?*

The work is $500 \times 10 = 5000$ foot pounds.

Example 9.21.4 *The force needed to stretch a spring x feet past its equilibrium position is kx. This is known as Hooke's law and is a good approximation as long as the spring is not stretched too far. If $k = 3$, how much work is needed to stretch the spring a distance of 2 feet beyond its equilibrium position? The constant, k is called the spring constant. Different springs would have different spring constants. The units on k are pounds/foot.*

This is a case of a variable force. To stretch the spring from x to $x+dx$ requires $3xdx$ foot pounds of work. Therefore, letting W denote the work up till time $x, dW = 3xdx$ and so the initial value problem is

$$\frac{dW}{dx} = 3x,\ W(0) = 0.$$

Thus $W(2) = \frac{3}{2}\left(2^2\right) = 6$ foot pounds because an antiderivative for $3x$ is $\frac{3}{2}x^2$.

9.22 Exercises

1. The main span of the Portage Lake lift bridge[3] weighs 4,400,000 pounds. How much work is done in raising this main span to a height of 100 feet?

[3]This is the heaviiest lift bridge in the world. It joins the towns of Houghton and Hancock in the upper peninsula of Michigan spanning portage lake. It provides 250 feet of clear channel for ships and can provide as much as 100 feet of vertical clearance. The lifting machinery is at the top of two massive towers 180 feet above the water. Aided by 1,100 ton counter weights on each tower, sixteen foot gears pull on 42 cables to raise the bridge. This usually creates impressive traffic jams on either side of the lake. The motion up and down of this span is quite slow.

2. A cylindrical storage tank having radius 20 feet and length 40 feet is filled with a fluid which weighs 50 pounds per cubic foot. This tank is lying on its side on the ground. Find the total force acting on the ends of the tank by the fluid.

3. Suppose the tank in Problem 2 is filled to a depth of 8 feet. Find the work needed to pump the fluid to a height of 50 feet.

4. A conical hole is filled with water. If the depth of the hole is 20 feet and the radius of the hole is 10 feet, how much work is needed to pump the water to a height of 10 feet above the ground?

5. Suppose the spring constant is 2 pounds per foot. Find the work needed to stretch the spring 3 feet beyond equilibrium.

6. A 20 foot chain lies on the ground. It weighs 5 pounds per foot. How much work is done to lift one end of the chain to a height of 20 feet?

7. A dam 500 feet high has a width at depth y equal to $4000 - 2y$ feet. What is the total force on the dam if it is filled?

8. When the bucket is filled with water it weighs 30 pounds and when empty it weighs 2 pounds and the person on top of a 100 foot building exerts a constant force of 40 pounds. The bucket is full at the bottom but leaks at the rate of .1 cubic feet per second. How much work does the person on the top of the building do in lifting the bucket to the top? Will the bucket be empty when it reaches the top? You can use Newton's law that force equals mass times acceleration.

9. In the situation of the above problem, suppose the person on the top maintains a constant velocity of 1 foot per second. How much work does he do and is the bucket empty when it reaches the top?

10. A silo is 10 feet in diameter and at a height of 30 feet there is a hemispherical top. The silage weighs 10 pounds per cubic foot. How much work was done in filling it to the very top?

11. A cylindrical storage tank having radius 10 feet is filled with water to a depth of 20 feet. If the storage tank stands upright on its circular base, what is the total force the water exerts on the sides of the tank? **Hint:** The pressure in the water at depth y is $62.5y$ pounds per square foot.

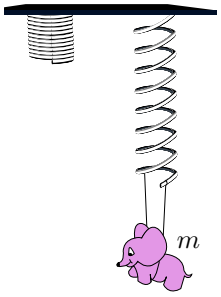
12. A spherical storage tank having radius 10 feet is filled with water. What is the total force the water exerts on the storage tank? **Hint:** The pressure in the water at depth y is $62.5y$ consider the area corresponding to a slice at height y. This is a surface of revolution and you know how to deal with these. The area of this slice times the pressure gives the total force acting on it.

13. A water barrel which is 11 inches in radius and 34 inches high is filled with water. If it is standing on end, what is the total force acting on the circular sides of the barrel?

14. Find the total force acting on the circular sides of the cylinder in Problem 2.

9.23 The Equations Of Undamped And Damped Oscillation

Consider a garage door spring. These springs exert a force which resists extension. Attach such a spring to the ceiling and attach a mass, m, to the bottom end of the spring as shown in the following picture. Any mass will do. It does not have to be a small elephant.



The weight of this mass, mg, is a downward force which extends the spring, moving the bottom end of the spring downward a distance l where the upward force exerted by the spring exactly balances the downward force exerted on the mass by gravity. It has been experimentally observed that as long as the extension, z, of such a spring is not too great, the restoring force exerted by the spring is of the form kz where k is some constant which depends on the spring. (It would be different for a slinky than for a garage door spring.) This is known as Hooke's law which is the simplest model for elastic materials. Therefore, $mg = kl$. Now let y be the displacement from this equilibrium position of the bottom of the spring with the positive direction being up. Thus the acceleration of the spring is y''. The extension of the spring in terms of y is $(l - y)$. Then Newton's second law[4] along with Hooke's law imply

$$my'' = k(l - y) - mg$$

and since $kl - mg = 0$, this yields

$$my'' + ky = 0.$$

Dividing by m and letting $\omega^2 = k/m$ yields the equation for undamped oscillation,

$$y'' + \omega^2 y = 0.$$

Based on physical reasoning just presented, there should be a solution to this equation. It is the displacement of the bottom end of a spring from the equilibrium position. However, it is not enough to base questions of existence in mathematics on physical intuition, although it is sometimes done. The following theorem gives the necessary existence and uniqueness results. The equation is the equation of undamped oscillations. It occurs in modeling a weight on a spring but it also occurs in many other physical settings.

Theorem 9.23.1 *The initial value problem,*

$$y'' + \omega^2 y = 0,\ y(0) = y_0, y'(0) = y_1 \tag{9.18}$$

has a unique solution and this solution is

$$y(t) = y_0 \cos(\omega t) + \frac{y_1}{\omega} \sin(\omega t). \tag{9.19}$$

[4]This important law will be discussed more thoroughly later. I assume you have seen it in a physics class by now.

Proof: You should verify that (9.19) does indeed provide a solution to the initial value problem. It only remains to verify uniqueness. Suppose then that y_1 and y_2 both solve the initial value problem, (9.18). Let $w = y_1 - y_2$. Then you should verify that $w'' + \omega^2 w = 0$, $w(0) = 0 = w'(0)$. Then multiplying both sides of the differential equation by w' it follows

$$w''w' + \omega^2 ww' = 0.$$

However, $w''w' = \frac{1}{2}\frac{d}{dt}(w')^2$ and $w'w = \frac{1}{2}\frac{d}{dt}(w)^2$ so the above equation reduces to

$$\frac{1}{2}\frac{d}{dt}\left((w')^2 + w^2\right) = 0.$$

Therefore, $(w')^2 + w^2$ is equal to some constant. However, when $t = 0$, this shows the constant must be zero. Therefore, $y_1 - y_2 = w = 0$. This proves the uniqueness.

Now consider another sort of differential equation,

$$y'' - a^2 y = 0,\ a > 0 \tag{9.20}$$

To give the complete solution, let $Dy \equiv y'$. Then the differential equation may be written as

$$(D + a)(D - a)\, y = 0.$$

Let $z = (D - a)\, y$. Thus $(D + a)\, z = 0$ and so $z(t) = C_1 e^{-at}$ from Theorem 9.19.1 on Page 247. Therefore,

$$(D - a)\, y \equiv y' - ay = C_1 e^{-at}.$$

Multiply both sides of this last equation by e^{-at}. By the product and chain rules,

$$\frac{d}{dt}\left(e^{-at}y\right) = C_1 e^{-2at}.$$

Therefore,

$$e^{-at}y = \frac{C_1}{-2a}e^{-2at} + C_2$$

and so

$$y = \frac{C_1}{-2a}e^{-at} + C_2 e^{at}.$$

Now since C_1 is arbitrary, it follows any solution of (9.20) is of the form $y = C_1 e^{-at} + C_2 e^{at}$. Now you should verify that any expression of this form actually solves the equation, (9.20). This proves most of the following theorem.

Theorem 9.23.2 *Every solution of the differential equation, $y'' - a^2 y = 0$ is of the form $C_1 e^{-at} + C_2 e^{at}$ for some constants C_1 and C_2 provided $a > 0$. In the case when $a = 0$, every solution of $y'' = 0$ is of the form $C_1 t + C_2$ for some constants, C_1 and C_2.*

All that remains of the proof is to do the part when $a = 0$ which is left as an exercise involving the mean value theorem.

Now consider the differential equation of damped oscillation. In the example of the object bobbing on the end of a spring,

$$my'' = -ky$$

where k was the spring constant and m was the mass of the object. Suppose the object is also attached to a dash pot. This is a device which resists motion like a shock absorber on a car. You know how these work. If the car is just sitting still the shock absorber applies

no force to the car. It only gives a force in response to up and down motion of the car and you assume this force is proportional to the velocity and opposite the velocity. Thus in our spring example, you would have

$$my'' = -ky - \delta^2 y'$$

where δ^2 is the constant of proportionality of the resisting force. Dividing by m and adjusting the coefficients,such damped oscillation satisfies an equation of the form,

$$y'' + 2by' + ay = 0. \tag{9.21}$$

Actually this is a general homogeneous second order equation, more general than what results from damped oscillation. Concerning the solutions to this equation, the following theorem is given. In this theorem the first case is referred to as the underdamped case. The second case is called the critically damped case and the third is called the overdamped case.

Theorem 9.23.3 *Suppose $b^2 - a < 0$. Then all solutions of (9.21) are of the form*

$$e^{-bt}\left(C_1 \cos(\omega t) + C_2 \sin(\omega t)\right). \tag{9.22}$$

where $\omega = \sqrt{a - b^2}$ and C_1 and C_2 are constants. In the case that $b^2 - a = 0$ the solutions of (9.21) are of the form

$$e^{-bt}\left(C_1 + C_2 t\right). \tag{9.23}$$

In the case that $b^2 - a > 0$ the solutions are of the form

$$e^{-bt}\left(C_1 e^{-rt} + C_2 e^{rt}\right), \tag{9.24}$$

where $r = \sqrt{b^2 - a}$.

Proof: Let $z = e^{bt}y$ and write (9.21) in terms of z. Thus, z is a solution to the equation

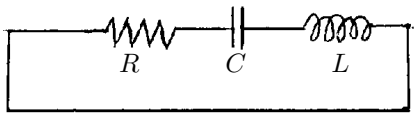
$$z'' + \left(a - b^2\right) z = 0. \tag{9.25}$$

If $b^2 - a < 0$, then by Theorem 9.23.1, $z(t) = C_1 \cos(\omega t) + C_2 \sin(\omega t)$ where $\omega = \sqrt{a - b^2}$. Therefore,

$$y = e^{-bt}\left(C_1 \cos(\omega t) + C_2 \sin(\omega t)\right)$$

as claimed. The other two cases are completely similar. They use Theorem 9.23.2 rather than Theorem 9.23.1.

Example 9.23.4 *An important example of these equations occurs in an electrical circuit having a capacitor, a resistor, and an inductor in series as shown in the following picture.*



The voltage drop across the inductor is $L\frac{di}{dt}$ where i is the current and L is the inductance. The voltage drop across the resistor is Ri where R is the resistance. This is according to Ohm's law. The voltage drop across the capacitor is $v = \frac{Q}{C}$ where Q is the charge on the capacitor and C is a constant called the capacitance. The current equals the rate of change of the charge on the capacitor. Thus $i = Q' = Cv'$. When these voltages are summed, you must get zero because there is no voltage source in the circuit. Thus $L\frac{di}{dt} + Ri + \frac{Q}{C} = 0$ and written in terms of the voltage drop across the capacitor, this becomes $LCv'' + CRv' + v = 0$, a second order linear differential equation of the sort discussed above.

9.24 Exercises

1. Verify that $y = C_1 e^{-at} + C_2 e^{at}$ solves the differential equation, (9.20).

2. Verify (9.25).

3. Verify that all solutions to the differential equation, $y'' = 0$ are of the form $y = C_1 t + C_2$.

4. Show that for all $x \geq 0$,
$$\sin x \leq x - \frac{x^3}{6} + \frac{x^5}{120}. \tag{9.26}$$
Hint: Let $f(x) = x - \frac{x^3}{6} + \frac{x^5}{120} - \sin x$. Then $f(0) = 0$. You need to show $f'(x) > 0$ for $x > 0$. You won't be able to do this directly. Consider $f'(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \cos x$. Then $f'(0) = 0$. You need to show $f''(x) > 0$. You won't be able to do this directly. Consider $f''(x) = -x + \frac{x^3}{6} + \sin x$. Then $f''(0) = 0$. You need to show $f'''(x) > 0$. You won't be able to do this directly. Continue this way. You will eventually have $f^{(k)}(0) = 0$ and it will be obvious that $f^{(k+1)}(x) > 0$.

5. Using Problem 4, estimate $\sin(.1)$ and $\cos(.1)$. Give upper and lower bounds for these numbers.

6. Using Problem 4 and Theorem 5.9.5 on Page 111, establish the limit,
$$\lim_{x \to 0} \frac{\sin(x)}{x} = 1.$$

7. A mass of ten Kilograms is suspended from a spring attached to the ceiling. This mass causes the end of the spring to be displaced a distance of 39.2 $cm.$ The mass end of the spring is then pulled down a distance of one $cm.$ and released. Find the displacement from the equilibrium position of the end of the spring as a function of time. Assume the acceleration of gravity is $9.8 meters/\sec^2$.

8. Keep everything the same in Problem 7 except suppose the suspended end of the spring is also attached to a dash pot which provides a force opposite the direction of the velocity having magnitude $10\sqrt{19}\,|v|$ Newtons for $|v|$ the speed. Give the displacement as before.

9. In (9.21) consider the equation in which $b = -1$ and $a = 3$. Explain why this equation describes a physical system which has some dubious properties.

10. Solve the initial value problem, $y'' + 5y' - y = 0$, $y(0) = 1, y'(0) = 0$.

11. Solve the initial value problem $y'' + 2y' + 2y = 0, y(0) = 0, y'(0) = 1$.

12. In the case of undamped oscillation show the solution can be written in the form $A\cos(\omega t - \phi)$ where ϕ is some angle called a phase shift and a constant, A, called the amplitude.

13. Using Problem 4 and Theorem 5.9.5, establish the limit,
$$\lim_{x \to 0} \frac{1 - \cos x}{x} = 0.$$

14. Is the derivative of a function always continuous? **Hint:** Consider

$$h(x) = \begin{cases} x^2 \sin\left(\frac{1}{x}\right) \text{ if } x \neq 0 \\ 0 \text{ if } x = 0 \end{cases}.$$

Show $h'(0) = 0$. What is $h'(x)$ for $x \neq 0$?

15. Suppose $f(t) = a\cos\omega t + b\sin\omega t$. Show there exists ϕ such that

$$f(t) = \sqrt{a^2+b^2}\sin(\omega t + \phi).$$

Hint: $f(t) = \sqrt{a^2+b^2}\left(\frac{a}{\sqrt{a^2+b^2}}\cos\omega t + \frac{b}{\sqrt{a^2+b^2}}\sin\omega t\right)$ and $\left(\frac{b}{\sqrt{a^2+b^2}}, \frac{a}{\sqrt{a^2+b^2}}\right)$ is a point on the unit circle so it is of the form $(\cos\phi, \sin\phi)$. Now recall the formulas for the sine and cosine of sums of angles. Can you also write $f(t) = \sqrt{a^2+b^2}\cos(\omega t + \phi)$? Explain.

The Integral

10.0.1 Outcomes

1. Describe upper and lower sums and give a correct definition of the integral.

2. Understand theorems about the existence of the integral of compositions of continuous functions with Riemann integrable functions.

3. Understand and be able to use intelligently all fundamental properties of the integral.

4. Understand the proof of the fundamental theorem of calculus and be able to use this theorem.

5. Understand a mathematically rigorous treatment of the logarithm.

6. Understand and be able to use various integration techniques in the context of definite integrals.

7. Describe the difference between an improper integral and an integral. Also tell what it means for an improper integral to converge and understand and use theorems which enable you to make this determination.

8. Understand and work problems involving improper integrals.

The integral originated in attempts to find areas of various shapes and the ideas involved in finding integrals are much older than the ideas related to finding derivatives. In fact, Archimedes[1] was finding areas of various curved shapes about 250 B.C. The integral is needed to remove some of the mathematical loose ends and also to enable the study of more general problems involving differential equations. It will also be useful for formulating other physical models. The technique used for finding the area of a circular segment presented early in the book was essentially that employed by Archimedes and contains the essential ideas for the integral. The main difference is that here the triangles will be replaced with rectangles. You may be wondering what the fuss is about. Areas have already been found as solutions of differential equations. However, there is a profound difference between what is about to be presented and what has just been done. It is related to the fundamental mathematical question of existence. As an illustration, consider the problem of finding the

[1] Archimedes 287-212 B.C. found areas of curved regions by stuffing them with simple shapes which he knew the area of and taking a limit. He also made fundamental contributions to physics. The story is told about how he determined that a gold smith had cheated the king by giving him a crown which was not solid gold as had been claimed. He did this by finding the amount of water displaced by the crown and comparing with the amount of water it should have displaced if it had been solid gold.

area between $y = e^{x^2}$ and the x axis for $x \in [0,1]$. As pointed out earlier, the area is obtained as a solution to the initial value problem,

$$A'(x) = e^{x^2},\ A(0) = 0.$$

So what is the solution to this initial value problem? By Theorem 9.1.1 there is at most one solution, but what is the solution? Does it even exist? More generally, for which functions, f does there exist a solution to the initial value problem, $y'(x) = f(x), y(0) = y_0$? These questions are typical of mathematics. There are usually two aspects to a mathematical concept. One is the question of existence and the other is how to find that which exists. The two questions are often very different and one can have a good understanding of one without having any idea how to go about considering the other. However, both are absolutely essential. In the preceding chapter the only thing considered was the second question.

10.1 Upper And Lower Sums

The Riemann integral pertains to bounded functions which are defined on a bounded interval. Let $[a,b]$ be a closed interval. A set of points in $[a,b]$, $\{x_0, \cdots, x_n\}$ is a partition if

$$a = x_0 < x_1 < \cdots < x_n = b.$$

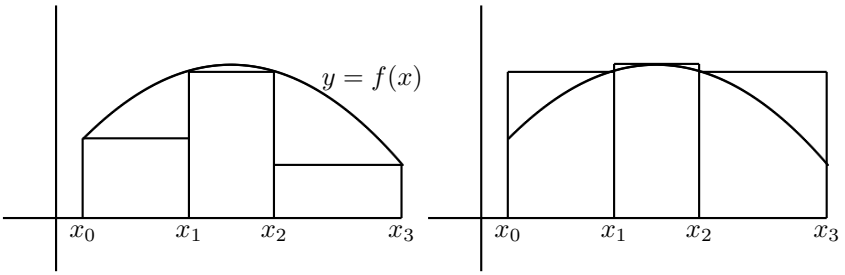
Such partitions are denoted by P or Q. For f a bounded function defined on $[a,b]$, let

$$M_i(f) \equiv \sup\{f(x) : x \in [x_{i-1}, x_i]\},$$
$$m_i(f) \equiv \inf\{f(x) : x \in [x_{i-1}, x_i]\}.$$

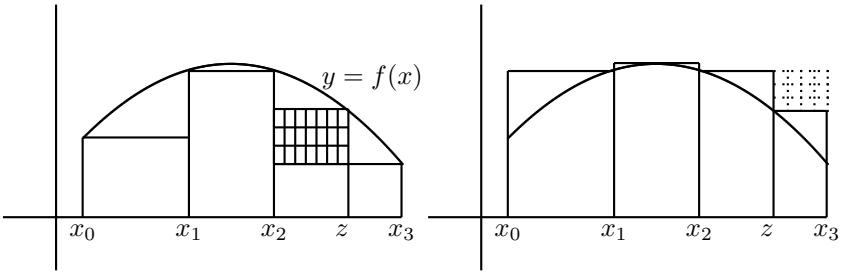
Also let $\Delta x_i \equiv x_i - x_{i-1}$. Then define upper and lower sums as

$$U(f,P) \equiv \sum_{i=1}^{n} M_i(f)\,\Delta x_i \text{ and } L(f,P) \equiv \sum_{i=1}^{n} m_i(f)\,\Delta x_i$$

respectively. The numbers, $M_i(f)$ and $m_i(f)$, are well defined real numbers because f is assumed to be bounded and $\mathbb{R}$ is complete. Thus the set $S = \{f(x) : x \in [x_{i-1}, x_i]\}$ is bounded above and below. In the following picture, the sum of the areas of the rectangles in the picture on the left is a lower sum for the function in the picture and the sum of the areas of the rectangles in the picture on the right is an upper sum for the same function which uses the same partition.



What happens when you add in more points in a partition? The following pictures illustrate in the context of the above example. In this example a single additional point, labeled z has been added in.



Note how the lower sum got larger by the amount of the area in the shaded rectangle and the upper sum got smaller by the amount in the rectangle shaded by dots. In general this is the way it works and this is shown in the following lemma.

Lemma 10.1.1 *If $P \subseteq Q$ then*

$$U(f,Q) \leq U(f,P), \text{ and } L(f,P) \leq L(f,Q).$$

Proof: This is verified by adding in one point at a time. Thus let $P = \{x_0, \cdots, x_n\}$ and let $Q = \{x_0, \cdots, x_k, y, x_{k+1}, \cdots, x_n\}$. Thus exactly one point, y, is added between x_k and x_{k+1}. Now the term in the upper sum which corresponds to the interval $[x_k, x_{k+1}]$ in $U(f,P)$ is

$$\sup\{f(x) : x \in [x_k, x_{k+1}]\}(x_{k+1} - x_k) \tag{10.1}$$

and the term which corresponds to the interval $[x_k, x_{k+1}]$ in $U(f,Q)$ is

$$\sup\{f(x) : x \in [x_k, y]\}(y - x_k) + \sup\{f(x) : x \in [y, x_{k+1}]\}(x_{k+1} - y) \tag{10.2}$$

$$\equiv M_1(y - x_k) + M_2(x_{k+1} - y) \tag{10.3}$$

All the other terms in the two sums coincide. Now $\sup\{f(x) : x \in [x_k, x_{k+1}]\} \geq \max(M_1, M_2)$ and so the expression in (10.2) is no larger than

$$\sup\{f(x) : x \in [x_k, x_{k+1}]\}(x_{k+1} - y) + \sup\{f(x) : x \in [x_k, x_{k+1}]\}(y - x_k)$$

$$= \sup\{f(x) : x \in [x_k, x_{k+1}]\}(x_{k+1} - x_k),$$

the term corresponding to the interval, $[x_k, x_{k+1}]$ and $U(f,P)$. This proves the first part of the lemma pertaining to upper sums because if $Q \supseteq P$, one can obtain Q from P by adding in one point at a time and each time a point is added, the corresponding upper sum either gets smaller or stays the same. The second part is similar and is left as an exercise.

Lemma 10.1.2 *If P and Q are two partitions, then*

$$L(f,P) \leq U(f,Q).$$

Proof: By Lemma 10.1.1,

$$L(f,P) \leq L(f,P \cup Q) \leq U(f,P \cup Q) \leq U(f,Q).$$

Definition 10.1.3

$$\overline{I} \equiv \inf\{U(f,Q) \text{ where } Q \text{ is a partition}\}$$

$$\underline{I} \equiv \sup\{L(f,P) \text{ where } P \text{ is a partition}\}.$$

Note that $\underline{I}$ and $\overline{I}$ are well defined real numbers.

Theorem 10.1.4 *$\underline{I} \leq \overline{I}$.*

Proof: From Lemma 10.1.2,

$$\underline{I} = \sup\{L(f,P) \text{ where } P \text{ is a partition}\} \leq U(f,Q)$$

because $U(f,Q)$ is an upper bound to the set of all lower sums and so it is no smaller than the least upper bound. Therefore, since Q is arbitrary,

$$\begin{aligned}\underline{I} &= \sup\{L(f,P) \text{ where } P \text{ is a partition}\} \\ &\leq \inf\{U(f,Q) \text{ where } Q \text{ is a partition}\} \equiv \overline{I}\end{aligned}$$

where the inequality holds because it was just shown that $\underline{I}$ is a lower bound to the set of all upper sums and so it is no larger than the greatest lower bound of this set. This proves the theorem.

Definition 10.1.5 *A bounded function f is Riemann integrable, written as*

$$f \in R([a,b])$$

if

$$\underline{I} = \overline{I}$$

and in this case,

$$\int_a^b f(x)\,dx \equiv \underline{I} = \overline{I}.$$

Thus, in words, the Riemann integral is the unique number which lies between all upper sums and all lower sums if there is such a unique number.

Recall Proposition 2.14.3. It is stated here for ease of reference.

Proposition 10.1.6 *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 $\delta > 0$,

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

This proposition implies the following theorem which is used to determine the question of Riemann integrability.

Theorem 10.1.7 *A bounded function f is Riemann integrable if and only if for all $\varepsilon > 0$, there exists a partition P such that*

$$U(f,P) - L(f,P) \ < \varepsilon. \tag{10.4}$$

Proof: First assume f is Riemann integrable. Then let P and Q be two partitions such that

$$U(f,Q) < \overline{I} + \varepsilon/2,\ L(f,P) > \underline{I} - \varepsilon/2.$$

Then since $\underline{I} = \overline{I}$,

$$U(f,Q\cup P) - L(f,P\cup Q) \leq U(f,Q) - L(f,P) < \overline{I} + \varepsilon/2 - (\underline{I} - \varepsilon/2) = \varepsilon.$$

Now suppose that for all $\varepsilon > 0$ there exists a partition such that (10.4) holds. Then for given ε and partition P corresponding to ε

$$\overline{I} - \underline{I} \leq U(f,P) - L(f,P) \leq \varepsilon.$$

Since ε is arbitrary, this shows $\underline{I} = \overline{I}$ and this proves the theorem.

The condition described in the theorem is called the Riemann criterion .

Not all bounded functions are Riemann integrable. For example, let

$$f(x) \equiv \begin{cases} 1 \text{ if } x \in \mathbb{Q} \\ 0 \text{ if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases} \tag{10.5}$$

Then if $[a,b] = [0,1]$ all upper sums for f equal 1 while all lower sums for f equal 0. Therefore the Riemann criterion is violated for $\varepsilon = 1/2$.

10.2 Exercises

1. Prove the second half of Lemma 10.1.1 about lower sums.
2. Verify that for f given in (10.5), the lower sums on the interval $[0,1]$ are all equal to zero while the upper sums are all equal to one.
3. Let $f(x) = 1 + x^2$ for $x \in [-1,3]$ and let $P = \left\{-1, -\frac{1}{3}, 0, \frac{1}{2}, 1, 2\right\}$. Find $U(f,P)$ and $L(f,P)$.
4. Show that if $f \in R([a,b])$, there exists a partition, $\{x_0, \cdots, x_n\}$ such that for any $z_k \in [x_k, x_{k+1}]$,
$$\left| \int_a^b f(x)\, dx - \sum_{k=1}^n f(z_k)(x_k - x_{k-1}) \right| < \varepsilon$$
This sum, $\sum_{k=1}^n f(z_k)(x_k - x_{k-1})$, is called a Riemann sum and this exercise shows that the integral can always be approximated by a Riemann sum.
5. Let $P = \left\{1, 1\frac{1}{4}, 1\frac{1}{2}, 1\frac{3}{4}, 2\right\}$. Find upper and lower sums for the function, $f(x) = \frac{1}{x}$ using this partition. What does this tell you about $\ln(2)$?
6. If $f \in R([a,b])$ and f is changed at finitely many points, show the new function is also in $R([a,b])$ and has the same integral as the unchanged function.
7. Consider the function, $y = x^2$ for $x \in [0,1]$. Show this function is Riemann integrable and find the integral using the definition and the formula
$$\sum_{k=1}^n k^2 = \frac{1}{3}(n+1)^3 - \frac{1}{2}(n+1)^2 + \frac{1}{6}(n+1)$$
which you should verify by using math induction. This is not a practical way to find integrals in general.
8. Define a "left sum" as
$$\sum_{k=1}^n f(x_{k-1})(x_k - x_{k-1})$$
and a "right sum",
$$\sum_{k=1}^n f(x_k)(x_k - x_{k-1}).$$

Also suppose that all partitions have the property that $x_k - x_{k-1}$ equals a constant, $(b-a)/n$ so the points in the partition are equally spaced, and define the integral to be the number these right and left sums get close to as n gets larger and larger. Show that for f given in (10.5), $\int_0^x f(t)\,dt = 1$ if x is rational and $\int_0^x f(t)\,dt = 0$ if x is irrational. It turns out that the correct answer should always equal zero for that function, regardless of whether x is rational. This is shown in more advanced courses when the Lebesgue integral is studied. This illustrates why the method of defining the integral in terms of left and right sums is total nonsense.

10.3 Functions Of Riemann Integrable Functions

It is often necessary to consider functions of Riemann integrable functions and a natural question is whether these are Riemann integrable. The following theorem gives a partial answer to this question. This is not the most general theorem which will relate to this question but it will be enough for the needs of this book.

Theorem 10.3.1 *Let f, g be bounded functions and let $f([a,b]) \subseteq [c_1,d_1]$ and $g([a,b]) \subseteq [c_2,d_2]$. Let $H : [c_1,d_1] \times [c_2,d_2] \to \mathbb{R}$ satisfy,*

$$|H(a_1,b_1) - H(a_2,b_2)| \le K[|a_1-a_2| + |b_1-b_2|]$$

for some constant K. Then if $f, g \in R([a,b])$ it follows that $H \circ (f,g) \in R([a,b])$.

Proof: In the following claim, $M_i(h)$ and $m_i(h)$ have the meanings assigned above with respect to some partition of $[a,b]$ for the function, h.

Claim: The following inequality holds.

$$|M_i(H\circ(f,g)) - m_i(H\circ(f,g))| \le$$
$$K[|M_i(f) - m_i(f)| + |M_i(g) - m_i(g)|].$$

Proof of the claim: By the above proposition, there exist $x_1, x_2 \in [x_{i-1}, x_i]$ be such that

$$H(f(x_1), g(x_1)) + \eta > M_i(H\circ(f,g)),$$

and

$$H(f(x_2), g(x_2)) - \eta < m_i(H\circ(f,g)).$$

Then

$$|M_i(H\circ(f,g)) - m_i(H\circ(f,g))|$$
$$< 2\eta + |H(f(x_1),g(x_1)) - H(f(x_2),g(x_2))|$$
$$< 2\eta + K[|f(x_1)-f(x_2)| + |g(x_1)-g(x_2)|]$$
$$\le 2\eta + K[|M_i(f) - m_i(f)| + |M_i(g) - m_i(g)|].$$

Since $\eta > 0$ is arbitrary, this proves the claim.

Now continuing with the proof of the theorem, let P be such that

$$\sum_{i=1}^n (M_i(f) - m_i(f))\Delta x_i < \frac{\varepsilon}{2K},\ \sum_{i=1}^n (M_i(g) - m_i(g))\Delta x_i < \frac{\varepsilon}{2K}.$$

Then from the claim,

$$\sum_{i=1}^n (M_i(H\circ(f,g)) - m_i(H\circ(f,g)))\Delta x_i$$

$$< \sum_{i=1}^{n} K \left[|M_i(f) - m_i(f)| + |M_i(g) - m_i(g)| \right] \Delta x_i < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this shows $H \circ (f,g)$ satisfies the Riemann criterion and hence $H \circ (f,g)$ is Riemann integrable as claimed. This proves the theorem.

This theorem implies that if f, g are Riemann integrable, then so is $af + bg, |f|, f^2$, along with infinitely many other such continuous combinations of Riemann integrable functions. For example, to see that $|f|$ is Riemann integrable, let $H(a,b) = |a|$. Clearly this function satisfies the conditions of the above theorem and so $|f| = H(f,f) \in R([a,b])$ as claimed. The following theorem gives an example of many functions which are Riemann integrable.

Theorem 10.3.2 *Let $f : [a,b] \to \mathbb{R}$ be either increasing or decreasing on $[a,b]$. Then $f \in R([a,b])$.*

Proof: Let $\varepsilon > 0$ be given and let

$$x_i = a + i\left(\frac{b-a}{n}\right), \ i = 0, \cdots, n.$$

Then since f is increasing,

$$\begin{aligned} U(f,P) - L(f,P) &= \sum_{i=1}^{n} (f(x_i) - f(x_{i-1})) \left(\frac{b-a}{n}\right) \\ &= (f(b) - f(a)) \left(\frac{b-a}{n}\right) < \varepsilon \end{aligned}$$

whenever n is large enough. Thus the Riemann criterion is satisfied and so the function is Riemann integrable. The proof for decreasing f is similar.

Corollary 10.3.3 *Let $[a,b]$ be a bounded closed interval and let $\phi : [a,b] \to \mathbb{R}$ be Lipschitz continuous. Then $\phi \in R([a,b])$. Recall that a function, ϕ, is Lipschitz continuous if there is a constant, K, such that for all x, y,*

$$|\phi(x) - \phi(y)| < K|x - y|.$$

Proof: Let $f(x) = x$. Then by Theorem 10.3.2, f is Riemann integrable. Let $H(a,b) \equiv \phi(a)$. Then by Theorem 10.3.1 $H \circ (f,f) = \phi \circ f = \phi$ is also Riemann integrable. This proves the corollary.

In fact, it is enough to assume ϕ is continuous, although this is harder. This is the content of the next theorem which is where the difficult theorems about continuity and uniform continuity are used.

Theorem 10.3.4 *Suppose $f : [a,b] \to \mathbb{R}$ is continuous. Then $f \in R([a,b])$.*

Proof: By Corollary 5.13.5 on Page 125, f is uniformly continuous on $[a,b]$. Therefore, if $\varepsilon > 0$ is given, there exists a $\delta > 0$ such that if $|x_i - x_{i-1}| < \delta$, then $M_i - m_i < \frac{\varepsilon}{b-a}$. Let

$$P \equiv \{x_0, \cdots, x_n\}$$

be a partition with $|x_i - x_{i-1}| < \delta$. Then

$$U(f,P) - L(f,P) < \sum_{i=1}^{n} (M_i - m_i)(x_i - x_{i-1}) < \frac{\varepsilon}{b-a}(b-a) = \varepsilon.$$

By the Riemann criterion, $f \in R([a,b])$. This proves the theorem.

10.4 Properties Of The Integral

The integral has many important algebraic properties. First here is a simple lemma.

Lemma 10.4.1 *Let S be a nonempty set which is bounded above and below. Then if $-S \equiv \{-x : x \in S\}$,*

$$\sup(-S) = -\inf(S) \tag{10.6}$$

and

$$\inf(-S) = -\sup(S). \tag{10.7}$$

Proof: Consider (10.6). Let $x \in S$. Then $-x \leq \sup(-S)$ and so $x \geq -\sup(-S)$. If follows that $-\sup(-S)$ is a lower bound for S and therefore, $-\sup(-S) \leq \inf(S)$. This implies $\sup(-S) \geq -\inf(S)$. Now let $-x \in -S$. Then $x \in S$ and so $x \geq \inf(S)$ which implies $-x \leq -\inf(S)$. Therefore, $-\inf(S)$ is an upper bound for $-S$ and so $-\inf(S) \geq \sup(-S)$. This shows (10.6). Formula (10.7) is similar and is left as an exercise.

In particular, the above lemma implies that for $M_i(f)$ and $m_i(f)$ defined above $M_i(-f) = -m_i(f)$, and $m_i(-f) = -M_i(f)$.

Lemma 10.4.2 *If $f \in R([a,b])$ then $-f \in R([a,b])$ and*

$$-\int_a^b f(x)\,dx = \int_a^b -f(x)\,dx.$$

Proof: The first part of the conclusion of this lemma follows from Theorem 10.3.2 since the function $\phi(y) \equiv -y$ is Lipschitz continuous. Now choose P such that

$$\int_a^b -f(x)\,dx - L(-f,P) < \varepsilon.$$

Then since $m_i(-f) = -M_i(f)$,

$$\varepsilon > \int_a^b -f(x)\,dx - \sum_{i=1}^n m_i(-f)\,\Delta x_i = \int_a^b -f(x)\,dx + \sum_{i=1}^n M_i(f)\,\Delta x_i$$

which implies

$$\varepsilon > \int_a^b -f(x)\,dx + \sum_{i=1}^n M_i(f)\,\Delta x_i \geq \int_a^b -f(x)\,dx + \int_a^b f(x)\,dx.$$

Thus, since ε is arbitrary,

$$\int_a^b -f(x)\,dx \leq -\int_a^b f(x)\,dx$$

whenever $f \in R([a,b])$. It follows

$$\int_a^b -f(x)\,dx \leq -\int_a^b f(x)\,dx = -\int_a^b -(-f(x))\,dx \leq \int_a^b -f(x)\,dx$$

and this proves the lemma.

Theorem 10.4.3 *The integral is linear,*

$$\int_a^b (\alpha f + \beta g)(x)\,dx = \alpha \int_a^b f(x)\,dx + \beta \int_a^b g(x)\,dx.$$

whenever $f, g \in R([a,b])$ and $\alpha, \beta \in \mathbb{R}$.

Proof: First note that by Theorem 10.3.1, $\alpha f + \beta g \in R([a,b])$. To begin with, consider the claim that if $f, g \in R([a,b])$ then

$$\int_a^b (f+g)(x)\, dx = \int_a^b f(x)\, dx + \int_a^b g(x)\, dx. \tag{10.8}$$

Let P_1,Q_1 be such that

$$U(f,Q_1) - L(f,Q_1) < \varepsilon/2,\ U(g,P_1) - L(g,P_1) < \varepsilon/2.$$

Then letting $P \equiv P_1 \cup Q_1$, Lemma 10.1.1 implies

$$U(f,P) - L(f,P) < \varepsilon/2, \text{ and } U(g,P) - U(g,P) < \varepsilon/2.$$

Next note that

$$m_i(f+g) \geq m_i(f) + m_i(g),\ M_i(f+g) \leq M_i(f) + M_i(g).$$

Therefore,

$$L(g+f,P) \geq L(f,P) + L(g,P),\ U(g+f,P) \leq U(f,P) + U(g,P).$$

For this partition,

$$\begin{aligned}\int_a^b (f+g)(x)\, dx &\in [L(f+g,P), U(f+g,P)] \\ &\subseteq [L(f,P) + L(g,P), U(f,P) + U(g,P)]\end{aligned}$$

and

$$\int_a^b f(x)\, dx + \int_a^b g(x)\, dx \in [L(f,P) + L(g,P), U(f,P) + U(g,P)].$$

Therefore,

$$\left| \int_a^b (f+g)(x)\, dx - \left(\int_a^b f(x)\, dx + \int_a^b g(x)\, dx \right) \right| \leq$$

$$U(f,P) + U(g,P) - (L(f,P) + L(g,P)) < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

This proves (10.8) since ε is arbitrary.

It remains to show that

$$\alpha \int_a^b f(x)\, dx = \int_a^b \alpha f(x)\, dx.$$

Suppose first that $\alpha \geq 0$. Then

$$\int_a^b \alpha f(x)\, dx \equiv \sup\{L(\alpha f, P) : P \text{ is a partition}\} =$$

$$\alpha \sup\{L(f,P) : P \text{ is a partition}\} \equiv \alpha \int_a^b f(x)\, dx.$$

If $\alpha < 0$, then this and Lemma 10.4.2 imply

$$\begin{aligned}\int_a^b \alpha f(x)\, dx &= \int_a^b (-\alpha)(-f(x))\, dx \\ &= (-\alpha) \int_a^b (-f(x))\, dx = \alpha \int_a^b f(x)\, dx.\end{aligned}$$

This proves the theorem.

Theorem 10.4.4 *If $f \in R([a,b])$ and $f \in R([b,c])$, then $f \in R([a,c])$ and*

$$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx. \tag{10.9}$$

Proof: Let P_1 be a partition of $[a,b]$ and P_2 be a partition of $[b,c]$ such that

$$U(f,P_i) - L(f,P_i) < \varepsilon/2,\ i = 1,2.$$

Let $P \equiv P_1 \cup P_2$. Then P is a partition of $[a,c]$ and

$$\begin{aligned} &U(f,P) - L(f,P) \\ &= U(f,P_1) - L(f,P_1) + U(f,P_2) - L(f,P_2) < \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned} \tag{10.10}$$

Thus, $f \in R([a,c])$ by the Riemann criterion and also for this partition,

$$\begin{aligned} \int_a^b f(x)\,dx + \int_b^c f(x)\,dx &\in [L(f,P_1) + L(f,P_2), U(f,P_1) + U(f,P_2)] \\ &= [L(f,P), U(f,P)] \end{aligned}$$

and

$$\int_a^c f(x)\,dx \in [L(f,P), U(f,P)].$$

Hence by (10.10),

$$\left| \int_a^c f(x)\,dx - \left(\int_a^b f(x)\,dx + \int_b^c f(x)\,dx \right) \right| < U(f,P) - L(f,P) < \varepsilon$$

which shows that since ε is arbitrary, (10.9) holds. This proves the theorem.

Corollary 10.4.5 *Let $[a,b]$ be a closed and bounded interval and suppose that*

$$a = y_1 < y_2 \cdots < y_l = b$$

and that f is a bounded function defined on $[a,b]$ which has the property that f is either increasing on $[y_j, y_{j+1}]$ or decreasing on $[y_j, y_{j+1}]$ for $j = 1, \cdots, l-1$. Then $f \in R([a,b])$.

Proof: This follows from Theorem 10.4.4 and Theorem 10.3.2.
The symbol, $\int_a^b f(x)\,dx$ when $a > b$ has not yet been defined.

Definition 10.4.6 *Let $[a,b]$ be an interval and let $f \in R([a,b])$. Then*

$$\int_b^a f(x)\,dx \equiv -\int_a^b f(x)\,dx.$$

Note that with this definition,

$$\int_a^a f(x)\,dx = -\int_a^a f(x)\,dx$$

and so

$$\int_a^a f(x)\,dx = 0.$$

Theorem 10.4.7 *Assuming all the integrals make sense,*

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx.$$

Proof: This follows from Theorem 10.4.4 and Definition 10.4.6. For example, assume

$$c \in (a,b).$$

Then from Theorem 10.4.4,

$$\int_a^c f(x)\,dx + \int_c^b f(x)\,dx = \int_a^b f(x)\,dx$$

and so by Definition 10.4.6,

$$\begin{aligned}\int_a^c f(x)\,dx &= \int_a^b f(x)\,dx - \int_c^b f(x)\,dx \\ &= \int_a^b f(x)\,dx + \int_b^c f(x)\,dx.\end{aligned}$$

The other cases are similar.

The following properties of the integral have either been established or they follow quickly from what has been shown so far.

$$\text{If } f \in R([a,b]) \text{ then if } c \in [a,b],\ f \in R([a,c]), \tag{10.11}$$

$$\int_a^b \alpha\,dx = \alpha(b-a), \tag{10.12}$$

$$\int_a^b (\alpha f + \beta g)(x)\,dx = \alpha\int_a^b f(x)\,dx + \beta\int_a^b g(x)\,dx, \tag{10.13}$$

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx, \tag{10.14}$$

$$\int_a^b f(x)\,dx \geq 0 \text{ if } f(x) \geq 0 \text{ and } a < b, \tag{10.15}$$

$$\left|\int_a^b f(x)\,dx\right| \leq \left|\int_a^b |f(x)|\,dx\right|. \tag{10.16}$$

The only one of these claims which may not be completely obvious is the last one. To show this one, note that

$$|f(x)| - f(x) \geq 0,\ |f(x)| + f(x) \geq 0.$$

Therefore, by (10.15) and (10.13), if $a < b$,

$$\int_a^b |f(x)|\,dx \geq \int_a^b f(x)\,dx$$

and

$$\int_a^b |f(x)|\,dx \geq -\int_a^b f(x)\,dx.$$

Therefore,

$$\int_a^b |f(x)|\,dx \geq \left|\int_a^b f(x)\,dx\right|.$$

If $b < a$ then the above inequality holds with a and b switched. This implies (10.16).

10.5 Fundamental Theorem Of Calculus

With these properties, it is easy to prove the fundamental theorem of calculus[2]. Let $f \in R([a,b])$. Then by (10.11) $f \in R([a,x])$ for each $x \in [a,b]$. The first version of the fundamental theorem of calculus is a statement about the derivative of the function

$$x \to \int_a^x f(t)\, dt.$$

Theorem 10.5.1 *Let $f \in R([a,b])$ and let*

$$F(x) \equiv \int_a^x f(t)\, dt.$$

Then if f is continuous at $x \in (a,b)$,

$$F'(x) = f(x).$$

Proof: Let $x \in (a,b)$ be a point of continuity of f and let h be small enough that $x+h \in [a,b]$. Then by using (10.14),

$$h^{-1}(F(x+h) - F(x)) = h^{-1}\int_x^{x+h} f(t)\, dt.$$

Also, using (10.12),

$$f(x) = h^{-1}\int_x^{x+h} f(x)\, dt.$$

Therefore, by (10.16),

$$\left|h^{-1}(F(x+h) - F(x)) - f(x)\right| = \left|h^{-1}\int_x^{x+h} (f(t) - f(x))\, dt\right|$$

$$\leq \left|h^{-1}\int_x^{x+h} |f(t) - f(x)|\, dt\right|.$$

Let $\varepsilon > 0$ and let $\delta > 0$ be small enough that if $|t-x| < \delta$, then

$$|f(t) - f(x)| < \varepsilon.$$

Therefore, if $|h| < \delta$, the above inequality and (10.12) shows that

$$\left|h^{-1}(F(x+h) - F(x)) - f(x)\right| \leq |h|^{-1}\varepsilon |h| = \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this shows

$$\lim_{h\to 0} h^{-1}(F(x+h) - F(x)) = f(x)$$

and this proves the theorem.

Note this gives existence for the initial value problem,

$$F'(x) = f(x),\ F(a) = 0$$

whenever f is Riemann integrable and continuous.[3]

The next theorem is also called the fundamental theorem of calculus.

[2]This theorem is why Newton and Liebnitz are credited with inventing calculus. The integral had been around for thousands of years and the derivative was by their time well known. However the connection between these two ideas had not been fully made although Newton's predecessor, Isaac Barrow had made some progress in this direction.

[3]Of course it was proved that if f is continuous on a closed interval, $[a,b]$, then $f \in R([a,b])$ but this is a hard theorem using the difficult result about uniform continuity.

Theorem 10.5.2 *Let $f \in R([a,b])$ and suppose there exists an antiderivative for f, G, such that*

$$G'(x) = f(x)$$

for every point of (a,b) and G is continuous on $[a,b]$. Then

$$\int_a^b f(x)\,dx = G(b) - G(a). \tag{10.17}$$

Proof: Let $P = \{x_0, \cdots, x_n\}$ be a partition satisfying

$$U(f,P) - L(f,P) < \varepsilon.$$

Then

$$\begin{aligned} G(b) - G(a) &= G(x_n) - G(x_0) \\ &= \sum_{i=1}^{n} G(x_i) - G(x_{i-1}). \end{aligned}$$

By the mean value theorem,

$$\begin{aligned} G(b) - G(a) &= \sum_{i=1}^{n} G'(z_i)(x_i - x_{i-1}) \\ &= \sum_{i=1}^{n} f(z_i)\,\Delta x_i \end{aligned}$$

where z_i is some point in $[x_{i-1}, x_i]$. It follows, since the above sum lies between the upper and lower sums, that

$$G(b) - G(a) \in [L(f,P), U(f,P)],$$

and also

$$\int_a^b f(x)\,dx \in [L(f,P), U(f,P)].$$

Therefore,

$$\left| G(b) - G(a) - \int_a^b f(x)\,dx \right| < U(f,P) - L(f,P) < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, (10.17) holds. This proves the theorem.

The following notation is often used in this context. Suppose F is an antiderivative of f as just described with F continuous on $[a,b]$ and $F' = f$ on (a,b). Then

$$\int_a^b f(x)\,dx = F(b) - F(a) \equiv F(x)\,|_a^b.$$

Recall how many interesting problems can be reduced to initial value problems. This was true of work, area, arc length and many other examples. The examples given were cooked uup to work out and you could actually solve the initial value problem using known functions. What if you couldn't do this? The next theorem is a significant existence theorem which tells you that solutions of the initial value problem exist.

Theorem 10.5.3 *Suppose f is a continuous function defined on an interval, (a,b), $c \in (a,b)$, and $y_0 \in \mathbb{R}$. Then there exists a unique solution to the initial value problem,*

$$F'(x) = f(x),\ F(c) = y_0.$$

This solution is given by

$$F(x) = y_0 + \int_c^x f(t)\,dt. \tag{10.18}$$

Proof: From Theorem 10.3.4, it follows the integral in (10.18) is well defined. Now by the fundamental theorem of calculus, $F'(x) = f(x)$. Therefore, F solves the given differential equation. Also, $F(c) = y_0 + \int_c^c f(t)\,dt = y_0$ so the initial condition is also satisfied. This establishes the existence part of the theorem.

Suppose F and G both solve the initial value problem. Then $F'(x) - G'(x) = f(x) - f(x) = 0$ and so $F(x) - G(x) = C$ for some constant, C. However, $F(c) - G(c) = y_0 - y_0 = 0$ and so the constant C can only equal 0. This proves the uniqueness part of the theorem.

Example 10.5.4 *Find the area between $y = x^2$ and $y = x^3$ for $x \in [0,1]$.*

You need to solve the initial value problem, $A'(x) = x^2 - x^3, A'(0) = 0$. The answer is then $A(1)$. By Theorem 10.5.3, $A(x) = 0 + \int_0^x \left(t^2 - t^3\right) dt$ and so the answer is

$$A(1) = \int_0^1 \left(t^2 - t^3\right) dt = \frac{1}{12}$$

Example 10.5.5 *A cylinder standing on its end is filled with water which weighs 62.5 pounds per cubic foot is 5 feet in radius and is 10 feet tall. Find the total force on the sides of the cylinder.*

Remember the force acting on the sides of a horizontal slice of the cylinder y feet below the top of the fluid is $(2\pi \times 5 \times 62.5 \times y)\,dy = dF$. Thus you need to find a solution to the initial value problem

$$\frac{dF}{dy} = (2\pi \times 5 \times 62.5 \times y),\ F(0) = 0$$

and in this case, the answer equals $F(10)$. By Theorem 10.5.3 and Theorem 10.5.2,

$$F(10) = \int_0^{10} (2\pi)(5)(62.5)(y)\,dy = 98175 \text{ pounds.}$$

An informal way of looking at this is $dF = (2\pi \times 5 \times 62.5 \times y)\,dy$ for y between 0 and 10. The dF is an "infinitesimal" piece of the total force. To get the total force you just sum the $dF's$. Thus $F = \int_0^{10} (2\pi)(5)(62.5)(y)\,dy$. The reason the integral sign looks like an S is that it is a sort of a sum. You are summing up "infinitesimal" contributions to obtain the total. This last statement is mathematical gobeldygook. I have told you nothing about infinitesimals. I have only used the term in an evocative manner. However, it turns out to be a useful way of thinking about things because it allows you to remember what to do without fussing with the initial value problems.

Example 10.5.6 *Find the area of the surface of revolution formed by revolving $y = x^2 - \frac{1}{8}\ln x$ about the x axis for $x \in [1,2]$.*

An infinitesimal contribution is of the form

$$dA = 2\pi\left(x^2 - \frac{1}{8}\ln x\right)\sqrt{1+\left(\frac{1}{8}\frac{16x^2-1}{x}\right)^2}dx = 2\pi\left(x^2 - \frac{1}{8}\ln x\right)\frac{1}{8}\frac{1+16x^2}{x}dx$$

Summing these with the integral and using Theorem 10.5.2 yields

$$\int_1^2 2\pi\left(x^2 - \frac{1}{8}\ln x\right)\frac{1}{8}\frac{1+16x^2}{x}dx = \frac{63}{4}\pi - \frac{1}{64}\pi\ln^2 2 - \pi\ln 2$$

Isn't it great that things worked out to get a nice answer? What if it hadn't. Sometimes things don't work out.

Example 10.5.7 *Find the area of the surface of revolution formed by revolving $y = x^2 - \ln x$ about the x axis for $x \in [1,2]$.*

In this case an infinitesimal contribution to this area is

$$dA = 2\pi\left(x^2 - \ln x\right)\frac{1}{x}\sqrt{4x^4 - 3x^2 + 1}dx$$

and so the total area is given by the integral,

$$A = \int_1^2 2\pi\left(x^2 - \ln x\right)\frac{1}{x}\sqrt{4x^4 - 3x^2 + 1}dx.$$

What now? Can you find an antiderivative in terms of functions you have names for and use Theorem 10.5.2 to evaluate it? If you want to waste lots of time, give it a try. If you can do this one, it is very easy to find important examples that you can't do by finding antiderivatives in terms of known functions. Examples are integrals which have integrands equal to $\sin x/x, e^{-x^2}, \sqrt{1+x^4}, \sin\left(x^2\right)$, and many others. Sometimes these occur in important applications. What do you do then when Theorem 10.5.2 has failed? The answer is you use a numerical method to compute the integral. The most rudimentary numerical method goes right to the definition of the integral. It is described in the following definition. A much better method is described in the exercises. An extensive study of numerical methods for finding integrals will not be attempted in this book but it is very important you be aware that such methods are available. When you use a calculator or computer algebra system to compute an integral numerically, the machine is using a numerical method, probably one which is much more sophisticated than any presented in beginning calculus.

Definition 10.5.8 *Let f be a bounded function defined on a closed interval $[a,b]$ and let $P \equiv \{x_0, \cdots, x_n\}$ be a partition of the interval. Suppose $z_i \in [x_{i-1}, x_i]$ is chosen. Then the sum*

$$\sum_{i=1}^n f\left(z_i\right)\left(x_i - x_{i-1}\right)$$

is known as a Riemann sum. Also,

$$||P|| \equiv \max\left\{|x_i - x_{i-1}| : i = 1, \cdots, n\right\}.$$

Proposition 10.5.9 *Suppose $f \in R([a,b])$. Then there exists a partition, $P \equiv \{x_0, \cdots, x_n\}$ with the property that for any choice of $z_k \in [x_{k-1}, x_k]$,*

$$\left|\int_a^b f(x)\,dx - \sum_{k=1}^n f\left(z_k\right)\left(x_k - x_{k-1}\right)\right| < \varepsilon.$$

Proof: Choose P such that $U(f,P) - L(f,P) < \varepsilon$ and then both $\int_a^b f(x)\,dx$ and $\sum_{k=1}^n f(z_k)(x_k - x_{k-1})$ are contained in $[L(f,P), U(f,P)]$ and so the claimed inequality must hold. This proves the proposition.

It is significant because it gives a way of approximating the integral.

Example 10.5.10 *Use a Riemann sum to approximate* $\int_0^1 \sin\left(x^2\right) dx$.

I will use the partition, $\left\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\right\}$ and I will pick the midpoint of each subinterval to evaluate the function. This particular Riemann sum is called the **midpoint sum**. Thus the Riemann sum approximating the integral is

$$\sin\left(\frac{1}{64}\right)\left(\frac{1}{4}\right) + \sin\left(\left(\frac{3}{8}\right)^2\right)\frac{1}{4} + \sin\left(\left(\frac{5}{8}\right)^2\right)\frac{1}{4} + \sin\left(\left(\frac{7}{8}\right)^2\right)\frac{1}{4} = .307\,39$$

Using a computer algebra system to evaluate this integral gives

$$\int_0^1 \sin\left(x^2\right) dx = .310\,27$$

so you see in this case, the primitive Riemann sum approach yielded a pretty good answer. Of course it is important to know how close you are to the true answer and this involves the concept of error estimates. The following is a rudimentary error estimate which tells how well such a mid point sum approximates an integral.

Theorem 10.5.11 *Let f be a continuous function defined on the interval, $[a,b]$ and suppose also that f' exists on (a,b) and there exists $K \geq |f'(x)|$ for all $x \in (a,b)$. Suppose also that $\{x_0, x_1, \cdots, x_n\}$ is a partition in which the points are equally spaced. Thus $x_k = a + k\frac{b-a}{n}$. Then*

$$\left| \textit{midpoint sum} - \int_a^b f(x)\,dx \right| < K\frac{(b-a)^2}{n}.$$

Proof: You estimate the error on each interval, $[x_{k-1}, x_k]$ and then sum up these estimates. The contribution of the midpoint sum on this interval equals $f\left(\frac{x_{k-1}+x_k}{2}\right)\left(\frac{b-a}{n}\right)$. Thus the error corresponding to this interval is $\left|f\left(\frac{x_{k-1}+x_k}{2}\right)\left(\frac{b-a}{n}\right) - \int_{x_{k-1}}^{x_k} f(t)\,dt\right|$. By the mean value theorem for integrals, Problem 11 on Page 279 there exists $z_k \in (x_{k-1}, x_k)$ such that

$$\int_{x_{k-1}}^{x_k} f(t)\,dt = f(z_k)(x_k - x_{k-1}) = f(z_k)\left(\frac{b-a}{n}\right).$$

Therefore, this error equals

$$\left|f\left(\frac{x_{k-1}+x_k}{2}\right) - f(z_k)\right|\left(\frac{b-a}{n}\right)$$

which by the mean value theorem is no more than

$$|f'(w_k)|\left|\frac{x_{k-1}+x_k}{2} - z_k\right|\left(\frac{b-a}{n}\right) \leq K\left(\frac{b-a}{2n}\right)\left(\frac{b-a}{n}\right).$$

Since there are n of these, it follows the total error is no more than $K\frac{(b-a)^2}{n}$ as claimed. This proves the theorem.

Note that to get an error estimate, you need information on the derivative of the function, something which has nothing to do with the existence of the integral. This is typical of error estimates in numerical methods. You need to know something extra beyond what you need to get existence.

10.6 The Riemann Integral

The definition of Riemann integrability given in this chapter is also called Darboux integrability and the integral defined as the unique number which lies between all upper sums and all lower sums which is given in this chapter is called the Darboux integral . The definition of the Riemann integral in terms of Riemann sums is given next.

Definition 10.6.1 *A bounded function, f defined on $[a,b]$ is said to be Riemann integrable if there exists a number, I with the property that for every $\varepsilon > 0$, there exists $\delta > 0$ such that if*

$$P \equiv \{x_0, x_1, \cdots, x_n\}$$

is any partition having $||P|| < \delta$, and $z_i \in [x_{i-1}, x_i]$,

$$\left| I - \sum_{i=1}^{n} f(z_i)(x_i - x_{i-1}) \right| < \varepsilon.$$

The number $\int_a^b f(x)\,dx$ is defined as I.

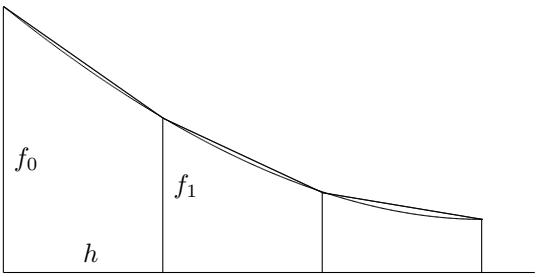
Thus, there are two definitions of the integral, this one in terms of Riemann sums and the earlier one which defined the integral to be the number which is between all the upper sums and lower sums. It turns out they are equivalent which is the following theorem of Darboux.

Theorem 10.6.2 *A bounded function defined on $[a,b]$ is Riemann integrable in the sense of Definition 10.6.1 if and only if it is integrable in the sense of Darboux. Furthermore the two integrals coincide.*

The proof of this theorem is left for the exercises in Problems 15 - 17. It isn't essential that you understand this theorem so if it does not interest you, leave it out. Note that it implies that given a Riemann integrable function f in either sense, it can be approximated by Riemann sums whenever $||P||$ is sufficiently small. Both versions of the integral are obsolete but entirely adequate for most applications and as a point of departure for a more up to date and satisfactory integral. The reason for using the Darboux approach to the integral is that all the existence theorems are easier to prove in this context.

10.7 Exercises

1. Let $F(x) = \int_{x^2}^{x^3} \frac{t^5+7}{t^7+87t^6+1}\,dt$. Find $F'(x)$.
2. Let $F(x) = \int_2^x \frac{1}{1+t^4}\,dt$. Sketch a graph of F and explain why it looks the way it does.
3. Use a midpoint sum with 4 subintervals to estimate $\int_1^2 \frac{1}{t} dt$ and give an estimate of how close your approximation is to the true answer.
4. Use a midpoint sum with 4 subintervals to estimate $\int_0^1 \frac{1}{t^2+1} dt$ and give an estimate of how close your approximation is to the true answer.
5. There is a general procedure for estimating the integral of a function, f on an interval, $[a,b]$. Form a uniform partition, $P = \{x_0, x_1, \cdots, x_n\}$ where for each j, $x_j - x_{j-1} = h$. Let $f_i = f(x_i)$ and assuming $f \geq 0$ on the interval $[x_{i-1}, x_i]$, approximate the area above this interval and under the curve with the area of a trapezoid having vertical sides, f_{i-1}, and f_i as shown in the following picture.



Thus $\frac{1}{2}\left(\frac{f_i+f_{i-1}}{2}\right)$ approximates the area under the curve. Show that adding these up yields

$$\frac{h}{2}\left[f_0+2f_1+\cdots+2f_{n-1}+f_n\right]$$

as an approximation to $\int_a^b f(x)\,dx$. This is known as the trapezoid rule. Verify that if $f(x)=mx+b$, the trapezoid rule gives the exact answer for the integral. Would this be true of upper and lower sums for such a function? Can you show that in the case of the function, $f(t)=1/t$ the trapezoid rule will always yield an answer which is too large for $\int_1^2 \frac{1}{t}dt$?

6. Let there be three equally spaced points, $x_{i-1}, x_{i-1}+h\equiv x_i$, and $x_i+2h\equiv x_{i+1}$. Suppose also a function, f, has the value f_{i-1} at x, f_i at $x+h$, and f_{i+1} at $x+2h$. Then consider

$$g_i(x)\equiv \frac{f_{i-1}}{2h^2}(x-x_i)(x-x_{i+1})-\frac{f_i}{h^2}(x-x_{i-1})(x-x_{i+1})+\frac{f_{i+1}}{2h^2}(x-x_{i-1})(x-x_i).$$

Check that this is a second degree polynomial which equals the values f_{i-1}, f_i, and f_{i+1} at the points x_{i-1}, x_i, and x_{i+1} respectively. The function, g_i is an approximation to the function, f on the interval $[x_{i-1}, x_{i+1}]$. Also,

$$\int_{x_{i-1}}^{x_{i+1}} g_i(x)\,dx$$

is an approximation to $\int_{x_{i-1}}^{x_{i+1}} f(x)\,dx$. Show $\int_{x_{i-1}}^{x_{i+1}} g_i(x)\,dx$ equals

$$\frac{hf_{i-1}}{3}+\frac{hf_i4}{3}+\frac{hf_{i+1}}{3}.$$

Now suppose n is even and $\{x_0, x_1, \cdots, x_n\}$ is a partition of the interval, $[a,b]$ and the values of a function, f defined on this interval are $f_i = f(x_i)$. Adding these approximations for the integral of f on the succession of intervals,

$$[x_0, x_2], [x_2, x_4], \cdots, [x_{n-2}, x_n],$$

show that an approximation to $\int_a^b f(x)\,dx$ is

$$\frac{h}{3}\left[f_0+4f_1+2f_2+4f_3+2f_2+\cdots+4f_{n-1}+f_n\right].$$

This is called Simpson's rule. Use Simpson's rule to compute an approximation to $\int_1^2 \frac{1}{t}dt$ letting $n = 4$. Compare with the answer from a calculator or computer.

7. A mine shaft has the shape $y = \cosh(.01x) - 5000$ where the units are in feet. Thus this mine shaft has a depth of a little less than one mile occuring when $x = 0$. The surface corresponds to $y = 0$ which occurs when $x = 921$ feet (approximately). A cable which weighs 5 pounds per foot extends from this point on the surface down to the very bottom of this mine shaft. Set up an integral which will give the work required to haul this cable up to the surface and use a computer algebra system or calculator to compute this integral. Next suppose there is a 5 ton ore skip on the bottom of this cable. How much work would be required to haul up the cable and the 5 ton skip of ore[4]?

8. Let a and b be positive numbers and consider the function,

$$F(x) = \int_0^{ax} \frac{1}{a^2 + t^2}\, dt + \int_b^{a/x} \frac{1}{a^2 + t^2}\, dt.$$

Show that F is a constant.

9. Solve the following initial value problem from ordinary differential equations which is to find a function y such that

$$y'(x) = \frac{x^7 + 1}{x^6 + 97x^5 + 7},\ y(10) = 5.$$

10. If $F, G \in \int f(x)\, dx$ for all $x \in \mathbb{R}$, show $F(x) = G(x) + C$ for some constant, C. Use this to give a different proof of the fundamental theorem of calculus which has for its conclusion $\int_a^b f(t)\, dt = G(b) - G(a)$ where $G'(x) = f(x)$.

11. Suppose f is continuous on $[a, b]$. Show there exists $c \in (a, b)$ such that

$$f(c) = \frac{1}{b-a} \int_a^b f(x)\ dx.$$

Hint: You might consider the function $F(x) \equiv \int_a^x f(t)\ dt$ and use the mean value theorem for derivatives and the fundamental theorem of calculus.

12. Suppose f and g are continuous functions on $[a, b]$ and that $g(x) \neq 0$ on (a, b). Show there exists $c \in (a, b)$ such that

$$f(c) \int_a^b g(x)\ dx = \int_a^b f(x)\, g(x)\ dx.$$

Hint: Define $F(x) \equiv \int_a^x f(t)\, g(t)\ dt$ and let $G(x) \equiv \int_a^x g(t)\ dt$. Then use the Cauchy mean value theorem on these two functions.

[4]In the uupper peninsula of Michigan, there are copper mines having inclined shafts which achieve a depth of roughly one mile. These shafts start off very steep and become less so toward the bottom to allow for the sagging in the cable. One such is shaft 2 for the Quincy mine in Hancock Michigan. The problem of hauling up so much heavy wire cable was solved by letting out cable and an empty ore skip on one of the dual skipways while taking up the cable and the full ore skip on the other. The Hoist which did the work of hauling the skip weighs 880 tons and is the largest seam hoist ever built. It was operated between 1920 and 1931. This hoist could bring up a 10 ton skip of ore at the rate of 16 meters per second. The huge steam hoist is housed in its own building and the cables from this hoist passed over head on large pulleys to the mine shaft. Once the cable broke. Imagine what happened at the top and at the bottom.

13. Consider the function
$$f(x) \equiv \begin{cases} \sin\left(\frac{1}{x}\right) \text{ if } x \neq 0 \\ 0 \text{ if } x = 0 \end{cases}.$$
Is f Riemann integrable? Explain why or why not.

14. Prove the second part of Theorem 10.3.2 about decreasing functions.

15. Suppose f is a bounded function defined on $[a,b]$ and $|f(x)| < M$ for all $x \in [a,b]$. Now let Q be a partition having n points, $\{x_0^*, \cdots, x_n^*\}$ and let P be any other partition. Show that
$$|U(f,P) - L(f,P)| \leq 2Mn\,||P|| + |U(f,Q) - L(f,Q)|.$$
Hint: Write the sum for $U(f,P) - L(f,P)$ and split this sum into two sums, the sum of terms for which $[x_{i-1}, x_i]$ contains at least one point of Q, and terms for which $[x_{i-1}, x_i]$ does not contain any points of Q. In the latter case, $[x_{i-1}, x_i]$ must be contained in some interval, $\left[x_{k-1}^*, x_k^*\right]$. Therefore, the sum of these terms should be no larger than $|U(f,Q) - L(f,Q)|$.

16. ↑ If $\varepsilon > 0$ is given and f is a Darboux integrable function defined on $[a,b]$, show there exists $\delta > 0$ such that whenever $||P|| < \delta$, then
$$|U(f,P) - L(f,P)| < \varepsilon.$$

17. ↑ Prove Theorem 10.6.2.

10.8 Return Of The Wild Assumption

The entire treatment of exponential functions and logarithms up till this time has been based on the Wild Assumption on Page 158. This was a totally unjustified assumption that exponential functions existed and were differentiable. Some pictures were drawn by a computer as evidence. Based on this outrageous wild assumption, logarithms were defined. It is now possible to establish Wild Assumption 7.3.1.

Define
$$L_1(x) \equiv \int_1^x \frac{1}{t}\,dt. \tag{10.19}$$

There is no problem in writing this integral because the function, $f(t) = 1/t$ is decreasing.

Theorem 10.8.1 *The function, $L_1 : (0,\infty) \to \mathbb{R}$ satisfies the following properties.*

$$L_1(xy) = L_1(x) + L_1(y),\ L_1(1) = 0, \tag{10.20}$$

The function, L_1 is one to one and onto, strictly increasing, and its graph is concave downward. In addition to this, whenever $\frac{m}{n} \in \mathbb{Q}$

$$L_1\left(\sqrt[n]{x^m}\right) = \frac{m}{n} L_1(x). \tag{10.21}$$

Proof: Fix $y > 0$ and let

$$f(x) = L_1(xy) - (L_1(x) + L_1(y))$$

Then by Theorem 6.2.6 on Page 134 and the Fundamental theorem of calculus, Theorem 10.5.1,

$$f'(x) = y\left(\frac{1}{xy}\right) - \frac{1}{x} = 0.$$

Therefore, by Corollary 6.8.4 on Page 147, $f(x)$ is a constant. However, $f(1) = 0$ and so this proves (10.20).

From the Fundamental theorem of calculus, Theorem 10.5.1, $L_1'(x) = \frac{1}{x} > 0$ and so L_1 is a strictly increasing function and is therefore one to one. Also the second derivative equals

$$L_1''(x) = \frac{-1}{x^2} < 0$$

showing that the graph of L_1 is concave down.

Now consider the assertion that L_1 is onto. First note that from the definition, $L_1(2) > 0$. In fact,

$$L_1(2) \geq 1/2$$

as can be seen by looking at a lower sum for $\int_1^2 (1/t)\,dt$. Now if $x > 0$

$$L_1(x \times x) = L_1 x + L_1 x = 2L_1 x.$$

Also,

$$\begin{aligned} L_1((x)(x)(x)) &= L_1((x)(x)) + L_1(x) \\ &= L_1(x) + L_1(x) + L_1(x) \\ &= 3L_1(x). \end{aligned}$$

Continuing in this way it follows that for any positive integer, n,

$$L_1(x^n) = nL_1(x). \tag{10.22}$$

Therefore, $L_1(x)$ achieves arbitrarily large values as x gets increasingly large because you can take $x = 2$ in (10.22) and use the definition of L_1 to verify that $L_1(2) > 0$. Now if $x > 0$

$$0 = L_1\left(\overbrace{\left(\frac{1}{x}\right)(x)}^{=1}\right) = L_1\left(\frac{1}{x}\right) + L_1(x)$$

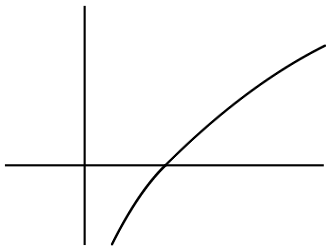
showing that

$$L_1\left(x^{-1}\right) = L_1\left(\frac{1}{x}\right) = -L_1 x. \tag{10.23}$$

You see that $\left(\frac{1}{2}\right)^n$ can be made as close to zero as desired by taking n sufficiently large. Also, from (10.22),

$$L_1\left(\frac{1}{2^n}\right) = nL_1\left(\frac{1}{2}\right) = -nL_1(2)$$

showing that $L_1(x)$ gets arbitrarily large in the negative direction provided that x is a sufficiently small positive number. Since L_1 is continuous, the intermediate value theorem may be used to fill in all the numbers in between. Thus the picture of the graph of L_1 looks like the following



where the graph approaches the y axis as x gets close to 0.

It only remains to verify the claim about raising x to a rational power. From (10.22) and (10.23), for any integer, n, positive or negative,

$$L_1(x^n) = nL_1(x).$$

Therefore, letting m, n be integers,

$$mL_1(x) = L_1(x^m) = L_1\left(\left(\sqrt[n]{x^m}\right)^n\right) = nL_1\left(\sqrt[n]{x^m}\right)$$

and so

$$\frac{m}{n}L_1(x) = L_1\left(\sqrt[n]{x^m}\right).$$

This proves the theorem.

Now Wild Assumption 7.3.1 on Page 158 can be fully justified.

Definition 10.8.2 *For $b > 0$ define*

$$\exp_b(x) \equiv L_1^{-1}(xL_1(b)). \tag{10.24}$$

Proposition 10.8.3 *The function just defined in (10.24) satisfies all conditions of Wild Assumption 7.3.1 on Page 158 and $L_1(b) = \ln b$ as defined on Page 158.*

Proof: First let $x = \frac{m}{n}$ where m and n are integers. To verify that $\exp_b(m/n) = \sqrt[n]{b^m}$,

$$\exp_b\left(\frac{m}{n}\right) \equiv L_1^{-1}\left(\frac{m}{n}L_1(b)\right) = L_1^{-1}\left(L_1\left(\sqrt[n]{b^m}\right)\right) = \sqrt[n]{b^m}$$

by (10.21).

That $\exp_b(x) > 0$ follows immediately from the definition of the inverse function. (L_1 is defined on positive real numbers and so L_1^{-1} has values in the positive real numbers.)

Consider the claim that if $h \neq 0$ and $b \neq 1$, then $\exp_b(h) \neq 1$. Suppose then that $\exp_b(h) = 1$. Then doing L_1 to both sides, $hL_1(b) = L_1(1) = 0$. Hence either $h = 0$ or $b = 1$ which are both excluded.

What of the laws of exponents for arbitrary values of x and y? As part of Wild Assumption 7.3.1, these were assumed to hold.

$$L_1\left(\exp_b\left(x+y\right)\right)=\left(x+y\right)L_1\left(b\right)$$

and

$$\begin{aligned}L_1\left(\exp_b\left(x\right)\exp_b\left(y\right)\right) &= L_1\left(\exp_b\left(x\right)\right)+L_1\left(\exp_b\left(y\right)\right)\\ &= xL_1\left(b\right)+yL_1\left(b\right)=\left(x+y\right)L_1\left(b\right).\end{aligned}$$

Since L_1 is one to one, this shows the first law of exponents holds,

$$\exp_b\left(x+y\right)=\exp_b\left(x\right)\exp_b\left(y\right).$$

From the definition, $\exp_b\left(1\right)=b$ and $\exp_b\left(0\right)=1$. Therefore,

$$1=\exp_b\left(1+\left(-1\right)\right)=\exp_b\left(-1\right)\exp_b\left(1\right)=\exp_b\left(-1\right)b$$

showing that $\exp_b\left(-1\right)=b^{-1}$. Now

$$L_1\left(\exp_{ab}\left(x\right)\right)=xL_1\left(ab\right)=xL_1\left(a\right)+xL_1\left(b\right)$$

while

$$\begin{aligned}L_1\left(\exp_a\left(x\right)\exp_b\left(x\right)\right) &= L_1\left(\exp_a\left(x\right)\right)+L_1\left(\exp_b\left(x\right)\right)\\ &= xL_1\left(a\right)+xL_1\left(b\right).\end{aligned}$$

Again, since L_1 is one to one, $\exp_{ab}\left(x\right)=\exp_a\left(x\right)\exp_b\left(x\right)$. Finally,

$$L_1\left(\exp_{\exp_b\left(x\right)}\left(y\right)\right)=yL_1\left(\exp_b\left(x\right)\right)=yxL_1\left(b\right)$$

while

$$L_1\left(\exp_b\left(xy\right)\right)=xyL_1\left(b\right)$$

and so since L_1 is one to one, $\exp_{\exp_b\left(x\right)}\left(y\right)=\exp_b\left(xy\right)$. This establishes all the laws of exponents for arbitrary real values of the exponent.

$\exp_b'$ exists by Theorem 8.1.6 on Page 169. Therefore, $\exp_b$ defined in (10.24) satisfies all the conditions of Wild Assumption 7.3.1.

It remains to consider the derivative of $\exp_b$ and verify $L_1\left(b\right)=\ln b$. First,

$$L_1\left(L_1^{-1}\left(x\right)\right)=x$$

and so

$$L_1'\left(L_1^{-1}\left(x\right)\right)\left(L_1^{-1}\right)'\left(x\right)=1.$$

By (10.19),

$$\frac{\left(L_1^{-1}\right)'\left(x\right)}{L_1^{-1}\left(x\right)}=1$$

showing that $\left(L_1^{-1}\right)'\left(x\right)=\left(L_1^{-1}\right)\left(x\right)$. Now by the chain rule,

$$\begin{aligned}\exp_b'\left(x\right) &= \left(L_1^{-1}\right)'\left(xL_1\left(b\right)\right)L_1\left(b\right)\\ &= L_1^{-1}\left(xL_1\left(b\right)\right)L_1\left(b\right)\\ &= \exp_b\left(x\right)L_1\left(b\right).\end{aligned}$$

From the definition of $\ln b$ on Page 158, it follows $\ln = L_1$.

10.9 Exercises

1. Show $\ln 2 \in [.5, 1]$.

2. Apply the trapezoid rule to estimate $\ln 2$ in the case where $h = 1/5$. Now use a calculator or table to find the exact value of $\ln 2$.

3. Suppose it is desired to find a function, $L : (0, \infty) \to \mathbb{R}$ which satisfies

$$L(xy) = Lx + Ly,\ L(1) = 0. \tag{10.25}$$

Show the only differentiable solutions to this equation are functions of the form $L_k(x) = \int_1^x \frac{k}{t}\, dt$. **Hint:** Fix $x > 0$ and differentiate both sides of the above equation with respect to y. Then let $y = 1$.

4. Recall that $\ln e = 1$. In fact, this was how e was defined. Show that

$$\lim_{y \to 0+} (1 + yx)^{1/y} = e^x.$$

Hint: Consider $\ln (1 + yx)^{1/y} = \frac{1}{y} \ln (1 + yx) = \frac{1}{y} \int_1^{1+yx} \frac{1}{t}\, dt$, use upper and lower sums and then the squeezing theorem to verify $\ln (1 + yx)^{1/y} \to x$. Recall that $x \to e^x$ is continuous.

5. Logarithms were invented before calculus, one of the inventors being Napier, a Scottish nobleman. His interest in logarithms was computational. Describe how one could use logarithms to find 7^{th} roots for example. Also describe how one could use logarithms to do computations involving large numbers. Describe how you could construct a table for $\log_{10} x$ for x various numbers. Next, how do you suppose Napier did it? Remember, he did not have calculus. **Hint:** $\log_b \left(35^{1/7}\right) = \frac{1}{7} \log_b (35)$. You can find the logarithm of the number you are after. If you had a table of logarithms to some base, you could then see what number this corresponded to. Napier essentially found a table of logarithms to the base .999999 or some such number like that. The reason for the strange choice is that successive integer powers of this number are quite close to each other. This is not true for a number like 10. Shortly after calculus was invented, the idea of considering the logarithms in terms of integrals became the best way to think of them.

10.10 Techniques Of Integration

The techniques for finding antiderivatives may be used to find integrals.

10.10.1 The Method Of Substitution

Recall

$$\int f(g(x))\, g'(x)\, dx = F(x) + C, \tag{10.26}$$

where $F'(y) = f(y)$.

How does this relate to finding definite integrals? This is based on the following formula in which all the functions are integrable and $F'(y) = f(y)$.

$$\int_a^b f(g(x))\, g'(x)\, dx = \int_{g(a)}^{g(b)} f(y)\, dy. \tag{10.27}$$

This formula follows from the observation that, by the fundamental theorem of calculus, both sides equal $F(g(b)) - F(g(a))$.

How can you remember this? The easiest way is to use the Leibniz notation. In (10.27) let $y = g(x)$. Then

$$\frac{dy}{dx} = g'(x)$$

and so formally $dy = g'(x)\,dx$. Then making the substitution

$$\int_a^b \overbrace{f(g(x))}^{f(y)} \overbrace{g'(x)\,dx}^{dy} = \int_?^? f(y)\,dy.$$

What should go in as the top and bottom limits of the integral? The important thing to remember is that **if you change the variable, you must change the limits!** When $x = a$, it follows that $y = g(a)$ to the bottom limit must equal $g(a)$. Similarly the top limit should be $g(b)$.

Example 10.10.1 *Find* $\int_1^2 x \sin\left(x^2\right)\,dx$

Let $u = x^2$ so $du = 2x\,dx$ and so $\frac{du}{2} = x\,dx$. Therefore, changing the variables gives

$$\int_1^2 x \sin\left(x^2\right)\,dx = \frac{1}{2}\int_1^4 \sin(u)\,du = -\frac{1}{2}\cos(4) + \frac{1}{2}\cos(1)$$

Sometimes people prefer not to worry about the limits. This is fine provided you don't write anything which is false. The above problem can be done in the following way.

$$\int x \sin\left(x^2\right)\,dx = -\frac{1}{2}\cos\left(x^2\right) + C$$

and so from an application of the fundamental theorem of calculus

$$\begin{aligned}\int_1^2 x \sin\left(x^2\right)\,dx &= -\frac{1}{2}\cos\left(x^2\right)|_1^2 \\ &= -\frac{1}{2}\cos(4) + \frac{1}{2}\cos(1).\end{aligned}$$

Example 10.10.2 *Find the area of the ellipse,*

$$\frac{(y-\beta)^2}{b^2} + \frac{(x-\alpha)^2}{a^2} = 1.$$

If you sketch the ellipse, you see that it suffices to find the area of the top right quarter for $y \geq \beta$ and $x \geq \alpha$ and multiply by 4 since the bottom half is just a reflection of the top half about the line, $y = \beta$ and the left top quarter is just the reflection of the top right quarter reflected about the line, $x = \alpha$. Thus the area of the ellipse is

$$4\int_\alpha^{\alpha+a} b\sqrt{1 - \frac{(x-\alpha)^2}{a^2}}\,dx$$

Change the variables, letting $u = \frac{x-\alpha}{a}$. Then $du = \frac{1}{a}\,dx$ and so upon changing the limits to correspond to the new variables, this equals

$$4ba\overbrace{\int_0^1 \sqrt{1-u^2}\,du}^{\pi/4} = 4 \times ab \times \frac{1}{4}\pi = \pi ab$$

because the integral in the above is just one quarter of the unit circle and so has area equal to $\pi/4$.

10.10.2 Integration By Parts

Recall the following proposition for finding antiderivative.

Proposition 10.10.3 *Let u and v be differentiable functions for which $\int u(x)\,v'(x)\,dx$ and $\int u'(x)\,v(x)\,dx$ are nonempty. Then*

$$uv - \int u'(x)\,v(x)\,dx = \int u(x)\,v'(x)\,dx. \tag{10.28}$$

In terms of integrals, this is stated in the following proposition.

Proposition 10.10.4 *Let u and v be differentiable functions on $[a,b]$ such that $uv', u'v \in R([a,b])$. Then*

$$\int_a^b u(x)\,v'(x)\,dx = uv(x)\,|_a^b - \int_a^b u'(x)\,v(x)\,dx \tag{10.29}$$

Proof: Use the product rule and properties of integrals to write

$$\begin{aligned}\int_a^b u(x)\,v'(x)\,dx &= \int_a^b (uv)'(x)\,dx - \int_a^b u'(x)\,v(x)\,dx \\ &= uv(x)\,|_a^b - \int_a^b u'(x)\,v(x)\,dx.\end{aligned}$$

This proves the proposition.

Example 10.10.5 *Find $\int_0^\pi x\sin(x)\,dx$*

Let $u(x) = x$ and $v'(x) = \sin(x)$. Then applying (10.28),

$$\begin{aligned}\int_0^\pi x\sin(x)\,dx &= (-\cos(x))\,x|_0^\pi - \int_0^\pi (-\cos(x))\,dx \\ &= -\pi\cos(\pi) = \pi.\end{aligned}$$

Example 10.10.6 *Find $\int_0^1 xe^{2x}\,dx$*

Let $u(x) = x$ and $v'(x) = e^{2x}$. Then from (10.29)

$$\begin{aligned}\int_0^1 xe^{2x}\,dx &= \frac{e^{2x}}{2}x|_0^1 - \int_0^1 \frac{e^{2x}}{2}\,dx \\ &= \frac{e^{2x}}{2}x|_0^1 - \frac{e^{2x}}{4}|_0^1 \\ &= \frac{e^2}{2} - \left(\frac{e^2}{4} - \frac{1}{4}\right) \\ &= \frac{e^2}{4} + \frac{1}{4}\end{aligned}$$

10.11 Exercises

1. Find the integrals.

   (a) $\int_0^4 xe^{-3x}\,dx$

(b) $\int_2^3 \frac{1}{x(\ln(|x|))^2}\,dx$

(c) $\int_0^1 x\sqrt{2-x}\,dx$

(d) $\int_2^3 (\ln|x|)^2\,dx$ **Hint:** Let $u(x) = (\ln|x|)^2$ and $v'(x) = 1$.

(e) $\int_0^\pi x^3\cos\left(x^2\right)\,dx$

2. Find $\int_1^2 x\ln\left(x^2\right)\,dx$

3. Find $\int_0^1 e^x\sin(x)\,dx$

4. Find $\int_0^1 2^x\cos(x)\,dx$

5. Find $\int_0^2 x^3\cos(x)\,dx$

6. Find the integrals.

(a) $\int_5^6 \frac{x}{\sqrt{2x-3}}\,dx$

(b) $\int_2^4 x\left(3x^2+6\right)^5\,dx$

(c) $\int_0^\pi x\sin\left(x^2\right)\,dx$

(d) $\int_0^{\pi/4}\sin^3(2x)\cos(2x)$

(e) $\int_0^7 \frac{1}{\sqrt{1+4x^2}}\,dx$ **Hint:** Remember the $\sinh^{-1}$ function and its derivative.

7. Find the integrals.

(a) $\int_0^{\pi/9}\sec(3x)\,dx$

(b) $\int_0^{\pi/9}\sec^2(3x)\tan(3x)\,dx$

(c) $\int_0^5 \frac{1}{3+5x^2}\,dx$

(d) $\int_0^1 \frac{1}{\sqrt{5-4x^2}}\,dx$

(e) $\int_2^6 \frac{3}{x\sqrt{4x^2-5}}\,dx$

8. Find the integrals.

(a) $\int_0^3 x\cosh\left(x^2+1\right)\,dx$

(b) $\int_0^2 x^3 5^{x^4}\,dx$

(c) $\int_{-\pi}^\pi \sin(x)\,7^{\cos(x)}\,dx$

(d) $\int_0^\pi x\sin\left(x^2\right)\,dx$

(e) $\int_1^2 x^5\sqrt{2x^2+1}\,dx$ **Hint:** Let $u = 2x^2+1$.

9. Find $\int_0^{\pi/4}\sin^2(x)\,dx$. **Hint:** Derive and use $\sin^2(x) = \frac{1-\cos(2x)}{2}$.

10. Find the area between the graphs of $y = \sin(2x)$ and $y = \cos(2x)$ for $x \in [0, 2\pi]$.

11. Find the following integrals.

(a) $\int_1^\pi x^2\sin\left(x^3\right)\,dx$

(b) $\int_1^6 \frac{x}{1+x^2}\,dx$

(c) $\int_0^{.5} \frac{1}{1+4x^2}\,dx$

(d) $\int_1^4 x^2\sqrt{1+x}\,dx$

12. The most important of all differential equations is the first order linear equation, $y' + p(t)\,y = f(t)$. Show the solution to the initial value problem consisting of this equation and the initial condition, $y(a) = y_a$ is

$$y(t) = e^{-P(t)}y_a + e^{-P(t)}\int_a^t e^{P(s)}f(s)\,ds,$$

where $P(t) = \int_a^t p(s)\,ds$. Give conditions under which everything is correct. **Hint:** You use the integrating factor approach. Multiply both sides by $e^{P(t)}$, verify the left side equals

$$\frac{d}{dt}\left(e^{P(t)}y(t)\right),$$

and then take the integral, $\int_a^t$ of both sides.

13. Suppose $x_0 \in (a,b)$ and that f is a function which has $n+1$ continuous derivatives on this interval. Consider the following.

$$\begin{aligned} f(x) &= f(x_0) + \int_{x_0}^x f'(t)\,dt \\ &= f(x_0) + (t-x)\,f'(t)\,|_{x_0}^x + \int_{x_0}^x (x-t)\,f''(t)\,dt \\ &= f(x_0) + f'(x_0)(x-x_0) + \int_{x_0}^x (x-t)\,f''(t)\,dt. \end{aligned}$$

Explain the above steps and continue the process to eventually obtain Taylor's formula,

$$f(x) = f(x_0) + \sum_{k=1}^n \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k + \frac{1}{n!}\int_{x_0}^x (x-t)^n f^{(n+1)}(t)\,dt$$

where $n! \equiv n(n-1)\cdots 3\cdot 2\cdot 1$ if $n \geq 1$ and $0! \equiv 1$.

14. In the above Taylor's formula, use Problem 12 on Page 279 to obtain the existence of some z between x_0 and x such that

$$f(x) = f(x_0) + \sum_{k=1}^n \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k + \frac{f^{(n+1)}(z)}{(n+1)!}(x-x_0)^{n+1}.$$

Hint: You might consider two cases, the case when $x > x_0$ and the case when $x < x_0$.

15. There is a general procedure for constructing these methods of approximate integration like the trapezoid rule and Simpson's rule. Consider $[0,1]$ and divide this interval into n pieces using a uniform partition, $\{x_0, \cdots, x_n\}$ where $x_i - x_{i-1} = 1/n$ for each i. The approximate integration scheme for a function, f, will be of the form

$$\left(\frac{1}{n}\right)\sum_{i=0}^n c_i f_i \approx \int_0^1 f(x)\,dx$$

where $f_i = f(x_i)$ and the constants, c_i are chosen in such a way that the above sum gives the exact answer for $\int_0^1 f(x)\,dx$ where $f(x) = 1, x, x^2, \cdots, x^n$. When this has been done, change variables to write

$$\begin{aligned}\int_a^b f(y)\,dy &= (b-a)\int_0^1 f(a+(b-a)x)\,dx \\ &\approx \frac{b-a}{n}\sum_{i=1}^n c_i f\left(a+(b-a)\left(\frac{i}{n}\right)\right) \\ &= \frac{b-a}{n}\sum_{i=1}^n c_i f_i\end{aligned}$$

where $f_i = f\left(a+(b-a)\left(\frac{i}{n}\right)\right)$. Consider the case where $n=1$. It is necessary to find constants c_0 and c_1 such that

$$\begin{aligned}c_0 + c_1 = 1 = \textstyle\int_0^1 1\,dx \\ 0c_0 + c_1 = 1/2 = \textstyle\int_0^1 x\,dx.\end{aligned}$$

Show that $c_0 = c_1 = 1/2$, and that this yields the trapezoid rule. Next take $n=2$ and show the above procedure yields Simpson's rule. Show also that if this integration scheme is applied to any polynomial of degree 3 the result will be exact. That is,

$$\frac{1}{2}\left(\frac{1}{3}f_0 + \frac{4}{3}f_1 + \frac{1}{3}f_2\right) = \int_0^1 f(x)\,dx$$

whenever $f(x)$ is a polynomial of degree three. Show that if f_i are the values of f at $a, \frac{a+b}{2}$, and b with $f_1 = f\left(\frac{a+b}{2}\right)$, it follows that the above formula gives $\int_a^b f(x)\,dx$ exactly whenever f is a polynomial of degree three. Obtain an integration scheme for $n=3$.

16. Let f have four continuous derivatives on $[x_{i-1}, x_{i+1}]$ where $x_{i+1} = x_{i-1} + 2h$ and $x_i = x_{i-1} + h$. Show using Problem 14, there exists a polynomial of degree three, $p_3(x)$, such that

$$f(x) = p_3(x) + \frac{1}{4!}f^{(4)}(\xi)(x-x_i)^4$$

Now use Problem 15 and Problem 6 on Page 278 to conclude

$$\left|\int_{x_{i-1}}^{x_{i+1}} f(x)\,dx - \left(\frac{hf_{i-1}}{3} + \frac{hf_i 4}{3} + \frac{hf_{i+1}}{3}\right)\right| < \frac{M}{4!}\frac{2h^5}{5},$$

where M satisfies, $M \geq \max\left\{\left|f^{(4)}(t)\right| : t \in [x_{i-1}, x_i]\right\}$. Now let $S(a,b,f,2m)$ denote the approximation to $\int_a^b f(x)\,dx$ obtained from Simpson's rule using $2m$ equally spaced points. Show

$$\left|\int_a^b f(x)\,dx - S(a,b,f,2m)\right| < \frac{M}{1920}(b-a)^5\frac{1}{m^4}$$

where $M \geq \max\left\{\left|f^{(4)}(t)\right| : t \in [a,b]\right\}$. Better estimates are available in numerical analysis books. However, these also have the error in the form $C\left(1/m^4\right)$.

17. A **regular Sturm Liouville problem** involves the differential equation, for an unknown function of x which is denoted here by y,

$$(p(x)y')' + (\lambda q(x) + r(x))y = 0,\ x \in [a,b]$$

and it is assumed that $p(t), q(t) > 0$ for any t along with boundary conditions,

$$\begin{aligned} C_1 y(a) + C_2 y'(a) &= 0 \\ C_3 y(b) + C_4 y'(b) &= 0 \end{aligned}$$

where

$$C_1^2 + C_2^2 > 0, \text{ and } C_3^2 + C_4^2 > 0.$$

There is an imense theory connected to these important problems. The constant, λ is called an eigenvalue. Show that if y is a solution to the above problem corresponding to$\lambda = \lambda_1$ and if z is a solution corresponding to $\lambda = \lambda_2 \neq \lambda_1$, then

$$\int_a^b q(x)y(x)z(x)\,dx = 0. \tag{10.30}$$

Hint: Do something like this:

$$\begin{aligned} (p(x)y')'z + (\lambda_1 q(x) + r(x))yz &= 0, \\ (p(x)z')'y + (\lambda_2 q(x) + r(x))zy &= 0. \end{aligned}$$

Now subtract and either use integration by parts or show

$$(p(x)y')'z - (p(x)z')'y = ((p(x)y')z - (p(x)z')y)'$$

and then integrate. Use the boundary conditions to show that $y'(a)z(a) - z'(a)y(a) = 0$ and $y'(b)z(b) - z'(b)y(b) = 0$. The formula, (10.30) is called an orthogonality relation and it makes possible an expansion in terms of certain functions called eigenfunctions.

18. Letting $[a,b] = [-\pi,\pi]$, consider an example of a regular Sturm Liouville problem which is of the form

$$y'' + \lambda y = 0, y(-\pi) = 0, y(\pi) = 0.$$

Show that if $\lambda = n^2$ and $y_n(x) = \sin(nx)$ for n a positive integer, then y_n is a solution to this regular Sturm Liouville problem. In this case, $q(x) = 1$ and so from Problem 17, it must be the case that

$$\int_{-\pi}^{\pi} \sin(nx)\sin(mx)\,dx = 0$$

if $n \neq m$. Show directly using integration by parts that the above equation is true.

10.12 Improper Integrals

The integral is only defined for certain bounded functions which are defined on closed and bounded intervals. Nevertheless people do consider things like the following: $\int_0^\infty f(t)\,dt$. Whenever things like this occur they require a special definition. They are called **improper integrals**. In this section a few types of improper integrals will be discussed.

Definition 10.12.1 *The symbol* $\int_a^\infty f(t)\,dt$ *is defined to equal*

$$\lim_{R\to\infty}\int_a^R f(t)\,dt$$

whenever this limit exists. If $\lim_{x\to a+} f(t) = \pm\infty$ *but* f *is integrable on* $[a+\delta, b]$ *for all small* δ*, then*

$$\int_a^b f(t)\,dt \equiv \lim_{\delta\to 0+}\int_{a+\delta}^b f(t)\,dt$$

whenever this limit exists. Similarly, if $\lim_{x\to b-} f(t) = \pm\infty$ *but* f *is integrable on* $[a, b-\delta]$ *for all small* δ*, then*

$$\int_a^b f(t)\,dt \equiv \lim_{\delta\to 0+}\int_a^{b-\delta} f(t)\,dt$$

whenever the limit exists. Finally, if $\lim_{x\to a+} f(t) = \pm\infty$*, then*

$$\int_a^\infty f(t)\,dt \equiv \lim_{R\to\infty}\int_a^R f(t)\,dt$$

where the improper integral, $\int_a^R f(t)\,dt$ *is defined above as* $\lim_{\delta\to 0+}\int_{a+\delta}^R f(t)\,dt$.

You can probably construct other examples of improper integrals such as integrals of the form $\int_{-\infty}^a f(t)\,dt$. The definitions are analogous to the above.

Example 10.12.2 *Find* $\int_0^\infty e^{-t}\,dt$.

From the definition, this equals $\lim_{R\to\infty}\int_0^R e^{-t}\,dt = \lim_{R\to\infty}\left(1-e^{-R}\right) = 1$.

Example 10.12.3 *Find* $\int_0^1 \frac{1}{\sqrt{x}}\,dx$.

From the definition this equals $\lim_{\delta\to 0+}\int_\delta^1 x^{-1/2}\,dx = \lim_{\delta\to 0+}\left(2-2\sqrt{\delta}\right) = 2$.

Sometimes you can argue the improper integral exists even though you can't find it. The following theorem is about this question of existence.

Theorem 10.12.4 *Suppose* $f(t)\geq 0$ *for all* $t\in[a,\infty)$ *and that* $f\in R([a+\delta, R])$ *whenever* $\delta>0$ *is small enough, for every* $R>a$*. Suppose also there exists a number,* M *such that for all* $R>a$*, the integral* $\int_a^R f(t)\,dt$ *exists and*

$$\int_a^R f(t)\,dt \leq M. \tag{10.31}$$

Then $\int_a^\infty f(t)\,dt$ *exists.*

If $f(t)\geq 0$ *for all* $t\in(a,b]$ *and there exists* M *such that*

$$\int_{a+\delta}^b f(t)\,dt \leq M \tag{10.32}$$

for all $\delta>0$*, then* $\int_a^b f(t)\,dt$ *exists.*

If $f(t)\geq 0$ *for all* $t\in[a,b)$ *and there exists* M *such that*

$$\int_a^{b-\delta} f(t)\,dt \leq M \tag{10.33}$$

for all $\delta>0$*, then* $\int_a^b f(t)\,dt$ *exists.*

Proof: Suppose (10.31). Then $I \equiv \sup\left\{\int_a^R f(t)\,dt : R > a\right\} \leq M$. It follows that if $\varepsilon > 0$ is given, there exists R_0 such that $\int_a^{R_0} f(t)\,dt \in (I-\varepsilon, I]$. Then, since $f(t) \geq 0$, it follows that for $R \geq R_0$, and small $\delta > 0$,

$$\int_{a+\delta}^{R} f(t)\,dt = \int_{a+\delta}^{R_0} f(t)\,dt + \int_{R_0}^{R} f(t)\,dt.$$

Letting $\delta \to 0+$,

$$\int_{a}^{R} f(t)\,dt = \int_{a}^{R_0} f(t)\,dt + \int_{R_0}^{R} f(t)\,dt \geq \int_{a}^{R_0} f(t)\,dt.$$

Therefore, whenever $R > R_0$, $\left|\int_a^R f(t)\,dt - I\right| < \varepsilon$. Since ε is arbitrary, the conditions for

$$I = \lim_{R\to\infty} \int_a^R f(t)\,dt$$

are satisfied and so $I = \int_a^\infty f(t)\,dt$.

Now suppose (10.32). Then $I \equiv \sup\left\{\int_{a+\delta}^b f(t)\,dt : \delta > 0\right\} \leq M$. It follows that if $\varepsilon > 0$ is given, there exists $\delta_0 > 0$ such that

$$\int_{a+\delta_0}^{b} f(t)\,dt \in (I-\varepsilon, I].$$

Therefore, if $\delta < \delta_0$,

$$I - \varepsilon < \int_{a+\delta_0}^{b} f(t)\,dt \leq \int_{a+\delta}^{b} f(t)\,dt \leq I$$

showing that for such δ, $\left|\int_{a+\delta}^b f(t)\,dt - I\right| < \varepsilon$. This is what is meant by the expression

$$\lim_{\delta\to 0+} \int_{a+\delta}^{b} f(t)\,dt = I$$

and so $I = \int_a^b f(t)\,dt$.

The last case is entirely similar to this one. This proves the theorem.

Example 10.12.5 *Does* $\int_0^1 \frac{1}{\sqrt{\sin x}}\,dx$ *exist?*

I don't know how to find an antiderivative for this function but the question of existence can still be resolved. Since $\lim_{x\to 0+} \frac{x}{\sin x} = 1$, it follows that for x small enough, $\frac{x}{\sin x} < \frac{3}{2}$, say for $x < \delta_1$. Then for such x, it follows

$$\frac{2}{3}x < \sin x$$

and so if $\delta < \delta_1$,

$$\int_\delta^1 \frac{1}{\sqrt{\sin x}}\,dx \leq \int_\delta^{\delta_1} \sqrt{\frac{3}{2}}\frac{1}{\sqrt{x}}\,dx + \int_{\delta_1}^1 \frac{1}{\sqrt{\sin\delta_1}}\,dx.$$

Now using the argument of Example 10.12.3, the first integral in the above is bounded above by $\left(\sqrt{\delta_1} - \sqrt{\delta}\right)\sqrt{6}$. The second integral equals $\frac{1-\delta_1}{\sqrt{\sin\delta_1}}$. Therefore, the improper integral exists because the conditions of Theorem 10.12.4 with $M = \frac{1-\delta_1}{\sqrt{\sin\delta_1}} + \left(\sqrt{\delta_1} - \sqrt{\delta}\right)\sqrt{6}$.

Example 10.12.6 *The gamma function is defined by* $\Gamma(\alpha) \equiv \int_0^\infty e^{-t}t^{\alpha-1}\,dt$ *whenever* $\alpha > 0$. *Does the improper integral exist?*

You should supply the details to the following estimate in which δ is a small positive number less than 1 and R is a large positive number.

$$\begin{aligned}\int_\delta^R e^{-t}t^{\alpha-1}\,dt &\leq \int_\delta^k e^{-t}t^{\alpha-1}\,dt + \int_k^R e^{-t}t^{\alpha-1}\,dt \\ &\leq \int_0^k t^{\alpha-1}\,dt + \int_k^\infty e^{-t/2}\,dt.\end{aligned}$$

Here k is chosen such that if $t \geq k$,

$$e^{-t}t^{\alpha-1} < e^{-t/2}.$$

Such a k exists because

$$\lim_{t\to\infty} \frac{e^{-t}t^{\alpha-1}}{e^{-t/2}} = 0.$$

Therefore, let $M \equiv \int_0^k t^{\alpha-1}\,dt + \int_k^\infty e^{-t/2}\,dt$ and this shows from the above theorem that

$$\int_0^R e^{-t}t^{\alpha-1}\,dt \leq M$$

for all large R and so $\int_0^\infty e^{-t}t^{\alpha-1}\,dt$ exists.

Sometimes the existence of the improper integral is a little more subtle. This is the case when functions are not all the same sign for example.

Example 10.12.7 *Does* $\int_0^\infty \frac{\sin x}{x}\,dx$ *exist?*

You should verify $\int_0^1 \frac{\sin x}{x}\,dx$ exists and that

$$\begin{aligned}\int_0^R \frac{\sin x}{x}\,dx &= \int_0^1 \frac{\sin x}{x}\,dx + \int_1^R \frac{\sin x}{x}\,dx \\ &= \int_0^1 \frac{\sin x}{x}\,dx + \cos 1 - \frac{\cos R}{R} - \int_1^R \frac{\cos x}{x^2}\,dx.\end{aligned}$$

Thus the improper integral exists if it can be shown that $\int_1^\infty \frac{\cos x}{x^2}\,dx$ exists. However,

$$\int_1^R \frac{\cos x}{x^2}\,dx = \int_1^R \frac{\cos x + |\cos x|}{x^2}\,dx - \int_1^R \frac{(|\cos x| - \cos x)}{x^2}\,dx \tag{10.34}$$

and

$$\int_1^R \frac{\cos x + |\cos x|}{x^2}\,dx \leq \int_1^R \frac{2}{x^2}\,dx \leq \int_0^\infty \frac{2}{x^2}\,dx < \infty$$

$$\int_1^R \frac{(|\cos x| - \cos x)}{x^2}\,dx \leq \int_1^R \frac{2}{x^2}\,dx \leq \int_0^\infty \frac{2}{x^2}\,dx < \infty.$$

Since both integrands are positive, Theorem 10.12.4 applies and the limits

$$\lim_{R\to\infty} \int_1^R \frac{(|\cos x| - \cos x)}{x^2}\,dx, \lim_{R\to\infty} \int_1^R \frac{\cos x + |\cos x|}{x^2}\,dx$$

both exist and so from (10.34) $\lim_{R\to\infty}\int_1^R \frac{\cos x}{x^2}\,dx$ also exists and so $\int_0^\infty \frac{\sin x}{x}\,dx$ exists.

This is an important example. There are at least two ways to show that $\int_0^\infty \frac{\sin x}{x}\,dx = \frac{1}{2}\pi$. However, they involve techniques which will not be discussed in this book. It is a standard problem in the subject of complex analysis. The above argument is a special case of the following corollary to Theorem 10.12.4.

Definition 10.12.8 *Let f be a real valued function. Then $f^+(t) \equiv \frac{|f(t)|+f(t)}{2}$ and $f^-(t) \equiv \frac{|f(t)|-f(t)}{2}$. Thus $|f(t)| = f^+(t) + f^-(t)$ and $f(t) = f^+(t) - f^-(t)$ while both f^+ and f^- are nonnegative functions.*

Corollary 10.12.9 *Suppose f is a real valued function, Riemann integrable on every finite interval, and the conditions of Theorem 10.12.4 hold for both f^+ and f^-. Then $\int_0^\infty f(t)\,dt$ exists.*

Corollary 10.12.10 *Suppose f is a real valued function, Riemann integrable on every finite interval, and $\int_0^\infty |f(x)|\,dx$ exists. Then $\int_0^\infty f(x)\,dx$ exists.*

Proof: $0 \le f^+(x) \le |f(x)|$, and $0 \le f^-(x) \le |f(x)|$ and so for every $R > 0$,

$$\int_0^R f^+(x)\,dx \le \int_0^\infty |f(x)|\,dx$$

and

$$\int_0^R f^-(x)\,dx \le \int_0^\infty |f(x)|\,dx$$

and so it follows from Theorem 10.12.4 that $\int_0^\infty f^+(x)\,dx$ and $\int_0^\infty f^-(x)\,dx$ both exist. Therefore,

$$\int_0^\infty f(x)\,dx \equiv \lim_{R\to\infty}\left(\int_0^R f^+(x)\,dx - \int_0^R f^-(x)\,dx\right)$$

also exists.

Example 10.12.11 *Does $\int_0^\infty \cos\left(x^2\right)\,dx$ exist?*

This is called a **Fresnel integral** and it has also been evaluated exactly using techniques from complex analysis. In fact $\int_0^\infty \cos\left(x^2\right)\,dx = \frac{1}{4}\sqrt{2}\sqrt{\pi}$. The verification that this integral exists is left to you. First change the variable letting $x^2 = u$ and then integrate by parts. You will eventually get an integral of the form $\int_0^\infty \frac{\sin u}{u^{3/2}}\,du$. Now consider $\int_\delta^1 \frac{\sin u}{u^{3/2}}du$ where δ is a small positive number. On $[\delta, 1]$, $\sin u$ is nonnegative. You also know that $\sin u \le u$. Therefore, $\frac{\sin u}{u^{3/2}} \le \frac{1}{u^{1/2}}$. Thus

$$\int_\delta^1 \frac{\sin u}{u^{3/2}}du \le \int_\delta^1 \frac{1}{u^{1/2}}du = 2 - 2\delta^{1/2} \le 2.$$

Therefore, $\int_0^1 \frac{\sin u}{u^{3/2}}du$ exists. Similarly, $\int_0^R \frac{\sin u}{u^{3/2}}du$ exists. Now

$$\int_0^R \frac{|\sin u|}{u^{3/2}} \le \int_0^1 \frac{|\sin u|}{u^{3/2}}du + \int_1^R \frac{1}{u^{3/2}}du \le 2 + 2$$

and so $\int_0^\infty \frac{|\sin u|}{u^{3/2}}du$ exists which shows by Corollary 10.12.10 that $\int_0^\infty \frac{\sin u}{u^{3/2}}du$ exists also. I have been a little sketchy on the details. You finish them.

10.13 Exercises

1. Verify all the details in Example 10.12.6.
2. Verify all the details of Example 10.12.7.
3. Verify all the details of Example 10.12.11.
4. Find the values of p for which $\int_1^\infty \frac{1}{t^p}dt$ exists and compute the integral when it does exist.
5. Find the values of p for which $\int_2^\infty \frac{1}{t(\ln t)^p}dt$ exists and compute the integral when it does exist.
6. Determine whether $\int_1^\infty \frac{\sin t}{\sqrt{t}}dt$ exists.
7. Determine whether $\int_3^\infty \frac{\sin t}{\ln t}dt$ exists. **Hint:** You might try integrating by parts.
8. Determine whether $\int_0^1 \frac{1}{\sqrt{x}+\sin x}dx$ exists.
9. Determine whether $\int_1^\infty \frac{1}{\sqrt{x+x^5}}dx$ exists.
10. Determine whether $\int_0^1 \frac{\sin t}{t}dt$ exists.
11. Determine whether $\int_0^1 \frac{\cot t}{t}dt$ exists.
12. Determine whether $\int_0^1 \frac{1}{1-x^3}dt$ exists.
13. Determine whether $\int_0^1 \frac{1}{1-\sqrt{x}}dx$ exists.
14. Determine whether $\int_0^1 \frac{1}{\sqrt{1-x}}dx$ exists.
15. Find $\int_0^1 \frac{1}{\sqrt{1-x}}dx$ if it exists.
16. Find $\int_0^1 \frac{1}{\sqrt[3]{1-x}}dx$ if it exists.
17. Find $\int_0^{\pi/2} \frac{\cos x}{\sqrt{1-\sin x}}dx$ if it exists.
18. Find $\int_0^\infty \frac{1}{4x^2+9}dx$ if it exists.
19. Define and find $\int_{-\infty}^0 e^x dx$. Note the lower limit of integration is $-\infty$.
20. Find $\int_0^\infty e^{-x}dx$ and then define and find the volume obtained by revolving the graph of $y = e^{-x}$ about the x axis. Define the surface area of the shape obtained by revolving about the x axis and determine whether it is finite.
21. When $\int_0^\infty f(x)\,dx$ and $\int_{-\infty}^0 f(x)\,dx$ both exist, it follows $\int_{-\infty}^\infty f(x)\,dx$ also exists and equals the sum of the two first integrals,

$$\int_{-\infty}^\infty f(x)\,dx = \int_{-\infty}^0 f(x)\,dx + \int_0^\infty f(x)\,dx.$$

The normal distribution function is $\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. In this formula, μ is the mean and σ is a positive number called the standard deviation. In statistics, there are

things called random variables. These are really just a kind of function and one of these random variables is said to be normally distributed if the probability that it has a value between a and b is given by the integral $\int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$. You may be wondering why $\sqrt{2\pi}$ occurs in this. It is because it is what is needed to have $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$. Later in the book you will learn how to show this. It is in Problem 24 on Page 736 and depends on changing variables in multiple integrals. For now, show $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$ exists. The importance of this distribution function cannot be over stated.

22. Show $\Gamma(1) = 1 = \Gamma(2)$. Next show $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ and prove that for n a non-negative integer, $\Gamma(n+1) = n!$.

23. It can be shown that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. Using this and Problem 22, find $\Gamma\left(\frac{5}{2}\right)$. What is the advantage of the gamma function over the notion of factorials? **Hint:** For what values of x is $\Gamma(x)$ defined?

24. Prove $\int_0^{\infty} \sin\left(x^2\right) dx$ exists. This is also called a Fresnel integral.

25. For $\alpha > 0$ find $\int_0^1 t^{\alpha} \ln t \, dt$ if it exists and if it does not exist, explain why.

26. Prove that for every $\alpha > 0$, $\int_0^1 t^{\alpha-1} dt$ exists and find the answer.

27. Prove that for every $\alpha > 0$, $\int_0^1 (\tan t)^{\alpha-1} dt$ exists.

28. Prove that for every $\alpha > 0$, $\int_0^1 (\sin t)^{\alpha-1} dt$ exists.

29. Recall the area of the surface obtained by revolving the graph of $y = f(x)$ about the x axis for $x \in [a,b]$ for f a positive continuous function having continuous derivative is given by
$$2\pi \int_a^b f(x) \sqrt{1 + (f'(x))^2} dx.$$
Also the volume of the solid obtained in the same way is given by the integral,
$$\pi \int_a^b (f(x))^2 dx.$$
It seems reasonable to define the surface area and volume for $x \in [a,\infty)$ in terms of an improper integral. Try it on the function, $f(x) = 1/x$ for $x \geq 1$. Show the resulting solid has finite volume but infinite surface area. Thus you could fill it but you couldn't paint it. Sometimes people call this Gabriel's horn.

30. Show $\int_0^{\infty} \frac{2x}{1+x^2} dx$ does not exist but that $\lim_{R\to\infty} \int_{-R}^{R} \frac{2x}{1+x^2} dx = 0$. This last limit is called the Cauchy principle value integral. It is not a very respectable thing. Later you might study a subject called complex analysis in which techniques for finding hard integrals are developed. These methods often give the Cauchy principle value. Try to show the following: For every number, A, there exist sequences $a_n, b_n \to \infty$ such that
$$\lim_{n\to\infty} \int_{-a_n}^{b_n} \frac{2x}{1+x^2} dx = A.$$
This is true and shows why such principle value integrals are somewhat disreputable.

31. Suppose f is a continuous function which is bounded and defined on $\mathbb{R}$. Show

$$\int_{-\infty}^{\infty} \frac{\varepsilon}{\pi\left(\varepsilon^2 + (x - x_1)^2\right)} dx = 1.$$

Next show that

$$\lim_{\varepsilon \to 0+} \int_{-\infty}^{\infty} \frac{\varepsilon f(x)}{\pi\left(\varepsilon^2 + (x - x_1)^2\right)} dx = f(x_1).$$

Infinite Series

11.0.1 Outcomes

1. Understand the use of Taylor polynomials for approximating a given function and understand a proof of some form for the remainder.
2. Recall and understand the meaning of a convergent series.
3. Recall, use and understand the various tests for determining convergence of series.
4. Recall and explain the difference between conditional and absolute convergence.
5. Be able to justify the interchange in order of summation in a double sum.
6. Understand the use of a Taylor series as the definition of a function.
7. Understand and use correctly various methods for determining the interval of convergence and radius of convergence of a Taylor series.
8. Recall and understand the various operations which can be used on Taylor series and be able to justify their use.

11.1 Approximation By Taylor Polynomials

By now, you have noticed there are two sorts of functions, those which come from a formula like $f(x) = x^2 + 2$ which are easy to evaluate by following a simple procedure, and those which come as short words; things like $\ln(x)$ or $\sin(x)$. This latter type of function is not so easy to evaluate. For example, what is $\sin 2$? Can you get it by doing a simple sequence of operations like you can with $f(x) = x^2 + 2$? How can you find $\sin 2$? It turns out there are many ways to do so. In this section, the method of Taylor polynomials is discussed. The following theorem is called Taylor's theorem. Before presenting it, recall the meaning of $n!$ for n a positive integer. Define $0! \equiv 1 = 1!$ and $(n+1)! \equiv (n+1)\, n!$ so that $n! = n(n-1)\cdots 1$. In particular, $2! = 2, 3! = 3 \times 2! = 6, 4! = 4 \times 3! = 24$, etc.

Theorem 11.1.1 *Suppose f has $n+1$ derivatives on an interval, (a,b) and let $c \in (a,b)$. Then if $x \in (a,b)$, there exists ξ between c and x such that*

$$f(x) = f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}.$$

(In this formula, the symbol $\sum_{k=1}^{0} a_k$ will denote the number 0.)

Proof: If $n = 0$ then the theorem is true because it is just the mean value theorem. Suppose the theorem is true for $n-1, n \geq 1$. It can be assumed $x \neq c$ because if $x = c$ there is nothing to show. Then there exists K such that

$$f(x) - \left(f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k + K(x-c)^{n+1} \right) = 0$$

In fact,

$$K = \frac{-f(x) + \left(f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k \right)}{(x-c)^{n+1}}.$$

Now define $F(t)$ for t in the closed interval determined by x and c by

$$F(t) \equiv f(x) - \left(f(t) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!} (x-t)^k + K(x-t)^{n+1} \right).$$

Therefore, $F(c) = 0$ and also $F(x) = 0$. By the mean value theorem or Rolle's theorem, there exists t_1 between x and c such that $F'(t_1) = 0$. Therefore,

$$\begin{aligned} 0 &= f'(t_1) - \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!} k (x-t_1)^{k-1} - K(n+1)(x-t_1)^n \\ &= f'(t_1) - \left(f'(c) + \sum_{k=1}^{n-1} \frac{f^{(k+1)}(c)}{k!} (x-t_1)^k \right) - K(n+1)(x-t_1)^n \\ &= f'(t_1) - \left(f'(c) + \sum_{k=1}^{n-1} \frac{f'^{(k)}(c)}{k!} (x-t_1)^k \right) - K(n+1)(x-t_1)^n \end{aligned}$$

By induction applied to f', there exists ξ between x and t_1 such that the above simplifies to

$$\begin{aligned} 0 &= \frac{f'^{(n)}(\xi)(x-t_1)^n}{n!} - K(n+1)(x-t_1)^n \\ &= \frac{f^{(n+1)}(\xi)(x-t_1)^n}{n!} - K(n+1)(x-t_1)^n \end{aligned}$$

therefore,

$$K = \frac{f^{(n+1)}(\xi)}{(n+1)n!} = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

and the formula is true for n. This proves the theorem.

The term $\frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}$, is called the remainder and this particular form of the remainder is called the Lagrange form of the remainder.

Example 11.1.2 *Approximate* $\sin x$ *for* x *in some open interval containing* 0.

Use Taylor's formula just presented and let $c = 0$. Then for $f(x) = \sin x$,

$$f'(x) = \cos x,\ f''(x) = -\sin x,\ f'''(x) = -\cos x,$$

etc. Therefore, $f(0) = 0, f'(0) = 1, f''(0) = 0, f'''(0) = -1$, etc. Thus the Taylor polynomial for $\sin x$ is of the form

$$x - \frac{x^3}{3!} + \cdots \pm \frac{x^{2n+1}}{(2n+1)!}$$

while the remainder is of the form

$$\frac{f^{(2n+2)}(\xi)}{(2n+2)!}$$

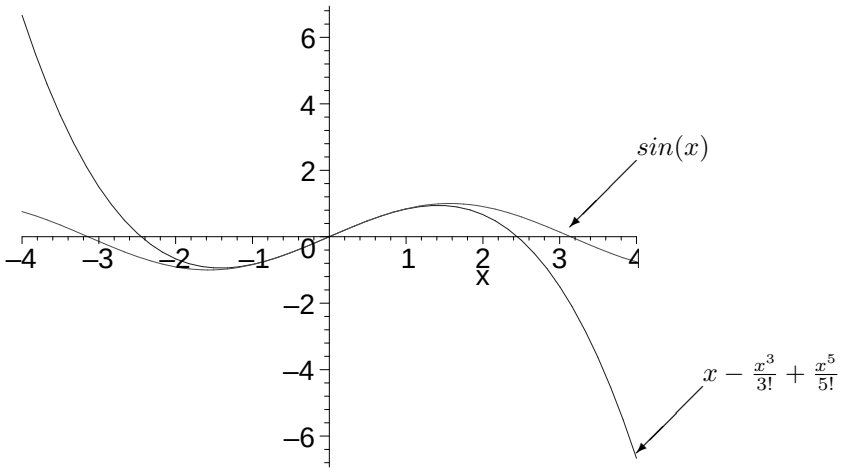
for some ξ between 0 and x. For $n=2$ in the above, the resulting polynomial is

$$x-\frac{x^3}{3!}+\frac{x^5}{5!}$$

and the error between this polynomial and $\sin x$ must be measured by the remainder term. Therefore,

$$\left|\sin x-\left(x-\frac{x^3}{3!}+\frac{x^5}{5!}\right)\right|\leq\left|\frac{f^{(6)}(\xi)\,x^6}{6!}\right|\leq\frac{x^6}{6!}.$$

For small x, this error is very small but if x is large, no such conclusion can be drawn. This is illustrated in the following picture.



You see from the picture that the polynomial is a very good approximation for the function, $\sin x$ as long as $|x|$ is small but that if $|x|$ gets very large, the approximation is lousy. The above estimate indicates the good approximation holds as long as $|x|$ is small and it quantifies how good the approximation is. Suppose for example, you wanted to find $\sin(.5)$. Then from the above error estimate,

$$\left|\sin(.5)-\left((.5)-\frac{(.5)^3}{3!}+\frac{(.5)^5}{5!}\right)\right|\leq\frac{(.5)^6}{6!}=\frac{1}{46\,080}$$

so difference between the approximation and $\sin(.5)$ is less than 10^{-4}. If this is used to find $\sin(.1)$ the polynomial approximation would be even closer.

11.2 Exercises

1. Let $p_n(x)=a_0+\sum_{k=1}^{n}a_k(x-c)^k$. Show that if you require that $p_n(c)=f(c), p_n'(c)=f'(c),\cdots,p_n^{(n)}(c)=f^{(n)}(c)$, then this requirement is achieved if and only if $a_0=f(c), a_1=f'(c),\cdots,a_n=\frac{f^{(n)}(c)}{n!}$. Thus the Taylor polynomial of degree n and its first n derivatives agree with the function and its first n derivatives when $x=c$.

2. Find the Taylor polynomials for $\cos x$ for x near 0 along with a formula for the remainder. Use your approximate polynomial to compute $\cos(.5)$ to 3 decimal places and prove your approximation is this good.

3. Find the Taylor polynomials for $x^4 + 2x^3 + x - 7$ for x near 1. Prove that the Taylor polynomial of degree 4 equals the function.

4. Find the Taylor polynomials for $3x^4 + 2x^3 + x^2 - 7$ for x near -1. Prove that the Taylor polynomial of degree 4 equals the function.

5. Find the Taylor polynomials for $\cosh x$ for x near 0.

6. Find the Taylor polynomials for $\sinh x$ for x near 0.

7. Find the Taylor polynomials for $\ln(1+x)$ for x near 0.

8. Find the Taylor polynomials for $\ln(1-x)$ for x near 0.

9. Find a Taylor polynomial for $\ln\left(\frac{1+x}{1-x}\right)$ and use it to compute $\ln 5$ to three decimal places.

10. Verify that $\lim_{n\to\infty} \frac{M^n}{n!} = 0$ whenever M is a positive real number. **Hint:** Prove by induction that $M^n/n! \leq (2M)^n / (\sqrt{n})^n$. Now consider what happens when $\sqrt{n}$ is much larger than $2M$.

11. Show that for every $x \in \mathbb{R}$, $\sin(x) = \lim_{n\to\infty} \sum_{k=1}^{n} (-1)^{n-1} \frac{x^{2n-1}}{(2n-1)!}$. **Hint:** Use the formula for the error to conclude that

$$\left| \sin x - \sum_{k=1}^{n} (-1)^{k-1} \frac{x^{2k-1}}{(2k-1)!} \right| \leq \frac{|x|^{2n}}{(2n)!}$$

and then use the result of Problem 10.

12. Show $\cos(x) = \lim_{n\to\infty} \sum_{k=1}^{n} (-1)^{n-1} \frac{x^{2n-2}}{(2n-2)!}$ by finding a suitable formula for the remainder and then using an argument similar to that done in Problem 11.

13. Using Problem 10, show $e^x = \lim_{n\to\infty} \sum_{k=1}^{n} \frac{x^k}{k!}$ for all $x \in \mathbb{R}$.

14. Find a_n such that $\arctan(x) = \lim_{n\to\infty} \sum_{k=0}^{n} a_k x^k$ for some values of x. Find the values of x for which the limit is true and prove your result. **Hint:** It is a good idea to use

$$\arctan(x) = \int_0^x \frac{1}{1+t^2}\, dt,$$

show

$$\frac{1}{1+t^2} = \sum_{k=0}^{n} (-1)^k t^{2k} \pm \frac{t^{2n+2}}{1+t^2},$$

and then integrate this finite sum from 0 to x. Thus the error would be no larger than

$$\left| \int_0^x \frac{t^{2n+2}}{1+t^2}\, dt \right| \leq \left| \int_0^x t^{2n+2}\, dt \right|.$$

15. If you did Problem 14 correctly, you found

$$\arctan x = \lim_{n\to\infty} \sum_{k=1}^{n} (-1)^{k-1} \frac{x^{2k-1}}{2k-1}$$

and that this limit will hold for $x \in [-1,1]$. Use this to verify that

$$\frac{\pi}{4} = \lim_{n\to\infty} \sum_{k=1}^{n} (-1)^{k-1} \frac{1}{2k-1}.$$

16. Do for $\ln(1+x)$ what was done for $\arctan(x)$ and find a formula of this sort for $\ln 2$. Use

$$\ln(1+x) = \int_0^x \frac{1}{1+t}\, dt.$$

17. Repeat 14 for the function $\ln(1-x)$.

18. Suppose a function $y(x)$ satisfies the initial value problem, $y' = y,\ y(0) = 1$. Find Taylor polynomials for this function. Do you know this function which satisfies the given initial value problem?

19. Suppose a function, $y(x)$ satisfies the initial value problem, $y'' + y = 0,\ y(0) = 0, y'(0) = 1$. Find Taylor polynomials for this function. Do you know this function which satisfies the given initial value problem?

11.3 Infinite Series Of Numbers

11.3.1 Basic Considerations

Earlier in Definition 5.11.1 on Page 114 the notion of limit of a sequence was discussed. There is a very closely related concept called an infinite series which is dealt with in this section.

Definition 11.3.1 *Define*

$$\sum_{k=m}^{\infty} a_k \equiv \lim_{n\to\infty} \sum_{k=m}^{n} a_k$$

whenever the limit exists and is finite. In this case the series is said to converge. If it does not converge, it is said to diverge. The sequence $\{\sum_{k=m}^{n} a_k\}_{n=m}^{\infty}$ *in the above is called the sequence of partial sums.*

From this definition, it should be clear that infinite sums do not always make sense. Sometimes they do and sometimes they don't, depending on the behavior of the partial sums. As an example, consider $\sum_{k=1}^{\infty} (-1)^k$. The partial sums corresponding to this symbol alternate between -1 and 0. Therefore, there is no limit for the sequence of partial sums. It follows the symbol just written is meaningless and the infinite sum diverges.

Example 11.3.2 *Find the infinite sum,* $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$.

Note $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$ and so $\sum_{n=1}^{N} \frac{1}{n(n+1)} = \sum_{n=1}^{N} \left(\frac{1}{n} - \frac{1}{n+1}\right) = -\frac{1}{N+1} + 1$. Therefore,

$$\lim_{N\to\infty} \sum_{n=1}^{N} \frac{1}{n(n+1)} = \lim_{N\to\infty} \left(-\frac{1}{N+1} + 1\right) = 1.$$

Proposition 11.3.3 *Let $a_k \geq 0$. Then $\{\sum_{k=m}^{n} a_k\}_{n=m}^{\infty}$ is an increasing sequence. If this sequence is bounded above, then $\sum_{k=m}^{\infty} a_k$ converges and its value equals*

$$\sup\left\{\sum_{k=m}^{n} a_k : n = m, m+1, \cdots\right\}.$$

When the sequence is not bounded above, $\sum_{k=m}^{\infty} a_k$ diverges.

Proof: It follows $\{\sum_{k=m}^{n} a_k\}_{n=m}^{\infty}$ is an increasing sequence because

$$\sum_{k=m}^{n+1} a_k - \sum_{k=m}^{n} a_k = a_{n+1} \geq 0.$$

If it is bounded above, then by the form of completeness found in Theorem 5.11.15 on Page 119 it follows the sequence of partial sums converges to $\sup\{\sum_{k=m}^{n} a_k : n = m, m+1, \cdots\}$. If the sequence of partial sums is not bounded, then it is not a Cauchy sequence and so it does not converge. See Theorem 5.11.13 on Page 118. This proves the proposition.

In the case where $a_k \geq 0$, the above proposition shows there are only two alternatives available. Either the sequence of partial sums is bounded above or it is not bounded above. In the first case convergence occurs and in the second case, the infinite series diverges. For this reason, people will sometimes write $\sum_{k=m}^{\infty} a_k < \infty$ to denote the case where convergence occurs and $\sum_{k=m}^{\infty} a_k = \infty$ for the case where divergence occurs. Be very careful you never think this way in the case where it is not true that all $a_k \geq 0$. For example, the partial sums of $\sum_{k=1}^{\infty} (-1)^k$ are bounded because they are all either -1 or 0 but the series does not converge.

One of the most important examples of a convergent series is the geometric series. This series is $\sum_{n=0}^{\infty} r^n$. The study of this series depends on simple High school algebra and Theorem 5.11.9 on Page 117. Let $S_n \equiv \sum_{k=0}^{n} r^k$. Then

$$S_n = \sum_{k=0}^{n} r^k, \; rS_n = \sum_{k=0}^{n} r^{k+1} = \sum_{k=1}^{n+1} r^k.$$

Therefore, subtracting the second equation from the first yields

$$(1-r)\, S_n = 1 - r^{n+1}$$

and so a formula for S_n is available. In fact, if $r \neq 1$,

$$S_n = \frac{1 - r^{n+1}}{1-r}.$$

By Theorem 5.11.9, $\lim_{n\to\infty} S_n = \frac{1}{1-r}$ in the case when $|r| < 1$. Now if $|r| \geq 1$, the limit clearly does not exist because S_n fails to be a Cauchy sequence (Why?). This shows the following.

Theorem 11.3.4 *The geometric series, $\sum_{n=0}^{\infty} r^n$ converges and equals $\frac{1}{1-r}$ if $|r| < 1$ and diverges if $|r| \geq 1$.*

If the series do converge, the following holds.

Theorem 11.3.5 *If $\sum_{k=m}^{\infty} a_k$ and $\sum_{k=m}^{\infty} b_k$ both converge and x, y are numbers, then*

$$\sum_{k=m}^{\infty} a_k = \sum_{k=m+j}^{\infty} a_{k-j} \tag{11.1}$$

$$\sum_{k=m}^{\infty} xa_k + yb_k = x\sum_{k=m}^{\infty} a_k + y\sum_{k=m}^{\infty} b_k \tag{11.2}$$

$$\left|\sum_{k=m}^{\infty} a_k\right| \leq \sum_{k=m}^{\infty} |a_k| \tag{11.3}$$

where in the last inequality, the last sum equals $+\infty$ if the partial sums are not bounded above.

Proof: The above theorem is really only a restatement of Theorem 5.11.6 on Page 116 and the above definitions of infinite series. Thus

$$\sum_{k=m}^{\infty} a_k = \lim_{n\to\infty}\sum_{k=m}^{n} a_k = \lim_{n\to\infty}\sum_{k=m+j}^{n+j} a_{k-j} = \sum_{k=m+j}^{\infty} a_{k-j}.$$

To establish (11.2), use Theorem 5.11.6 on Page 116 to write

$$\begin{aligned}\sum_{k=m}^{\infty} xa_k + yb_k &= \lim_{n\to\infty}\sum_{k=m}^{n} xa_k + yb_k \\ &= \lim_{n\to\infty}\left(x\sum_{k=m}^{n} a_k + y\sum_{k=m}^{n} b_k\right) \\ &= x\sum_{k=m}^{\infty} a_k + y\sum_{k=m}^{\infty} b_k.\end{aligned}$$

Formula (11.3) follows from the observation that, from the triangle inequality,

$$\left|\sum_{k=m}^{n} a_k\right| \leq \sum_{k=m}^{\infty} |a_k|$$

and so

$$\left|\sum_{k=m}^{\infty} a_k\right| = \lim_{n\to\infty}\left|\sum_{k=m}^{n} a_k\right| \leq \sum_{k=m}^{\infty} |a_k|.$$

Example 11.3.6 *Find $\sum_{n=0}^{\infty}\left(\frac{5}{2^n}+\frac{6}{3^n}\right)$.*

From the above theorem and Theorem 11.3.4,

$$\begin{aligned}\sum_{n=0}^{\infty}\left(\frac{5}{2^n}+\frac{6}{3^n}\right) &= 5\sum_{n=0}^{\infty}\frac{1}{2^n} + 6\sum_{n=0}^{\infty}\frac{1}{3^n} \\ &= 5\frac{1}{1-(1/2)} + 6\frac{1}{1-(1/3)} = 19.\end{aligned}$$

The following criterion is useful in checking convergence.

Theorem 11.3.7 *The sum $\sum_{k=m}^{\infty} a_k$ converges if and only if for all $\varepsilon > 0$, there exists n_ε such that if $q \geq p \geq n_\varepsilon$, then*

$$\left|\sum_{k=p}^{q} a_k\right| < \varepsilon. \tag{11.4}$$

Proof: Suppose first that the series converges. Then $\{\sum_{k=m}^{n} a_k\}_{n=m}^{\infty}$ is a Cauchy sequence by Theorem 5.11.13 on Page 118. Therefore, there exists $n_\varepsilon > m$ such that if $q \geq p-1 \geq n_\varepsilon > m$,

$$\left|\sum_{k=m}^{q} a_k - \sum_{k=m}^{p-1} a_k\right| = \left|\sum_{k=p}^{q} a_k\right| < \varepsilon. \tag{11.5}$$

Next suppose (11.4) holds. Then from (11.5) it follows upon letting p be replaced with $p+1$ that $\{\sum_{k=m}^{n} a_k\}_{n=m}^{\infty}$ is a Cauchy sequence and so, by the completeness axiom, it converges. By the definition of infinite series, this shows the infinite sum converges as claimed.

Definition 11.3.8 *A series*

$$\sum_{k=m}^{\infty} a_k$$

is said to converge absolutely if

$$\sum_{k=m}^{\infty} |a_k|$$

converges. If the series does converge but does not converge absolutely, then it is said to converge conditionally.

Theorem 11.3.9 *If $\sum_{k=m}^{\infty} a_k$ converges absolutely, then it converges.*

Proof: Let $\varepsilon > 0$ be given. Then by assumption and Theorem 11.3.7, there exists n_ε such that whenever $q \geq p \geq n_\varepsilon$,

$$\sum_{k=p}^{q} |a_k| < \varepsilon.$$

Therefore, from the triangle inequality,

$$\varepsilon > \sum_{k=p}^{q} |a_k| \geq \left|\sum_{k=p}^{q} a_k\right|.$$

By Theorem 11.3.7, $\sum_{k=m}^{\infty} a_k$ converges and this proves the theorem.

In fact, the above theorem is really another version of the completeness axiom. Thus its validity implies completeness. You might try to show this.

Theorem 11.3.10 *(comparison test) Suppose $\{a_n\}$ and $\{b_n\}$ are sequences of non negative real numbers and suppose for all n sufficiently large, $a_n \leq b_n$. Then*

1. *If $\sum_{n=k}^{\infty} b_n$ converges, then $\sum_{n=m}^{\infty} a_n$ converges.*
2. *If $\sum_{n=k}^{\infty} a_n$ diverges, then $\sum_{n=m}^{\infty} b_n$ diverges.*

Proof: Consider the first claim. From the assumption there exists n^* such that $n^* > \max(k, m)$ and for all $n \geq n^*$ $b_n \geq a_n$. Then if $p \geq n^*$,

$$\begin{aligned}\sum_{n=m}^{p} a_n &\leq \sum_{n=m}^{n^*} a_n + \sum_{n=n^*+1}^{k} b_n \\ &\leq \sum_{n=m}^{n^*} a_n + \sum_{n=k}^{\infty} b_n.\end{aligned}$$

Thus the sequence,$\{\sum_{n=m}^{p} a_n\}_{p=m}^{\infty}$ is bounded above and increasing. Therefore, it converges by completeness. The second claim is left as an exercise.

Example 11.3.11 *Determine the convergence of* $\sum_{n=1}^{\infty} \frac{1}{n^2}$.

For $n > 1$,

$$\frac{1}{n^2} \leq \frac{1}{n\,(n-1)}.$$

Now

$$\begin{aligned} \sum_{n=2}^{p} \frac{1}{n\,(n-1)} &= \sum_{n=2}^{p} \left[\frac{1}{n-1} - \frac{1}{n}\right] \\ &= 1 - \frac{1}{p} \to 1 \end{aligned}$$

Therefore, letting $a_n = \frac{1}{n^2}$ and $b_n = \frac{1}{n(n-1)}$

A convenient way to implement the comparison test is to use the limit comparison test. This is considered next.

Theorem 11.3.12 *Let* $a_n, b_n > 0$ *and suppose for all* n *large enough,*

$$0 < a < \frac{a_n}{b_n} \leq \frac{a_n}{b_n} < b < \infty.$$

Then $\sum a_n$ *and* $\sum b_n$ *converge or diverge together.*

Proof: Let n^* be such that $n \geq n^*$, then

$$\frac{a_n}{b_n} > a \text{ and } \frac{a_n}{b_n} < b$$

and so for all such n,

$$ab_n < a_n < bb_n$$

and so the conclusion follows from the comparison test.

Example 11.3.13 *Determine the convergence of* $\sum_{k=1}^{\infty} \frac{1}{\sqrt{n^4+2n+7}}$.

This series converges by the limit comparison test. Compare with the series of Example 11.3.11.

$$\begin{aligned} \lim_{n\to\infty} \frac{\left(\frac{1}{n^2}\right)}{\left(\frac{1}{\sqrt{n^4+2n+7}}\right)} &= \lim_{n\to\infty} \frac{\sqrt{n^4+2n+7}}{n^2} \\ &= \lim_{n\to\infty} \sqrt{1 + \frac{2}{n^3} + \frac{7}{n^4}} = 1. \end{aligned}$$

Therefore, the series converges with the series of Example 11.3.11. How did I know what to compare with? I noticed that $\sqrt{n^4+2n+7}$ is essentially like $\sqrt{n^4} = n^2$ for large enough n. You see, the higher order term, n^4 dominates the other terms in n^4+2n+7. Therefore, reasoning that $1/\sqrt{n^4+2n+7}$ is a lot like $1/n^2$ for large n, it was easy to see what to compare with. Of course this is not always easy and there is room for acquiring skill through practice.

To really exploit this limit comparison test, it is desirable to get lots of examples of series, some which converge and some which do not. The tool for obtaining these examples here will be the following wonderful theorem known as the Cauchy condensation test.

Theorem 11.3.14 *Let $a_n \geq 0$ and suppose the terms of the sequence $\{a_n\}$ are decreasing. Thus $a_n \geq a_{n+1}$ for all n. Then*

$$\sum_{n=1}^{\infty} a_n \text{ and } \sum_{n=0}^{\infty} 2^n a_{2^n}$$

converge or diverge together.

Proof: This follows from the inequality of the following claim.
Claim:

$$\sum_{k=1}^{n} 2^k a_{2^{k-1}} \geq \sum_{k=1}^{2^n} a_k \geq \sum_{k=0}^{n} 2^{k-1} a_{2^k}.$$

Proof of the Claim: Note the claim is true for $n = 1$. Suppose the claim is true for n. Then, since $2^{n+1} - 2^n = 2^n$, and the terms, a_n, are decreasing,

$$\sum_{k=1}^{n+1} 2^k a_{2^{k-1}} = 2^{n+1} a_{2^n} + \sum_{k=1}^{n} 2^k a_{2^{k-1}} \geq 2^{n+1} a_{2^n} + \sum_{k=1}^{2^n} a_k$$

$$\geq \sum_{k=1}^{2^{n+1}} a_k \geq 2^n a_{2^{n+1}} + \sum_{k=1}^{2^n} a_k \geq 2^n a_{2^{n+1}} + \sum_{k=0}^{n} 2^{k-1} a_{2^k} = \sum_{k=0}^{n+1} 2^{k-1} a_{2^k}.$$

Example 11.3.15 *Determine the convergence of $\sum_{k=1}^{\infty} \frac{1}{k^p}$ where p is a positive number. These are called the p series.*

Let $a_n = \frac{1}{n^p}$. Then $a_{2^n} = \left(\frac{1}{2^p}\right)^n$. From the Cauchy condensation test the two series

$$\sum_{n=1}^{\infty} \frac{1}{n^p} \text{ and } \sum_{n=0}^{\infty} 2^n \left(\frac{1}{2^p}\right)^n = \sum_{n=0}^{\infty} \left(2^{(1-p)}\right)^n$$

converge or diverge together. If $p > 1$, the last series above is a geometric series having common ratio less than 1 and so it converges. If $p \leq 1$, it is still a geometric series but in this case the common ratio is either 1 or greater than 1 so the series diverges. It follows that the p series converges if $p > 1$ and diverges if $p \leq 1$. In particular, $\sum_{n=1}^{\infty} n^{-1}$ diverges while $\sum_{n=1}^{\infty} n^{-2}$ converges.

Example 11.3.16 *Determine the convergence of $\sum_{k=1}^{\infty} \frac{1}{\sqrt{n^2+100n}}$.*

Use the limit comparison test.

$$\lim_{n\to\infty} \frac{\left(\frac{1}{n}\right)}{\left(\frac{1}{\sqrt{n^2+100n}}\right)} = 1$$

and so this series diverges with $\sum_{k=1}^{\infty} \frac{1}{k}$.

Example 11.3.17 *Determine the convergence of $\sum_{k=2}^{\infty} \frac{1}{k \ln k}$.*

Use the Cauchy condensation test. The above series does the same thing in terms of convergence as the series

$$\sum_{n=1}^{\infty} 2^n \frac{1}{2^n \ln(2^n)} = \sum_{n=1}^{\infty} \frac{1}{n \ln 2}$$

and this series diverges by limit comparison with the series $\sum \frac{1}{n}$.

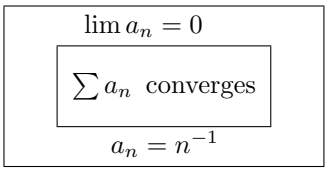
Sometimes it is good to be able to say a series does not converge. The n^{th} term test gives such a condition which is sufficient for this. It is really a corollary of Theorem 11.3.7.

Theorem 11.3.18 *If* $\sum_{n=m}^{\infty} a_n$ *converges, then* $\lim_{n\to\infty} a_n = 0$.

Proof: Apply Theorem 11.3.7 to conclude that

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \sum_{k=n}^{n} a_k = 0.$$

It is very important to observe that this theorem goes only in one direction. That is, you cannot conclude the series converges if $\lim_{n\to\infty} a_n = 0$. If this happens, you don't know anything from this information. Recall $\lim_{n\to\infty} n^{-1} = 0$ but $\sum_{n=1}^{\infty} n^{-1}$ diverges. The following picture is descriptive of the situation.



11.4 Exercises

1. Determine whether the following series converge and give reasons for your answers.

   (a) $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n^2+n+1}}$

   (b) $\sum_{n=1}^{\infty} \left(\sqrt{n+1} - \sqrt{n}\right)$

   (c) $\sum_{n=1}^{\infty} \frac{(n!)^2}{(2n)!}$

   (d) $\sum_{n=1}^{\infty} \frac{(2n)!}{(n!)^2}$

   (e) $\sum_{n=1}^{\infty} \frac{1}{2n+2}$

   (f) $\sum_{n=1}^{\infty} \left(\frac{n}{n+1}\right)^n$

   (g) $\sum_{n=1}^{\infty} \left(\frac{n}{n+1}\right)^{n^2}$

2. Determine whether the following series converge give reasons for your answers.

   (a) $\sum_{n=1}^{\infty} \frac{\ln\left(k^5\right)}{k}$

   (b) $\sum_{n=1}^{\infty} \frac{\ln\left(k^5\right)}{k^{1.01}}$

   (c) $\sum_{n=1}^{\infty} \sin\left(\frac{1}{n}\right)$

   (d) $\sum_{n=1}^{\infty} \tan\left(\frac{1}{n^2}\right)$

   (e) $\sum_{n=1}^{\infty} \cos\left(\frac{1}{n^2}\right)$

   (f) $\sum_{n=1}^{\infty} \sin\left(\frac{\sqrt{n}}{n^2+1}\right)$

3. Determine whether the following series converge and give reasons for your answers.

   (a) $\sum_{n=1}^{\infty} \frac{2^n+n}{n2^n}$

(b) $\sum_{n=1}^{\infty} \frac{2^n+n}{n^2 2^n}$

(c) $\sum_{n=1}^{\infty} \frac{n}{2n+1}$

(d) $\sum_{n=1}^{\infty} \frac{n^{100}}{1.01^n}$

(e) $\sum_{n=1}^{\infty} \frac{\ln n}{n^2}$

4. Find the exact values of the following infinite series if they converge.

(a) $\sum_{k=3}^{\infty} \frac{1}{k(k-2)}$

(b) $\sum_{k=1}^{\infty} \frac{1}{k(k+1)}$

(c) $\sum_{k=3}^{\infty} \frac{1}{(k+1)(k-2)}$

(d) $\sum_{k=1}^{N} \left(\frac{1}{\sqrt{k}} - \frac{1}{\sqrt{k+1}} \right)$

(e) $\sum_{n=1}^{\infty} \ln \left(\frac{(n+1)^2}{n(n+2)} \right)$

5. Suppose $\sum_{k=1}^{\infty} a_k$ converges and each $a_k \geq 0$. Does it follow that $\sum_{k=1}^{\infty} a_k^2$ also converges?

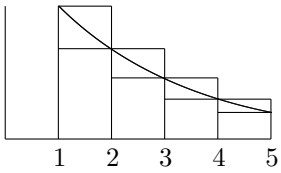
6. Find a series which diverges using one test but converges using another if possible. If this is not possible, tell why.

7. If $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ both converge and a_n, b_n are nonnegative, can you conclude the sum, $\sum_{n=1}^{\infty} a_n b_n$ converges?

8. If $\sum_{n=1}^{\infty} a_n$ converges and $a_n \geq 0$ for all n and b_n is bounded, can you conclude $\sum_{n=1}^{\infty} a_n b_n$ converges?

9. The logarithm test states the following. Suppose $a_k \geq 0$ and $a_k > 0$ for large k and that $p = \lim_{k\to\infty} \frac{\ln\left(\frac{1}{a_k}\right)}{\ln k}$ exists. If $p > 1$, then $\sum_{k=1}^{\infty} a_k$ converges. If $p < 1$, then the series, $\sum_{k=1}^{\infty} a_k$ does not converge. Prove this theorem.

10. Suppose f is a nonnegative continuous decreasing function defined on $[1, \infty)$. Show the improper integral, $\int_1^{\infty} f(t)\, dt$ and the sum $\sum_{k=1}^{\infty} f(k)$ converge or diverge together. This is called the integral test. Use this test to verify convergence of $\sum_{k=1}^{\infty} \frac{1}{k^{\alpha}}$ whenever $\alpha > 1$ and divergence whenever $\alpha \leq 1$. In showing this integral test, it might be helpful to consider the following picture.



In this picture, the graph of the continuous decreasing function is represented. There are rectangles below this curve and rectangles above it. $f(1)$ is the area of the first rectangle on the left which is above the curve while $f(2)$ is the area of the first rectangle on the left which lies below the curve. From the picture, $\int_1^5 f(t)\, dt$ lies between $f(2)+f(3)+f(4)+f(5)$ and $f(1)+f(2)+f(3)+f(4)$. Generalize to conclude that for any $n \in \mathbb{N}$, it follows that $\sum_{k=2}^{n} f(k) \leq \int_1^n f(t)\, dt \leq \sum_{k=1}^{n-1} f(k)$. Now explain why this inequality implies the integral and sum have the same convergence properties.

11. Show using either the Cauchy condensation test or the integral test that $\sum_{k=4}^{\infty} \frac{1}{\ln(\ln(k))\ln(k)k}$ diverges. Estimate how big N must be in order that $\sum_{k=4}^{N} \frac{1}{\ln(\ln(k))\ln(k)k} > 10$. What does this tell you about the wisdom of attempting to determine questions of convergence through experimentation?

12. For p a positive number, determine the convergence of $\sum_{n=2}^{\infty} \frac{1}{n(\ln(n))^p}$ for various values of p.

13. For p a positive number, determine the convergence of $\sum_{n=2}^{\infty} \frac{\ln n}{n^p}$ for various values of p.

14. Determine the convergence of the series $\sum_{n=1}^{\infty} \left(\sum_{k=1}^{n} \frac{1}{k}\right)^{-n/2}$.

11.4.1 More Tests For Convergence

So far, the tests for convergence have been applied to non negative terms only. Sometimes, a series converges, not because the terms of the series get small fast enough, but because of cancellation taking place between positive and negative terms. A discussion of this involves some simple algebra.

Let $\{a_n\}$ and $\{b_n\}$ be sequences and let

$$A_n \equiv \sum_{k=1}^{n} a_k,\ A_{-1} \equiv A_0 \equiv 0.$$

Then if $p < q$

$$\sum_{n=p}^{q} a_n b_n = \sum_{n=p}^{q} b_n \left(A_n - A_{n-1}\right) = \sum_{n=p}^{q} b_n A_n - \sum_{n=p}^{q} b_n A_{n-1}$$

$$= \sum_{n=p}^{q} b_n A_n - \sum_{n=p-1}^{q-1} b_{n+1} A_n = b_q A_q - b_p A_{p-1} + \sum_{n=p}^{q-1} A_n \left(b_n - b_{n+1}\right)$$

This formula is called the partial summation formula. It is just like integration by parts.

Theorem 11.4.1 *(Dirichlet's test) Suppose A_n is bounded and $\lim_{n\to\infty} b_n = 0$, with $b_n \geq b_{n+1}$. Then*

$$\sum_{n=1}^{\infty} a_n b_n$$

converges.

Proof: This follows quickly from Theorem 11.3.7. Indeed, letting $|A_n| \leq C$, and using the partial summation formula above along with the assumption that the b_n are decreasing,

$$\left|\sum_{n=p}^{q} a_n b_n\right| = \left|b_q A_q - b_p A_{p-1} + \sum_{n=p}^{q-1} A_n \left(b_n - b_{n+1}\right)\right|$$

$$\leq C\left(|b_q| + |b_p|\right) + C\sum_{n=p}^{q-1} \left(b_n - b_{n+1}\right)$$

$$= C\left(|b_q| + |b_p|\right) + C\left(b_p - b_q\right)$$

and by assumption, this last expression is small whenever p and q are sufficiently large. This proves the theorem.

Definition 11.4.2 *If $b_n > 0$ for all n, a series of the form $\sum_k (-1)^k b_k$ or $\sum_k (-1)^{k-1} b_k$ is known as an* ***alternating series****.*

The following corollary is known as the alternating series test.

Corollary 11.4.3 *(alternating series test) If $\lim_{n\to\infty} b_n = 0$, with $b_n \geq b_{n+1}$, then $\sum_{n=1}^{\infty} (-1)^n b_n$ converges.*

Proof: Let $a_n = (-1)^n$. Then the partial sums of $\sum_n a_n$ are bounded and so Theorem 11.4.1 applies.

In the situation of Corollary 11.4.3 there is a convenient error estimate available.

Theorem 11.4.4 *Let $b_n > 0$ for all n such that $b_n \geq b_{n+1}$ for all n and $\lim_{n\to\infty} b_n = 0$ and consider either $\sum_{n=1}^{\infty} (-1)^n b_n$ or $\sum_{n=1}^{\infty} (-1)^{n-1} b_n$. Then*

$$\left|\sum_{n=1}^{\infty} (-1)^n b_n - \sum_{n=1}^{N} (-1)^n b_n\right| \leq |b_{N+1}|, \left|\sum_{n=1}^{\infty} (-1)^{n-1} b_n - \sum_{n=1}^{N} (-1)^{n-1} b_n\right| \leq |b_{N+1}|$$

Example 11.4.5 *How many terms must I take in the sum, $\sum_{n=1}^{\infty} (-1)^n \frac{1}{n^2+1}$ to be closer than $\frac{1}{10}$ to $\sum_{n=1}^{\infty} (-1)^n \frac{1}{n^2+1}$?*

From Theorem 11.4.4, I need to find n such that $\frac{1}{n^2+1} \leq \frac{1}{10}$ and then $n-1$ is the desired value. Thus $n = 3$ and so

$$\left|\sum_{n=1}^{\infty} (-1)^n \frac{1}{n^2+1} - \sum_{n=1}^{2} (-1)^n \frac{1}{n^2+1}\right| \leq \frac{1}{10}$$

A favorite test for convergence is the ratio test. This is discussed next.

Theorem 11.4.6 *Suppose $|a_n| > 0$ for all n and suppose*

$$\lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} = r.$$

Then

$$\sum_{n=1}^{\infty} a_n \begin{cases} \text{diverges if } r > 1 \\ \text{converges absolutely if } r < 1 \\ \text{test fails if } r = 1 \end{cases}.$$

Proof: Suppose $r < 1$. Then there exists n_1 such that if $n \geq n_1$, then

$$0 < \left|\frac{a_{n+1}}{a_n}\right| < R$$

where $r < R < 1$. Then

$$|a_{n+1}| < R\,|a_n|$$

for all such n. Therefore,

$$|a_{n_1+p}| < R\,|a_{n_1+p-1}| < R^2\,|a_{n_1+p-2}| < \cdots < R^p\,|a_{n_1}| \tag{11.6}$$

and so if $m > n$, then $|a_m| < R^{m-n_1}\,|a_{n_1}|$. By the comparison test and the theorem on geometric series, $\sum |a_n|$ converges. This proves the convergence part of the theorem.

To verify the divergence part, note that if $r > 1$, then (11.6) can be turned around for some $R > 1$. Showing $\lim_{n\to\infty} |a_n| = \infty$. Since the n^{th} term fails to converge to 0, it follows the series diverges.

To see the test fails if $r = 1$, consider $\sum n^{-1}$ and $\sum n^{-2}$. The first series diverges while the second one converges but in both cases, $r = 1$. (Be sure to check this last claim.)

The ratio test is very useful for many different examples but it is somewhat unsatisfactory mathematically. One reason for this is the assumption that $a_n > 0$, necessitated by the need to divide by a_n, and the other reason is the possibility that the limit might not exist. The next test, called the root test removes both of these objections.

Theorem 11.4.7 *Suppose* $|a_n|^{1/n} < R < 1$ *for all* n *sufficiently large. Then*

$$\sum_{n=1}^{\infty} a_n \text{ converges absolutely.}$$

If there are infinitely many values of n *such that* $|a_n|^{1/n} \geq 1$, *then*

$$\sum_{n=1}^{\infty} a_n \text{ diverges.}$$

Proof: Suppose first that $|a_n|^{1/n} < R < 1$ for all n sufficiently large. Say this holds for all $n \geq n_R$. Then for such n,

$$\sqrt[n]{|a_n|} < R.$$

Therefore, for such n,

$$|a_n| \leq R^n$$

and so the comparison test with a geometric series applies and gives absolute convergence as claimed.

Next suppose $|a_n|^{1/n} \geq 1$ for infinitely many values of n. Then for those values of n, $|a_n| \geq 1$ and so the series fails to converge by the n^{th} term test.

Corollary 11.4.8 *Suppose* $\lim_{n\to\infty} |a_n|^{1/n}$ *exists and equals* r. *Then*

$$\sum_{k=m}^{\infty} a_k \begin{cases} \text{converges absolutely if } r < 1 \\ \text{test fails if } r = 1 \\ \text{diverges if } r > 1 \end{cases}$$

Proof: The first and last alternatives follow from Theorem 11.4.7. To see the test fails if $r = 1$, consider the two series $\sum_{n=1}^{\infty} \frac{1}{n}$ and $\sum_{n=1}^{\infty} \frac{1}{n^2}$ both of which have $r = 1$ but having different convergence properties.

Example 11.4.9 *Show that for all* $x \in \mathbb{R}$,

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}. \tag{11.7}$$

By Taylor's theorem

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + e^{\xi_n} \frac{x^{n+1}}{(n+1)!} \\ &= \sum_{k=0}^{n} \frac{x^k}{k!} + e^{\xi_n} \frac{x^{n+1}}{(n+1)!} \end{aligned} \tag{11.8}$$

where $|\xi_n| \leq |x|$. Now for any $x \in \mathbb{R}$

$$\left| e^{\xi_n} \frac{x^{n+1}}{(n+1)!} \right| \leq e^{|x|} \frac{|x|^{n+1}}{(n+1)!}$$

and an application of the ratio test shows

$$\sum_{n=0}^{\infty} e^{|x|} \frac{|x|^{n+1}}{(n+1)!} < \infty.$$

Therefore, the n^{th} term converges to zero by the n^{th} term test and so for each $x \in \mathbb{R}$,

$$\lim_{n\to\infty} e^{\xi_n} \frac{x^{n+1}}{(n+1)!} = 0.$$

Therefore, taking the limit in (11.8) it follows (11.7) holds.

11.4.2 Double Series*

Sometimes it is required to consider double series which are of the form

$$\sum_{k=m}^{\infty} \sum_{j=m}^{\infty} a_{jk} \equiv \sum_{k=m}^{\infty} \left(\sum_{j=m}^{\infty} a_{jk} \right).$$

In other words, first sum on j yielding something which depends on k and then sum these. The major consideration for these double series is the question of when

$$\sum_{k=m}^{\infty} \sum_{j=m}^{\infty} a_{jk} = \sum_{j=m}^{\infty} \sum_{k=m}^{\infty} a_{jk}.$$

In other words, when does it make no difference which subscript is summed over first? In the case of finite sums there is no issue here. You can always write

$$\sum_{k=m}^{M} \sum_{j=m}^{N} a_{jk} = \sum_{j=m}^{N} \sum_{k=m}^{M} a_{jk}$$

because addition is commutative. However, there are limits involved with infinite sums and the interchange in order of summation involves taking limits in a different order. Therefore, it is not always true that it is permissible to interchange the two sums. A general rule of thumb is this: If something involves changing the order in which two limits are taken, you may not do it without agonizing over the question. In general, limits foul up algebra and also introduce things which are counter intuitive. Here is an example. This example is a little technical. It is placed here just to prove conclusively there is a question which needs to be considered.

Example 11.4.10 *Consider the following picture which depicts some of the ordered pairs (m, n) where m, n are positive integers.*

The numbers next to the point are the values of a_{mn}. You see $a_{nn} = 0$ for all n, $a_{21} = a, a_{12} = b, a_{mn} = c$ for (m,n) on the line $y = 1 + x$ whenever $m > 1$, and $a_{mn} = -c$ for all (m,n) on the line $y = x - 1$ whenever $m > 2$.

Then $\sum_{m=1}^{\infty} a_{mn} = a$ if $n = 1$, $\sum_{m=1}^{\infty} a_{mn} = b - c$ if $n = 2$ and if $n > 2, \sum_{m=1}^{\infty} a_{mn} = 0$. Therefore,

$$\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_{mn} = a + b - c.$$

Next observe that $\sum_{n=1}^{\infty} a_{mn} = b$ if $m = 1, \sum_{n=1}^{\infty} a_{mn} = a + c$ if $m = 2$, and $\sum_{n=1}^{\infty} a_{mn} = 0$ if $m > 2$. Therefore,

$$\sum_{m=1}^{\infty}\sum_{n=1}^{\infty} a_{mn} = b + a + c$$

and so the two sums are different. Moreover, you can see that by assigning different values of a, b, and c, you can get an example for any two different numbers desired.

Don't become upset by this. It happens because, as indicated above, limits are taken in two different orders. An infinite sum always involves a limit and this illustrates why you must always remember this. This example in no way violates the commutative law of addition which has nothing to do with limits. However, it turns out that if $a_{ij} \geq 0$ for all i, j, then you can always interchange the order of summation. This is shown next and is based on the following lemma. First, some notation should be discussed.

Definition 11.4.11 *Let $f(a,b) \in [-\infty, \infty]$ for $a \in A$ and $b \in B$ where A, B are sets which means that $f(a,b)$ is either a number, ∞, or $-\infty$. The symbol, $+\infty$ is interpreted as a point out at the end of the number line which is larger than every real number. Of course there is*

no such number. That is why it is called ∞. *The symbol,* $-\infty$ *is interpreted similarly. Then* $\sup_{a\in A} f(a,b)$ *means* $\sup(S_b)$ *where* $S_b \equiv \{f(a,b) : a \in A\}$.

Unlike limits, you can take the sup in different orders.

Lemma 11.4.12 *Let* $f(a,b) \in [-\infty,\infty]$ *for* $a \in A$ *and* $b \in B$ *where* A, B *are sets. Then*

$$\sup_{a\in A}\sup_{b\in B} f(a,b) = \sup_{b\in B}\sup_{a\in A} f(a,b).$$

Proof: Note that for all a, b, $f(a,b) \le \sup_{b\in B}\sup_{a\in A} f(a,b)$ and therefore, for all a, $\sup_{b\in B} f(a,b) \le \sup_{b\in B}\sup_{a\in A} f(a,b)$. Therefore,

$$\sup_{a\in A}\sup_{b\in B} f(a,b) \le \sup_{b\in B}\sup_{a\in A} f(a,b).$$

Repeat the same argument interchanging a and b, to get the conclusion of the lemma.

Lemma 11.4.13 *If* $\{A_n\}$ *is an increasing sequence in* $[-\infty,\infty]$, *then* $\sup\{A_n\} = \lim_{n\to\infty} A_n$.

Proof: Let $\sup(\{A_n : n \in \mathbb{N}\}) = r$. In the first case, suppose $r < \infty$. Then letting $\varepsilon > 0$ be given, there exists n such that $A_n \in (r-\varepsilon, r]$. Since $\{A_n\}$ is increasing, it follows if $m > n$, then $r - \varepsilon < A_n \le A_m \le r$ and so $\lim_{n\to\infty} A_n = r$ as claimed. In the case where $r = \infty$, then if a is a real number, there exists n such that $A_n > a$. Since $\{A_k\}$ is increasing, it follows that if $m > n$, $A_m > a$. But this is what is meant by $\lim_{n\to\infty} A_n = \infty$. The other case is that $r = -\infty$. But in this case, $A_n = -\infty$ for all n and so $\lim_{n\to\infty} A_n = -\infty$.

Theorem 11.4.14 *Let* $a_{ij} \ge 0$. *Then* $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}$.

Proof: First note there is no trouble in defining these sums because the a_{ij} are all nonnegative. If a sum diverges, it only diverges to ∞ and so ∞ is the value of the sum. Next note that

$$\sum_{j=r}^{\infty}\sum_{i=r}^{\infty} a_{ij} \ge \sup_n \sum_{j=r}^{\infty}\sum_{i=r}^{n} a_{ij}$$

because for all j,

$$\sum_{i=r}^{\infty} a_{ij} \ge \sum_{i=r}^{n} a_{ij}.$$

Therefore,

$$\sum_{j=r}^{\infty}\sum_{i=r}^{\infty} a_{ij} \ge \sup_n \sum_{j=r}^{\infty}\sum_{i=r}^{n} a_{ij} = \sup_n \lim_{m\to\infty} \sum_{j=r}^{m}\sum_{i=r}^{n} a_{ij}$$

$$= \sup_n \lim_{m\to\infty} \sum_{i=r}^{n}\sum_{j=r}^{m} a_{ij} = \sup_n \sum_{i=r}^{n} \lim_{m\to\infty} \sum_{j=r}^{m} a_{ij}$$

$$= \sup_n \sum_{i=r}^{n}\sum_{j=r}^{\infty} a_{ij} = \lim_{n\to\infty} \sum_{i=r}^{n}\sum_{j=r}^{\infty} a_{ij} = \sum_{i=r}^{\infty}\sum_{j=r}^{\infty} a_{ij}$$

Interchanging the i and j in the above argument proves the theorem.

The following is the fundamental result on double sums.

Theorem 11.4.15 *Let a_{ij} be a number and suppose*

$$\sum_{i=r}^{\infty}\sum_{j=r}^{\infty}|a_{ij}|<\infty\ .$$

Then

$$\sum_{i=r}^{\infty}\sum_{j=r}^{\infty}a_{ij}=\sum_{j=r}^{\infty}\sum_{i=r}^{\infty}a_{ij}$$

and every infinite sum encountered in the above equation converges.

Proof: By Theorem 11.4.14

$$\sum_{j=r}^{\infty}\sum_{i=r}^{\infty}|a_{ij}|=\sum_{i=r}^{\infty}\sum_{j=r}^{\infty}|a_{ij}|<\infty$$

Therefore, for each j, $\sum_{i=r}^{\infty}|a_{ij}|<\infty$ and for each i, $\sum_{j=r}^{\infty}|a_{ij}|<\infty$. By Theorem 11.3.9 on Page 306, $\sum_{i=r}^{\infty}a_{ij}$, $\sum_{j=r}^{\infty}a_{ij}$ both converge, the first one for every j and the second for every i. Also,

$$\sum_{j=r}^{\infty}\left|\sum_{i=r}^{\infty}a_{ij}\right|\leq\sum_{j=r}^{\infty}\sum_{i=r}^{\infty}|a_{ij}|<\infty$$

and

$$\sum_{i=r}^{\infty}\left|\sum_{j=r}^{\infty}a_{ij}\right|\leq\sum_{i=r}^{\infty}\sum_{j=r}^{\infty}|a_{ij}|<\infty$$

so by Theorem 11.3.9 again,

$$\sum_{j=r}^{\infty}\sum_{i=r}^{\infty}a_{ij},\ \sum_{i=r}^{\infty}\sum_{j=r}^{\infty}a_{ij}$$

both exist. It only remains to verify they are equal. Note $0\leq(|a_{ij}|+a_{ij})\leq|a_{ij}|$. Therefore, by Theorem 11.4.14 and Theorem 11.3.5 on Page 304

$$\begin{aligned}\sum_{j=r}^{\infty}\sum_{i=r}^{\infty}|a_{ij}|+\sum_{j=r}^{\infty}\sum_{i=r}^{\infty}a_{ij} &= \sum_{j=r}^{\infty}\sum_{i=r}^{\infty}(|a_{ij}|+a_{ij})\\ &= \sum_{i=r}^{\infty}\sum_{j=r}^{\infty}(|a_{ij}|+a_{ij})\\ &= \sum_{i=r}^{\infty}\sum_{j=r}^{\infty}|a_{ij}|+\sum_{i=r}^{\infty}\sum_{j=r}^{\infty}a_{ij}\\ &= \sum_{j=r}^{\infty}\sum_{i=r}^{\infty}|a_{ij}|+\sum_{i=r}^{\infty}\sum_{j=r}^{\infty}a_{ij}\end{aligned}$$

and so $\sum_{j=r}^{\infty}\sum_{i=r}^{\infty}a_{ij}=\sum_{i=r}^{\infty}\sum_{j=r}^{\infty}a_{ij}$ as claimed. This proves the theorem.

One of the most important applications of this theorem is to the problem of multiplication of series.

Definition 11.4.16 *Let $\sum_{i=r}^{\infty}a_i$ and $\sum_{i=r}^{\infty}b_i$ be two series. For $n\geq r$, define*

$$c_n\equiv\sum_{k=r}^{n}a_kb_{n-k+r}.$$

The series $\sum_{n=r}^{\infty}c_n$ is called the Cauchy product of the two series.

It isn't hard to see where this comes from. Formally write the following in the case $r = 0$:

$$(a_0 + a_1 + a_2 + a_3 \cdots)(b_0 + b_1 + b_2 + b_3 \cdots)$$

and start multiplying in the usual way. This yields

$$a_0b_0 + (a_0b_1 + b_0a_1) + (a_0b_2 + a_1b_1 + a_2b_0) + \cdots$$

and you see the expressions in parentheses above are just the c_n for $n = 0, 1, 2, \cdots$. Therefore, it is reasonable to conjecture that

$$\sum_{i=r}^{\infty} a_i \sum_{j=r}^{\infty} b_j = \sum_{n=r}^{\infty} c_n$$

and of course there would be no problem with this in the case of finite sums but in the case of infinite sums, it is necessary to prove a theorem. The following is a special case of Merten's theorem.

Theorem 11.4.17 *Suppose $\sum_{i=r}^{\infty} a_i$ and $\sum_{j=r}^{\infty} b_j$ both converge absolutely*[1]*. Then*

$$\left(\sum_{i=r}^{\infty} a_i\right)\left(\sum_{j=r}^{\infty} b_j\right) = \sum_{n=r}^{\infty} c_n$$

where

$$c_n = \sum_{k=r}^{n} a_k b_{n-k+r}.$$

Proof: Let $p_{nk} = 1$ if $r \leq k \leq n$ and $p_{nk} = 0$ if $k > n$. Then

$$c_n = \sum_{k=r}^{\infty} p_{nk} a_k b_{n-k+r}.$$

Also,

$$\begin{aligned}
\sum_{k=r}^{\infty}\sum_{n=r}^{\infty} p_{nk}\,|a_k|\,|b_{n-k+r}| &= \sum_{k=r}^{\infty}|a_k|\sum_{n=r}^{\infty} p_{nk}\,|b_{n-k+r}| \\
&= \sum_{k=r}^{\infty}|a_k|\sum_{n=k}^{\infty}|b_{n-k+r}| \\
&= \sum_{k=r}^{\infty}|a_k|\sum_{n=k}^{\infty}\left|b_{n-(k-r)}\right| \\
&= \sum_{k=r}^{\infty}|a_k|\sum_{m=r}^{\infty}|b_m| < \infty.
\end{aligned}$$

Therefore, by Theorem 11.4.15

$$\begin{aligned}
\sum_{n=r}^{\infty} c_n &= \sum_{n=r}^{\infty}\sum_{k=r}^{n} a_k b_{n-k+r} = \sum_{n=r}^{\infty}\sum_{k=r}^{\infty} p_{nk} a_k b_{n-k+r} \\
&= \sum_{k=r}^{\infty} a_k \sum_{n=r}^{\infty} p_{nk} b_{n-k+r} = \sum_{k=r}^{\infty} a_k \sum_{n=k}^{\infty} b_{n-k+r} \\
&= \sum_{k=r}^{\infty} a_k \sum_{m=r}^{\infty} b_m
\end{aligned}$$

[1] Actually, it is only necessary to assume one of the series converges and the other converges absolutely. This is known as Merten's theorem and may be read in the 1974 book by Apostol listed in the bibliography.

and this proves the theorem.

11.5 Exercises

1. Determine whether the following series converge absolutely, conditionally, or not at all and give reasons for your answers.

   (a) $\sum_{n=1}^{\infty}(-1)^n \frac{1}{\sqrt{n^2+n+1}}$
   (b) $\sum_{n=1}^{\infty}(-1)^n \left(\sqrt{n+1}-\sqrt{n}\right)$
   (c) $\sum_{n=1}^{\infty}(-1)^n \frac{(n!)^2}{(2n)!}$
   (d) $\sum_{n=1}^{\infty}(-1)^n \frac{(2n)!}{(n!)^2}$
   (e) $\sum_{n=1}^{\infty} \frac{(-1)^n}{2n+2}$
   (f) $\sum_{n=1}^{\infty}(-1)^n \left(\frac{n}{n+1}\right)^n$
   (g) $\sum_{n=1}^{\infty}(-1)^n \left(\frac{n}{n+1}\right)^{n^2}$

2. Determine whether the following series converge absolutely, conditionally, or not at all and give reasons for your answers.

   (a) $\sum_{n=1}^{\infty}(-1)^n \frac{\ln\left(k^5\right)}{k}$
   (b) $\sum_{n=1}^{\infty}(-1)^n \frac{\ln\left(k^5\right)}{k^{1.01}}$
   (c) $\sum_{n=1}^{\infty}(-1)^n \frac{10^n}{(1.01)^n}$
   (d) $\sum_{n=1}^{\infty}(-1)^n \sin\left(\frac{1}{n}\right)$
   (e) $\sum_{n=1}^{\infty}(-1)^n \tan\left(\frac{1}{n^2}\right)$
   (f) $\sum_{n=1}^{\infty}(-1)^n \cos\left(\frac{1}{n^2}\right)$
   (g) $\sum_{n=1}^{\infty}(-1)^n \sin\left(\frac{\sqrt{n}}{n^2+1}\right)$

3. Determine whether the following series converge absolutely, conditionally, or not at all and give reasons for your answers.

   (a) $\sum_{n=1}^{\infty}(-1)^n \frac{2^n+n}{n2^n}$
   (b) $\sum_{n=1}^{\infty}(-1)^n \frac{2^n+n}{n^2 2^n}$
   (c) $\sum_{n=1}^{\infty}(-1)^n \frac{n}{2n+1}$
   (d) $\sum_{n=1}^{\infty}(-1)^n \frac{10^n}{n!}$
   (e) $\sum_{n=1}^{\infty}(-1)^n \frac{n^{100}}{1.01^n}$
   (f) $\sum_{n=1}^{\infty}(-1)^n \frac{\ln n}{n^2}$
   (g) $\sum_{n=1}^{\infty}(-1)^n \frac{3^n}{n^3}$
   (h) $\sum_{n=1}^{\infty}(-1)^n \frac{n^3}{3^n}$
   (i) $\sum_{n=1}^{\infty}(-1)^n \frac{n^3}{n!}$
   (j) $\sum_{n=1}^{\infty}(-1)^n \frac{n!}{n^{100}}$

4. Find the exact values of the following infinite series if they converge.

   (a) $\sum_{k=3}^{\infty} \frac{1}{k(k-2)}$

   (b) $\sum_{k=1}^{\infty} \frac{1}{k(k+1)}$

   (c) $\sum_{k=3}^{\infty} \frac{1}{(k+1)(k-2)}$

   (d) $\sum_{k=1}^{N} \left(\frac{1}{\sqrt{k}} - \frac{1}{\sqrt{k+1}} \right)$

   (e) $\sum_{n=1}^{\infty} \ln \left(\frac{(n+1)^2}{n(n+2)} \right)$

5. Suppose $\sum_{n=1}^{\infty} a_n$ converges absolutely. Can the same thing be said about $\sum_{n=1}^{\infty} a_n^2$? Explain.

6. A person says a series converges conditionally by the ratio test. Explain why his statement is total nonsense.

7. A person says a series diverges by the alternating series test. Explain why his statement is total nonsense.

8. Find a series which diverges using one test but converges using another if possible. If this is not possible, tell why.

9. If $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ both converge, can you conclude the sum, $\sum_{n=1}^{\infty} a_n b_n$ converges?

10. If $\sum_{n=1}^{\infty} a_n$ converges absolutely, and b_n is bounded, can you conclude $\sum_{n=1}^{\infty} a_n b_n$ converges? What if it is only the case that $\sum_{n=1}^{\infty} a_n$ converges?

11. The logarithm test states the following. Suppose $a_k \neq 0$ for large k and that $p = \lim_{k\to\infty} \frac{\ln\left(\frac{1}{|a_k|}\right)}{\ln k}$ exists. If $p > 1$, then $\sum_{k=1}^{\infty} a_k$ converges absolutely. If $p < 1$, then the series, $\sum_{k=1}^{\infty} a_k$ does not converge absolutely. Prove this theorem.

12. Prove Theorem 11.4.4. **Hint:** For $\sum_{k=1}^{\infty} (-1)^n b_n$, show the odd partial sums are all at least as small as $\sum_{k=1}^{\infty} (-1)^n b_n$ and are increasing while the even partial sums are at least as large as $\sum_{k=1}^{\infty} (-1)^n b_n$ and are decreasing.

13. Use Theorem 11.4.4 in the following alternating series to tell how large n must be so that $\left| \sum_{k=1}^{\infty} (-1)^k a_k - \sum_{k=1}^{n} (-1)^k a_k \right|$ is no larger than the given number.

    (a) $\sum_{k=1}^{\infty} (-1)^k \frac{1}{k}$, .001

    (b) $\sum_{k=1}^{\infty} (-1)^k \frac{1}{k^2}$, .001

    (c) $\sum_{k=1}^{\infty} (-1)^k \sin\left(\frac{1}{k}\right)$, .001

    (d) $\sum_{k=1}^{\infty} (-1)^{k-1} \frac{1}{\sqrt{k}}$, .001

    (e) $\sum_{k=1}^{\infty} (-1)^k \frac{\ln k}{k}$, .001 Note that this one satisfies b_k is decreasing if $k > 3$ but not for all k. Does this matter?

14. For $1 \geq x \geq 0$, and $p \geq 1$, show that $(1-x)^p \geq 1 - px$. **Hint:** This can be done using the mean value theorem from calculus. Define $f(x) \equiv (1-x)^p - 1 + px$ and show that $f(0) = 0$ while $f'(x) \geq 0$ for all $x \in (0,1)$.

15. Using the result of Problem 14 establish Raabe's Test, an interesting variation on the ratio test. This test says the following. Suppose there exists a constant, C and a number p such that
$$\left|\frac{a_{k+1}}{a_k}\right| \leq 1 - \frac{p}{k+C}$$
for all k large enough. Then if $p > 1$, it follows that $\sum_{k=1}^{\infty} a_k$ converges absolutely. **Hint:** Let $b_k \equiv k - 1 + C$ and note that for all k large enough, $b_k > 1$. Now conclude that there exists an integer, k_0 such that $b_{k_0} > 1$ and for all $k \geq k_0$ the given inequality above holds. Use Problem 14 to conclude that
$$\left|\frac{a_{k+1}}{a_k}\right| \leq 1 - \frac{p}{k+C} \leq \left(1 - \frac{1}{k+C}\right)^p = \left(\frac{b_k}{b_{k+1}}\right)^p$$
showing that $|a_k| b_k^p$ is decreasing for $k \geq k_0$. Thus $|a_k| \leq C/b_k^p = C/(k-1+C)^p$. Now use comparison theorems and the p series to obtain the conclusion of the theorem.

16. Consider the series $\sum_{k=0}^{\infty} (-1)^n \frac{1}{\sqrt{n+1}}$. Show this series converges and so it makes sense to write $\left(\sum_{k=0}^{\infty} (-1)^n \frac{1}{\sqrt{n+1}}\right)^2$. What about the Cauchy product of this series? Does it even converge? What does this mean about using algebra on infinite sums as though they were finite sums?

17. Verify Theorem 11.4.17 on the two series $\sum_{k=0}^{\infty} 2^{-k}$ and $\sum_{k=0}^{\infty} 3^{-k}$.

18. You can define infinite series of complex numbers in exactly the same way as infinite series of real numbers. That is $w = \sum_{k=1}^{\infty} z_k$ means: For every $\varepsilon > 0$ there exists N such that if $n \geq N$, then $\left|w - \sum_{k=1}^{N} z_k\right| < \varepsilon$. Here the absolute value is the one which applies to complex numbers. That is, $|a+ib| = \sqrt{a^2+b^2}$. Show that if $\{a_n\}$ is a decreasing sequence of nonnegative numbers with the property that $\lim_{n\to\infty} a_n = 0$ and if ω is any complex number which is not equal to 1 but which satisfies $|\omega| = 1$, then $\sum_{k=1}^{\infty} \omega^n a_n$ must converge. Note a sequence of complex numbers, $\{a_n + ib_n\}$ converges to $a + ib$ if and only if $a_n \to a$ and $b_n \to b$. See Problem 9 on Page 122. There are quite a few things in this problem you should think about.

11.6 Taylor Series

Earlier Taylor polynomials were used to approximate known functions such as $\sin x$ and $\ln(1+x)$. A much more exciting idea is to use infinite series of known functions as definitions of possibly new functions.

Definition 11.6.1 *Let $\{a_k\}_{k=0}^{\infty}$ be a sequence of numbers. The expression,*
$$\sum_{k=0}^{\infty} a_k (x-a)^k \tag{11.9}$$
is called a Taylor series centered at a. This is also called a power series centered at a.

In the above definition, x is a variable. Thus you can put in various values of x and ask whether the resulting series of numbers converges. Defining, D to be the set of all values of x such that the resulting series does converge, define a new function, f defined on D as
$$f(x) \equiv \sum_{k=0}^{\infty} a_k (x-a)^k.$$

This might be a totally new function, one which has no name. Nevertheless, much can be said about such functions. The following lemma is fundamental in considering the form of D which always turns out to be an interval centered at a which may or may not contain either end point.

Lemma 11.6.2 *Suppose $z \in D$. Then if $|x-a| < |z-a|$, then $x \in D$ also and furthermore, the series $\sum_{k=0}^{\infty} |a_k| |x-a|^k$ converges.*

Proof: Let $1 > r = |x-a| / |z-a|$. The n^{th} term test implies

$$\lim_{n\to\infty} |a_n| |z-a|^n = 0$$

and so for all n large enough,

$$|a_n| |z-a|^n < 1$$

so for such n,

$$|a_n| |x-a|^n = |a_n| |z-a|^n \frac{|x-a|^n}{|z-a|^n} \leq \frac{|x-a|^n}{|z-a|^n} < r^n$$

Therefore, $\sum_{k=0}^{\infty} |a_k| |x-a|^k$ converges by comparison with the geometric series, $\sum r^n$.

With this lemma, the following fundamental theorem is obtained.

Theorem 11.6.3 *Let $\sum_{k=0}^{\infty} a_k (x-a)^k$ be a Taylor series. Then there exists $r \leq \infty$ such that the Taylor series converges absolutely if $|x-a| < r$. Furthermore, if $|x-a| > r$, the Taylor series diverges.*

Proof: Let

$$r \equiv \sup \{|y-a| : y \in D\}.$$

Then if $|x-a| < r$, it follows there exists $z \in D$ such that $|z-a| > |x-a|$ since otherwise, r wouldn't be as defined. In fact $|x-a|$ would then be an upper bound to $\{|y-a| : y \in D\}$. Therefore, by the above lemma $\sum_{k=0}^{\infty} |a_k| |x-a|^k$ converges and this proves the first part of this theorem.

Now suppose $|x-a| > r$. If $\sum_{k=0}^{\infty} a_k (x-a)^k$ converges then by the above lemma, r fails to be an upper bound to $\{|y-a| : y \in D\}$ and so the Taylor series must diverge as claimed. This proves the theorem.

From now on D will be referred to as the interval of convergence and r of the above theorem as the radius of convergence. Determining which points of $\{x : |x-a| = r\}$ are in D requires the use of specific convergence tests and can be quite hard. However, the determination of r tends to be pretty easy.

Example 11.6.4 *Find the interval of convergence of the Taylor series $\sum_{n=1}^{\infty} \frac{x^n}{n}$.*

Use Corollary 11.4.8.

$$\lim_{n\to\infty} \left(\frac{|x|^n}{n} \right)^{1/n} = \lim_{n\to\infty} \frac{|x|}{\sqrt[n]{n}} = |x|$$

because $\lim_{n\to\infty} \sqrt[n]{n} = 1$ and so if $|x| < 1$ the series converges. The endpoints require special attention. When $x = 1$ the series diverges because it reduces to $\sum_{n=1}^{\infty} \frac{1}{n}$. At the other endpoint, however, the series converges because it reduces to $\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$ and the alternating series test applies and gives convergence.

Example 11.6.5 *Find the radius of convergence of $\sum_{n=1}^{\infty} \frac{n^n}{n!} x^n$.*

Apply the ratio test. Taking the ratio of the absolute values of the $(n+1)^{th}$ and the n^{th} terms

$$\frac{\frac{(n+1)^{(n+1)}}{(n+1)n!}|x|^{n+1}}{\frac{n^n}{n!}|x|^n} = (n+1)^n |x| n^{-n} = |x| \left(1+\frac{1}{n}\right)^n \to |x| e$$

Therefore the series converges absolutely if $|x| e < 1$ and diverges if $|x| e > 1$. Consequently, $r = 1/e$.

11.6.1 Operations On Power Series

It is desirable to be able to differentiate, integrate, and multiply power series. The following theorem says one can differentiate power series in the most natural way on the interval of convergence, just as you would differentiate a polynomial. This theorem may seem obvious, but it is a serious mistake to think this. You usually cannot differentiate an infinite series whose terms are functions even if the functions are themselves polynomials. The following is special and pertains to power series. It is another example of the interchange of two limits, in this case, the limit involved in taking the derivative and the limit of the sequence of finite sums.

Theorem 11.6.6 *Let $\sum_{n=0}^{\infty} a_n (x-a)^n$ be a Taylor series having radius of convergence $r > 0$ and let*

$$f(x) \equiv \sum_{n=0}^{\infty} a_n (x-a)^n \tag{11.10}$$

for $|x-a| < r$. Then

$$f'(x) = \sum_{n=0}^{\infty} a_n n (x-a)^{n-1} = \sum_{n=1}^{\infty} a_n n (x-a)^{n-1} \tag{11.11}$$

and this new differentiated power series, the derived series, has radius of convergence equal to r.

Proof: First it will be shown that the series on the right in (11.11) has the same radius of convergence as the original series. Thus let $|x-a| < r$ and pick y such that

$$|x-a| < |y-a| < r.$$

Then

$$\lim_{n\to\infty} |a_n| |y-a|^{n-1} = \lim_{n\to\infty} |a_n| |y-a|^n = 0$$

because

$$\sum_{n=0}^{\infty} |a_n| |y-a|^n < \infty$$

and so, for n large enough,

$$|a_n| |y-a|^{n-1} < 1.$$

Therefore, for large enough n,

$$\begin{aligned} |a_n| n |x-a|^{n-1} &= |a_n| |y-a|^{n-1} n \left|\frac{x-a}{y-a}\right|^{n-1} \\ &\leq n \left|\frac{x-a}{y-a}\right|^{n-1} \end{aligned}$$

and

$$\sum_{n=1}^{\infty} n \left| \frac{x-a}{y-a} \right|^{n-1}$$

converges by the ratio test. By the comparison test, it follows $\sum_{n=1}^{\infty} a_n n (x-a)^{n-1}$ converges absolutely for any x satisfying $|x-a| < r$. Therefore, the radius of convergence of the derived series is at least as large as that of the original series. On the other hand, if $\sum_{n=1}^{\infty} |a_n| n |x-a|^{n-1}$ converges then by the comparison test, $\sum_{n=1}^{\infty} |a_n| |x-a|^{n-1}$ and therefore $\sum_{n=1}^{\infty} |a_n| |x-a|^n$ also converges which shows the radius of convergence of the derived series is no larger than that of the original series. It remains to verify the assertion about the derivative.

Let $|x-a| < r$ and let $r_1 < r$ be close enough to r that

$$x \in (a-r_1, a+r_1) \subseteq [a-r_1, a+r_1] \subseteq (a-r, a+r).$$

Thus, letting $r_2 \in (r_1, r)$,

$$\sum_{n=0}^{\infty} |a_n| r_1^n, \sum_{n=0}^{\infty} |a_n| r_2^n < \infty \tag{11.12}$$

Letting y be close enough to x, it follows both x and y are in $[a-r_1, a+r_1]$. Then considering the difference quotient,

$$\begin{aligned} \frac{f(y)-f(x)}{y-x} &= \sum_{n=0}^{\infty} a_n (y-x)^{-1} \left[(y-a)^n - (x-a)^n \right] \\ &= \sum_{n=1}^{\infty} a_n n z_n^{n-1} \end{aligned} \tag{11.13}$$

where the last equation follows from the mean value theorem and z_n is some point between $x-a$ and $y-a$. Therefore,

$$\frac{f(y)-f(x)}{y-x} = \sum_{n=1}^{\infty} a_n n z_n^{n-1} =$$

$$\begin{aligned} &= \sum_{n=1}^{\infty} a_n n \left(z_n^{n-1} - (x-a)^{n-1} \right) + \sum_{n=1}^{\infty} a_n n (x-a)^{n-1} \\ &= \sum_{n=2}^{\infty} a_n n (n-1) w_n^{n-2} (z_n - (x-a)) + \sum_{n=1}^{\infty} a_n n (x-a)^{n-1} \end{aligned} \tag{11.14}$$

where w_n is between z_n and $x-a$. Thus w_n is between $x-a$ and $y-a$ and so

$$w_n + a \in [a-r_1, a+r_1]$$

which implies $|w_n| \leq r_1$. The first sum on the right in (11.14) therefore satisfies

$$\begin{aligned} \left| \sum_{n=2}^{\infty} a_n n (n-1) w_n^{n-2} (z_n - (x-a)) \right| &\leq |y-x| \sum_{n=2}^{\infty} |a_n| n (n-1) |w_n|^{n-2} \\ &\leq |y-x| \sum_{n=2}^{\infty} |a_n| n (n-1) r_1^{n-2} \end{aligned}$$

$$= |y-x| \sum_{n=2}^{\infty} |a_n| \, r_2^{n-2} n (n-1) \left(\frac{r_1}{r_2} \right)^{n-2}$$

Now from (11.12), $|a_n| \, r_2^{n-2} < 1$ for all n large enough. Therefore, for such n,

$$|a_n| \, r_2^{n-2} n (n-1) \left(\frac{r_1}{r_2} \right)^{n-2} \leq n (n-1) \left(\frac{r_1}{r_2} \right)^{n-2}$$

and the series $\sum n (n-1) \left(\frac{r_1}{r_2} \right)^{n-2}$ converges by the ratio test. Therefore, there exists a constant, C independent of y such that

$$\sum_{n=2}^{\infty} |a_n| \, n (n-1) \, r_1^{n-2} = C < \infty$$

Consequently, from (11.14)

$$\left| \frac{f(y) - f(x)}{y-x} - \sum_{n=1}^{\infty} a_n n (x-a)^{n-1} \right| \leq C |y-x| .$$

Taking the limit as $y \to x$ (11.11) follows. This proves the theorem.

As an immediate corollary, it is possible to characterize the coefficients of a Taylor series.

Corollary 11.6.7 *Let $\sum_{n=0}^{\infty} a_n (x-a)^n$ be a Taylor series with radius of convergence $r > 0$ and let*

$$f(x) \equiv \sum_{n=0}^{\infty} a_n (x-a)^n . \tag{11.15}$$

Then

$$a_n = \frac{f^{(n)}(a)}{n!}. \tag{11.16}$$

Proof: From (11.15), $f(a) = a_0 \equiv f^{(0)}(a)/0!$. From Theorem 11.6.6,

$$f'(x) = \sum_{n=1}^{\infty} a_n n (x-a)^{n-1} = a_1 + \sum_{n=2}^{\infty} a_n n (x-a)^{n-1} .$$

Now let $x = a$ and obtain that $f'(a) = a_1 = f'(a)/1!$. Next use Theorem 11.6.6 again to take the second derivative and obtain

$$f''(x) = 2a_2 + \sum_{n=3}^{\infty} a_n n (n-1) (x-a)^{n-2}$$

let $x = a$ in this equation and obtain $a_2 = f''(a)/2 = f''(a)/2!$. Continuing this way proves the corollary.

This also shows the coefficients of a Taylor series are unique. That is, if

$$\sum_{k=0}^{\infty} a_k (x-a)^k = \sum_{k=0}^{\infty} b_k (x-a)^k$$

for all x in some interval, then $a_k = b_k$ for all k.

Example 11.6.8 *Find the power series for $\sin(x)$, and $\cos(x)$ centered at 0 and give the interval of convergence.*

First consider $f(x) = \sin(x)$. Then $f'(x) = \cos(x), f''(x) = -\sin(x), f'''(x) = -\cos(x)$ etc. Therefore, from Taylor's formula, Theorem 11.1.1 on Page 299,

$$f(x) = 0 + x + 0 - \frac{x^3}{3!} + 0 + \frac{x^5}{5!} + \cdots + \frac{x^{2n+1}}{(2n+1)!} + \frac{f^{(2n+2)}(\xi_n)}{(2n+2)!}$$

where ξ_n is some number between 0 and x. Furthermore, this equals either $\pm\sin(\xi_n)$ or $\pm\cos(\xi_n)$ and so its absolute value is no larger than 1. Thus

$$\left|\frac{f^{(2n+2)}(\xi_n)}{(2n+2)!}\right| \leq \frac{1}{(2n+2)!}.$$

By the ratio test, it follows that

$$\sum_{n=0}^{\infty} \frac{1}{(2n+2)!} < \infty$$

and so by the comparison test,

$$\sum_{n=0}^{\infty} \left|\frac{f^{(2n+2)}(\xi_n)}{(2n+2)!}\right| < \infty$$

also. Therefore, by the n^{th} term test $\lim_{n\to\infty} \frac{f^{(2n+2)}(\xi_n)}{(2n+2)!} = 0$. This implies

$$\sin(x) = \sum_{k=0}^{n} (-1)^k \frac{x^{2k+1}}{(2k+1)!} + \frac{f^{(2n+2)}(\xi_n)}{(2n+2)!}$$

and the last term converges to zero as $n \to \infty$ for any value of x and therefore,

$$\sin(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}$$

for all $x \in \mathbb{R}$. By Theorem 11.6.6, you can differentiate both sides, doing the series term by term and obtain

$$\cos(x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}$$

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

Example 11.6.9 *Find the sum* $\sum_{k=1}^{\infty} k2^{-k}$.

It may not be obvious what this sum equals but with the above theorem it is easy to find. From the formula for the sum of a geometric series, $\frac{1}{1-t} = \sum_{k=0}^{\infty} t^k$ if $|t| < 1$. Differentiate both sides to obtain

$$(1-t)^{-2} = \sum_{k=1}^{\infty} kt^{k-1}$$

whenever $|t| < 1$. Let $t = 1/2$. Then

$$4 = \frac{1}{(1-(1/2))^2} = \sum_{k=1}^{\infty} k2^{-(k-1)}$$

and so if you multiply both sides by 2^{-1},

$$2 = \sum_{k=1}^{\infty} k2^{-k}.$$

The following is a very important example known as the binomial series.

Example 11.6.10 *Find a Taylor series for the function* $(1+x)^{\alpha}$ *centered at 0 valid for* $|x|<1$.

Use Theorem 11.6.6 to do this. First note that if $y(x) \equiv (1+x)^{\alpha}$, then y is a solution of the following initial value problem.

$$y' - \frac{\alpha}{(1+x)} y = 0,\ y(0) = 1. \tag{11.17}$$

Next it is necessary to observe there is only one solution to this initial value problem. To see this, multiply both sides of the differential equation in (11.17) by $(1+x)^{-\alpha}$. When this is done one obtains

$$\frac{d}{dx}\left((1+x)^{-\alpha} y\right) = (1+x)^{-\alpha}\left(y' - \frac{\alpha}{(1+x)} y\right) = 0. \tag{11.18}$$

Therefore, from (11.18), there must exist a constant, C, such that

$$(1+x)^{-\alpha} y = C.$$

However, $y(0) = 1$ and so it must be that $C = 1$. Therefore, there is exactly one solution to the initial value problem in (11.17) and it is $y(x) = (1+x)^{\alpha}$. The strategy for finding the Taylor series of this function consists of finding a series which solves the initial value problem above. Let

$$y(x) \equiv \sum_{n=0}^{\infty} a_n x^n \tag{11.19}$$

be a solution to (11.17). Of course it is not known at this time whether such a series exists. However, the process of finding it will demonstrate its existence. From Theorem 11.6.6 and the initial value problem,

$$(1+x)\sum_{n=0}^{\infty} a_n n x^{n-1} - \sum_{n=0}^{\infty} \alpha a_n x^n = 0$$

and so

$$\sum_{n=1}^{\infty} a_n n x^{n-1} + \sum_{n=0}^{\infty} a_n (n-\alpha) x^n = 0$$

Changing the order variable of summation in the first sum,

$$\sum_{n=0}^{\infty} a_{n+1}(n+1) x^n + \sum_{n=0}^{\infty} a_n (n-\alpha) x^n = 0$$

and from Corollary 11.6.7 and the initial condition for (11.17) this requires

$$a_{n+1} = \frac{a_n (\alpha - n)}{n+1}, a_0 = 1. \tag{11.20}$$

Therefore, from (11.20) and letting $n = 0$, $a_1 = \alpha$. Then using (11.20) again along with this information, $a_2 = \frac{\alpha(\alpha-1)}{2}$. Using the same process, $a_3 = \frac{\left(\frac{\alpha(\alpha-1)}{2}\right)(\alpha-2)}{3} = \frac{\alpha(\alpha-1)(\alpha-2)}{3!}$. By now you can spot the pattern. In general,

$$a_n = \frac{\overbrace{\alpha(\alpha-1)\cdots(\alpha-n+1)}^{n \text{ of these factors}}}{n!}.$$

Therefore, our candidate for the Taylor series is

$$y(x) = \sum_{n=0}^{\infty} \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!} x^n.$$

Furthermore, the above discussion shows this series solves the initial value problem on its interval of convergence. It only remains to show the radius of convergence of this series equals 1. It will then follow that this series equals $(1+x)^\alpha$ because of uniqueness of the initial value problem. To find the radius of convergence, use the ratio test. Thus the ratio of the absolute values of $(n+1)^{st}$ term to the absolute value of the n^{th} term is

$$\frac{\left|\frac{\alpha(\alpha-1)\cdots(\alpha-n+1)(\alpha-n)}{(n+1)n!}\right| |x|^{n+1}}{\left|\frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}\right| |x|^n} = |x| \frac{|\alpha-n|}{n+1} \to |x|$$

showing that the radius of convergence is 1 since the series converges if $|x|<1$ and diverges if $|x|>1$.

The expression, $\frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}$ is often denoted as $\binom{\alpha}{n}$. With this notation, the following theorem has been established.

Theorem 11.6.11 *Let α be a real number and let $|x|<1$. Then*

$$(1+x)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n.$$

There is a very interesting issue related to the above theorem which illustrates the limitation of power series. The function $f(x) = (1+x)^\alpha$ makes sense for all $x > -1$ but one is only able to describe it with a power series on the interval $(-1,1)$. Think about this. The above technique is a standard one for obtaining solutions of differential equations and this example illustrates a deficiency in the method. To completely understand power series, it is necessary to take a course in complex analysis. You may have noticed the prominent role played by geometric series. This is no accident. It turns out that the right way to consider Taylor series is through the use of geometric series and something called the Cauchy integral formula of complex analysis. However, these are topics for another course.

You can also integrate power series on their interval of convergence.

Theorem 11.6.12 *Let $f(x) = \sum_{n=0}^{\infty} a_n (x-a)^n$ and suppose the interval of convergence is $r>0$. Then if $|y-a|<r$,*

$$\int_a^y f(x)\,dx = \sum_{n=0}^{\infty} \int_a^y a_n (x-a)^n\,dx = \sum_{n=0}^{\infty} \frac{a_n (y-a)^{n+1}}{n+1}.$$

Proof: Define $F(y) \equiv \int_a^y f(x)\,dx$ and $G(y) \equiv \sum_{n=0}^{\infty} \frac{a_n (y-a)^{n+1}}{n+1}$. By Theorem 11.6.6 and the Fundamental theorem of calculus,

$$G'(y) = \sum_{n=0}^{\infty} a_n (y-a)^n = f(y) = F'(y).$$

Therefore, $G(y) - F(y) = C$ for some constant. But $C=0$ because $F(a) - G(a) = 0$. This proves the theorem.

Next consider the problem of multiplying two power series.

Theorem 11.6.13 *Let $\sum_{n=0}^{\infty} a_n (x-a)^n$ and $\sum_{n=0}^{\infty} b_n (x-a)^n$ be two power series having radii of convergence r_1 and r_2, both positive. Then*

$$\left(\sum_{n=0}^{\infty} a_n (x-a)^n\right)\left(\sum_{n=0}^{\infty} b_n (x-a)^n\right) = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} a_k b_{n-k}\right)(x-a)^n$$

whenever $|x-a| < r \equiv \min(r_1, r_2)$.

Proof: By Theorem 11.6.3 both series converge absolutely if $|x-a| < r$. Therefore, by Theorem 11.4.17

$$\left(\sum_{n=0}^{\infty} a_n (x-a)^n\right)\left(\sum_{n=0}^{\infty} b_n (x-a)^n\right) =$$

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n} a_k (x-a)^k b_{n-k} (x-a)^{n-k} = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} a_k b_{n-k}\right)(x-a)^n.$$

This proves the theorem.

The significance of this theorem in terms of applications is that it states you can multiply power series just as you would multiply polynomials and everything will be all right on the common interval of convergence.

This theorem can be used to find Taylor series which would perhaps be hard to find without it. Here is an example.

Example 11.6.14 *Find the Taylor series for $e^x \sin x$ centered at $x = 0$.*

Using Problems 11 - 13 on Page 302 or Example 11.6.8 on Page 325, and Example 11.4.9 on Page 313, all that is required is to multiply

$$\left(\overbrace{1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}\cdots}^{e^x}\right)\left(\overbrace{x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots}^{\sin x}\right)$$

From the above theorem the result should be

$$x + x^2 + \left(-\frac{1}{3!} + \frac{1}{2!}\right)x^3 + \cdots$$

$$= x + x^2 + \frac{1}{3}x^3 + \cdots$$

You can continue this way and get the following to a few more terms.

$$x + x^2 + \frac{1}{3}x^3 - \frac{1}{30}x^5 - \frac{1}{90}x^6 - \frac{1}{630}x^7 + \cdots$$

I don't see a pattern in these coefficients but I can go on generating them as long as I want. (In practice this tends to not be very long.) I also know the resulting power series will converge for all x because both the series for e^x and the one for $\sin x$ converge for all x.

Example 11.6.15 *Find the Taylor series for $\tan x$ centered at $x = 0$.*

Lets suppose it has a Taylor series $a_0 + a_1x + a_2x^2 + \cdots$. Then

$$\left(a_0 + a_1x + a_2x^2 + \cdots\right)\left(\overbrace{1 - \frac{x^2}{2} + \frac{x^4}{4!} + \cdots}^{\cos x}\right) = \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots\right).$$

Using the above, $a_0 = 0, a_1x = x$ so $a_1 = 1, \left(0\left(\frac{-1}{2}\right) + a_2\right)x^2 = 0$ so $a_2 = 0$. $\left(a_3 - \frac{a_1}{2}\right)x^3 = \frac{-1}{3!}x^3$ so $a_3 - \frac{1}{2} = -\frac{1}{6}$ so $a_3 = \frac{1}{3}$. Clearly one can continue in this manner. Thus the first several terms of the power series for tan are

$$\tan x = x + \frac{1}{3}x^3 + \cdots.$$

You can go on calculating these terms and find the next two yielding

$$\tan x = x + \frac{1}{3}x^3 + \frac{2}{15}x^5 + \frac{17}{315}x^7 + \cdots$$

This is a very significant technique because, as you see, there does not appear to be a very simple pattern for the coefficients of the power series for $\tan x$. Of course there are some issues here about whether $\tan x$ even has a power series, but if it does, the above must be it. In fact, $\tan(x)$ will have a power series valid on some interval centered at 0 and this becomes completely obvious when one uses methods from complex analysis but it isn't too obvious at this point. If you are interested in this issue, read the last section of the chapter. Note also that what has been accomplished is to divide the power series for $\sin x$ by the power series for $\cos x$ just like they were polynomials.

11.7 Exercises

1. Find the radius of convergence of the following.
   (a) $\sum_{k=1}^{\infty}\left(\frac{x}{2}\right)^n$
   (b) $\sum_{k=1}^{\infty}\sin\left(\frac{1}{n}\right)3^n x^n$
   (c) $\sum_{k=0}^{\infty}k!x^k$
   (d) $\sum_{n=0}^{\infty}\frac{(3n)^n}{(3n)!}x^n$
   (e) $\sum_{n=0}^{\infty}\frac{(2n)^n}{(2n)!}x^n$
2. Find $\sum_{k=1}^{\infty}k2^{-k}$.
3. Find $\sum_{k=1}^{\infty}k^2 3^{-k}$.
4. Find $\sum_{k=1}^{\infty}\frac{2^{-k}}{k}$.
5. Find $\sum_{k=1}^{\infty}\frac{3^{-k}}{k}$.
6. Find the power series centered at 0 for the function $1/\left(1+x^2\right)$ and give the radius of convergence.
7. Use the power series technique which was applied in Example 11.6.10 to consider the initial value problem $y' = y, y(0) = 1$. This yields another way to obtain the power series for e^x.

8. Use the power series technique on the initial value problem $y' + y = 0,\ y(0) = 1$. What is the solution to this initial value problem?

9. Use the power series technique to find solutions in terms of power series to the initial value problem
$$y'' + xy = 0,\ y(0) = 0, y'(0) = 1.$$
Tell where your solution gives a valid description of a solution for the initial value problem. **Hint:** This is a little different but you proceed the same way as in Example 11.6.10. The main difference is you have to do two differentiations of the power series instead of one.

10. Suppose the function, e^x is defined in terms of a power series, $e^x \equiv \sum_{k=0}^{\infty} \frac{x^k}{k!}$. Use Theorem 11.4.17 on Page 318 to show directly the usual law of exponents,
$$e^{x+y} = e^x e^y.$$
Be sure to check all the hypotheses.

11. Define the following function[2]:
$$f(x) \equiv \begin{cases} e^{-(1/x^2)} \text{ if } x \neq 0 \\ 0 \text{ if } x = 0 \end{cases}.$$
Show that $f^{(k)}(x)$ exists for all k and for all x. Show also that $f^{(k)}(0) = 0$ for all $k \in \mathbb{N}$. Therefore, the power series for $f(x)$ is of the form $\sum_{k=0}^{\infty} 0x^k$ and it converges for all values of x. However, it fails to converge to $f(x)$ except at the single point, $x = 0$.

12. Let $f_n(x) \equiv \left(\frac{1}{n} + x^2\right)^{1/2}$. Show that for all x,
$$||x| - f_n(x)| \leq \frac{1}{\sqrt{n}}.$$
Now show $f_n'(0) = 0$ for all n and so $f_n'(0) \to 0$. However, the function, $f(x) \equiv |x|$ has no derivative at $x = 0$. Thus even though $f_n(x) \to f(x)$ for all x, you cannot say that $f_n'(0) \to f'(0)$.

13. Let the functions, $f_n(x)$ be given in Problem 12 and consider
$$g_1(x) = f_1(x),\ g_n(x) = f_n(x) - f_{n-1}(x) \text{ if } n > 1.$$
Show that for all x,
$$\sum_{k=0}^{\infty} g_k(x) = |x|$$
and that $g_k'(0) = 0$ for all k. Therefore, you can't differentiate the series term by term and get the right answer[3].

[2]Surprisingly, this function is very important to those who use modern techniques to study differential equations. One needs to consider test functions which have the property they have infinitely many derivatives but vanish outside of some interval. The theory of complex variables can be used to show there are no examples of such functions if they have a valid power series expansion. It even becomes a little questionable whether such strange functions even exist at all. Nevertheless, they do, there are enough of them, and it is this very example which is used to show this.

[3]How bad can this get? It can be much worse than this. In fact, there are functions which are continuous everywhere and differentiable nowhere. We typically don't have names for them but they are there just the same. Every such function can be written as an infinite sum of polynomials which of course have derivatives at every point. Thus it is nonsense to differentiate an infinite sum term by term without a theorem of some sort.

14. Use the theorem about the binomial series to give a proof of the binomial theorem

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

whenever n is a positive integer.

15. You know $\int_0^x \frac{1}{t+1}\,dt = \ln|1+x|$. Use this and Theorem 11.6.12 to find the power series for $\ln|1+x|$ centered at 0. Where does this power series converge? Where does it converge to the function, $\ln|1+x|$?

16. You know $\int_0^x \frac{1}{t^2+1}\,dt = \arctan x$. Use this and Theorem 11.6.12 to find the power series for $\arctan x$ centered at 0. Where does this power series converge? Where does it converge to the function, $\arctan x$?

17. Find the power series for $\sin\left(x^2\right)$ by plugging in x^2 where ever there is an x in the power series for $\sin x$. How do you know this is the power series for $\sin\left(x^2\right)$?

18. Find the first several terms of the power series for $\sin^2(x)$ by multiplying the power series for $\sin(x)$. Next use the trig. identity, $\sin^2(x) = \frac{1-\cos(2x)}{2}$ and the power series for $\cos(2x)$ to find the power series.

19. Find the power series for $f(x) = \frac{1}{\sqrt{1-x^2}}$.

20. It is hard to find $\int_0^1 e^{x^2}\,dx$ because you don't have a convenient antiderivative for the integrand. Replace e^{x^2} with an appropriate power series and estimate this integral.

21. Do the same as the previous problem for $\int_0^1 \sin\left(x^2\right)\,dx$.

22. Find $\lim_{x\to 0} \frac{\tan(\sin x)-\sin(\tan x)}{x^7}$.[4]

23. Consider the function, $S(x) \equiv \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^{2n-1}}{(2n-1)!}$. This is the power series for $\sin(x)$ but pretend you don't know this. Show that the series for $S(x)$ converges for all $x \in \mathbb{R}$. Also show that S satisfies the initial value problem $y'' + y = 0, y(0) = 0, y'(0) = 1$.

24. Consider the function, $C(x) \equiv \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}$. This is the power series for $\cos(x)$ but pretend you don't know this. Show that the series for $C(x)$ converges for all $x \in \mathbb{R}$. Also show that S satisfies the initial value problem $y'' + y = 0, y(0) = 1, y'(0) = 0$.

25. Show there is at most one solution to the initial value problem, $y'' + y = 0, y(0) = a, y'(0) = b$ and find the solution to this problem in terms of $C(x)$ and $S(x)$. Also show directly from the series descriptions for $C(x)$ and $S(x)$ that $S'(x) = C(x)$ and $C'(x) = -S(x)$.

26. Using problem 25 about uniqueness of the initial value problem, show that $C(x+y) = C(x)C(y) - S(x)S(y)$ and $S(x+y) = S(x)C(y) + S(y)C(x)$. Do this in the following way: Fix y and consider the function $f(x) \equiv C(x+y)$ and $g(x) = C(x)C(y) - S(x)S(y)$. Then show both f and g satisfy the same initial value problem and so they must be equal. Do the other identity the same way. Also show $S(-x) = -S(x)$ and $C(-x) = C(x)$ and $S(x)^2 + C(x)^2 = 1$. This last claim is really easy. Just take the derivative and see $S^2 + C^2$ must be constant.

[4] This is a wonderful example. You should plug in small values of x using a calculator and see what you get using modern technology.

27. You know $S(0) = 0$ and $C(0) = 1$. Show there exists $T > 0$ such that on $(0, T)$ both $S(x)$ and $C(x)$ are positive but $C(T) = 0$ while $S(T) = 1$. (We usually refer to T as $\frac{\pi}{2}$.) To do this, note that $S'(0) > 0$ and so S is an increasing function on some interval. Therefore, C is a decreasing function on that interval because of $S^2 + C^2 = 1$. If C is bounded below by some positive number, then S must be unbounded because $S' = C$. However this would contradict $S^2 + C^2 = 1$. Therefore, $C(T) = 0$ for some T. Let T be the first time this occurs. You fill in the mathematical details of this argument. Next show that on $(T, 2T)$, $S(x) > 0$ and $C(x) < 0$ and on $(2T, 3T)$, both $C(x)$ and $S(x)$ are negative. Finally, show that on $(3T, 4T)$, $C(x) > 0$ and $S(x) < 0$. Also show $C(x + 2T) = C(x)$ and $S(x + 2T) = S(x)$. Do all this without resorting to identifying $S(x)$ with $\sin x$ and $C(x)$ with $\cos x$. Finally explain why $\sin x = S(x)$ for all x and $C(x) = \cos x$ for all x.

Note: Problems 23 - 27 outline a way to define the circular functions with no reference to plane geometry. The job is not finished because these circular functions were defined as the x and y coordinates of a point on the unit circle where the angle was measured in terms of arc length and I have not yet tied it in to arc length. This is very easy to do later. Taking the pythagorean theorem as the **definition** of length, a precise description of what is meant by arc length in terms of integrals can be presented. When this is done, it is possible to **define** $\sin x$ and $\cos x$ in terms of these power series described above and totally eliminate all references to plane geometry. This approach is vastly superior to the traditional approach presented earlier in this book. Calculus is different than geometry and so it is desirable to obtain descriptions of the important functions which are free of geometry.

11.8 Some Other Theorems

First recall Theorem 11.4.17 on Page 318. For convenience, the version of this theorem which is of interest here is listed below.

Theorem 11.8.1 *Suppose $\sum_{i=0}^{\infty} a_i$ and $\sum_{j=0}^{\infty} b_j$ both converge absolutely. Then*

$$\left(\sum_{i=0}^{\infty} a_i\right)\left(\sum_{j=0}^{\infty} b_j\right) = \sum_{n=0}^{\infty} c_n$$

where

$$c_n = \sum_{k=0}^{n} a_k b_{n-k}.$$

Furthermore, $\sum_{n=0}^{\infty} c_n$ converges absolutely.

Proof: It only remains to verify the last series converges absolutely. By Theorem 11.4.14 on Page 316 and letting p_{nk} be as defined there,

$$\begin{aligned}\sum_{n=0}^{\infty}|c_n| &= \sum_{n=0}^{\infty}\left|\sum_{k=0}^{n}a_k b_{n-k}\right| \\ &\leq \sum_{n=0}^{\infty}\sum_{k=0}^{n}|a_k|\,|b_{n-k}| = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty}p_{nk}\,|a_k|\,|b_{n-k}| \\ &= \sum_{k=0}^{\infty}\sum_{n=0}^{\infty}p_{nk}\,|a_k|\,|b_{n-k}| = \sum_{k=0}^{\infty}\sum_{n=k}^{\infty}|a_k|\,|b_{n-k}| \\ &= \sum_{k=0}^{\infty}|a_k|\sum_{n=0}^{\infty}|b_n| < \infty.\end{aligned}$$

This proves the theorem.

The theorem is about multiplying two series. What if you wanted to consider

$$\left(\sum_{n=0}^{\infty}a_n\right)^p$$

where p is a positive integer maybe larger than 2? Is there a similar theorem to the above?

Definition 11.8.2 *Define*

$$\sum_{k_1+\cdots+k_p=m}a_{k_1}a_{k_2}\cdots a_{k_p}$$

as follows. Consider all ordered lists of nonnegative integers $k_1,\cdots,k_p$ which have the property that $\sum_{i=1}^{p}k_i=m$. For each such list of integers, form the product, $a_{k_1}a_{k_2}\cdots a_{k_p}$ and then add all these products.

Note that

$$\sum_{k=0}^{n}a_k a_{n-k} = \sum_{k_1+k_2=n}a_{k_1}a_{k_2}$$

Therefore, from the above theorem, if $\sum a_i$ converges absolutely, it follows

$$\left(\sum_{i=0}^{\infty}a_i\right)^2 = \sum_{n=0}^{\infty}\left(\sum_{k_1+k_2=n}a_{k_1}a_{k_2}\right).$$

It turns out a similar theorem holds for replacing 2 with p.

Theorem 11.8.3 *Suppose $\sum_{n=0}^{\infty}a_n$ converges absolutely. Then*

$$\left(\sum_{n=0}^{\infty}a_n\right)^p = \sum_{m=0}^{\infty}c_{mp}$$

where

$$c_{mp}\equiv\sum_{k_1+\cdots+k_p=m}a_{k_1}\cdots a_{k_p}.$$

Proof: First note this is obviously true if $p = 1$ and is also true if $p = 2$ from the above theorem. Now suppose this is true for p and consider $\left(\sum_{n=0}^{\infty} a_n\right)^{p+1}$. By the induction hypothesis and the above theorem on the Cauchy product,

$$\begin{aligned}
\left(\sum_{n=0}^{\infty} a_n\right)^{p+1} &= \left(\sum_{n=0}^{\infty} a_n\right)^{p} \left(\sum_{n=0}^{\infty} a_n\right) \\
&= \left(\sum_{m=0}^{\infty} c_{mp}\right) \left(\sum_{n=0}^{\infty} a_n\right) \\
&= \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} c_{kp} a_{n-k}\right) \\
&= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \sum_{k_1+\cdots+k_p=k} a_{k_1} \cdots a_{k_p} a_{n-k} \\
&= \sum_{n=0}^{\infty} \sum_{k_1+\cdots+k_{p+1}=n} a_{k_1} \cdots a_{k_{p+1}}
\end{aligned}$$

and this proves the theorem.

This theorem implies the following corollary for power series.

Corollary 11.8.4 *Let*

$$\sum_{n=0}^{\infty} a_n (x-a)^n$$

be a power series having radius of convergence, $r > 0$. Then if $|x-a| < r$,

$$\left(\sum_{n=0}^{\infty} a_n (x-a)^n\right)^{p} = \sum_{n=0}^{\infty} b_{np} (x-a)^n$$

where

$$b_{np} \equiv \sum_{k_1+\cdots+k_p=n} a_{k_1} \cdots a_{k_p}.$$

Proof: Since $|x-a| < r$, the series, $\sum_{n=0}^{\infty} a_n (x-a)^n$, converges absolutely. Therefore, the above theorem applies and

$$\left(\sum_{n=0}^{\infty} a_n (x-a)^n\right)^{p} =$$

$$\sum_{n=0}^{\infty} \left(\sum_{k_1+\cdots+k_p=n} a_{k_1} (x-a)^{k_1} \cdots a_{k_p} (x-a)^{k_p}\right) =$$

$$\sum_{n=0}^{\infty} \left(\sum_{k_1+\cdots+k_p=n} a_{k_1} \cdots a_{k_p}\right) (x-a)^n.$$

With this theorem it is possible to consider the question raised in Example 11.6.15 on Page 329 about the existence of the power series for $\tan x$. This question is clearly included in the more general question of when

$$\left(\sum_{n=0}^{\infty} a_n (x-a)^n\right)^{-1}$$

has a power series.

Lemma 11.8.5 *Let $f(x)=\sum_{n=0}^{\infty}a_n(x-a)^n$, a power series having radius of convergence $r>0$. Suppose also that $f(a)=1$. Then there exists $r_1>0$ and $\{b_n\}$ such that for all $|x-a|<r_1$,*

$$\frac{1}{f(x)}=\sum_{n=0}^{\infty}b_n(x-a)^n.$$

Proof: By continuity, there exists $r_1>0$ such that if $|x-a|<r_1$, then

$$\sum_{n=1}^{\infty}|a_n||x-a|^n<1.$$

Now pick such an x. Then

$$\begin{aligned}\frac{1}{f(x)} &= \frac{1}{1+\sum_{n=1}^{\infty}a_n(x-a)^n}\\ &= \frac{1}{1+\sum_{n=0}^{\infty}c_n(x-a)^n}\end{aligned}$$

where $c_n=a_n$ if $n>0$ and $c_0=0$. Then

$$\left|\sum_{n=1}^{\infty}a_n(x-a)^n\right|\le\sum_{n=1}^{\infty}|a_n||x-a|^n<1 \tag{11.21}$$

and so from the formula for the sum of a geometric series,

$$\frac{1}{f(x)}=\sum_{p=0}^{\infty}\left(\sum_{n=0}^{\infty}c_n(x-a)^n\right)^p.$$

By Corollary 11.8.4, this equals

$$\sum_{p=0}^{\infty}\sum_{n=0}^{\infty}b_{np}(x-a)^n \tag{11.22}$$

where

$$b_{np}=\sum_{k_1+\cdots+k_p=n}c_{k_1}\cdots c_{k_p}.$$

Thus $|b_{np}|\le\sum_{k_1+\cdots+k_p=n}|c_{k_1}|\cdots|c_{k_p}|\equiv B_{np}$ and so by Theorem 11.8.3,

$$\begin{aligned}\sum_{p=0}^{\infty}\sum_{n=0}^{\infty}|b_{np}||x-a|^n &\le \sum_{p=0}^{\infty}\sum_{n=0}^{\infty}B_{np}|x-a|^n\\ &= \sum_{p=0}^{\infty}\left(\sum_{n=0}^{\infty}|c_n||x-a|^n\right)^p<\infty\end{aligned}$$

by (11.21) and the formula for the sum of a geometric series. Since the series of (11.22) converges absolutely, Theorem 11.4.14 on Page 316 implies the series in (11.22) equals

$$\sum_{n=0}^{\infty}\left(\sum_{p=0}^{\infty}b_{np}\right)(x-a)^n$$

and so, letting $\sum_{p=0}^{\infty}b_{np}\equiv b_n$, this proves the lemma.

With this lemma, the following theorem is easy to obtain.

Theorem 11.8.6 *Let $f(x) = \sum_{n=0}^{\infty} a_n (x-a)^n$, a power series having radius of convergence $r > 0$. Suppose also that $f(a) \neq 0$. Then there exists $r_1 > 0$ and $\{b_n\}$ such that for all $|x-a| < r_1$,*

$$\frac{1}{f(x)} = \sum_{n=0}^{\infty} b_n (x-a)^n .$$

Proof: Let $g(x) \equiv f(x)/f(a)$ so that $g(x)$ satisfies the conditions of the above lemma. Then by that lemma, there exists $r_1 > 0$ and a sequence, $\{b_n\}$ such that

$$\frac{f(a)}{f(x)} = \sum_{n=0}^{\infty} b_n (x-a)^n$$

for all $|x-a| < r_1$. Then

$$\frac{1}{f(x)} = \sum_{n=0}^{\infty} \widetilde{b_n} (x-a)^n$$

where $\widetilde{b_n} = b_n/f(a)$. This proves the theorem.

There is a very interesting question related to r_1 in this theorem. One might think that if $|x-a| < r$, the radius of convergence of $f(x)$ and if $f(x) \neq 0$ it should be possible to write $1/f(x)$ as a power series centered at a. Unfortunately this is not true. Consider $f(x) = 1 + x^2$. In this case $r = \infty$ but the power series for $1/f(x)$ converges only if $|x| < 1$. What happens is this, $1/f(x)$ will have a power series that will converge for $|x-a| < r_1$ where r_1 is the distance between a and the nearest singularity or zero of $f(x)$ in the complex plane. In the case of $f(x) = 1 + x^2$ this function has a zero at $x = \pm i$. This is just another instance of why the natural setting for the study of power series is the complex plane. To read more on power series, you should see the book by Apostol [3] or any text on complex variable.

Part III

Basic Linear Algebra

Fundamentals

12.0.1 Outcomes

1. Describe $\mathbb{R}^n$ and do algebra with vectors in $\mathbb{R}^n$.
2. Represent a line in 3 space by a vector parameterization, a set of scalar parametric equations or using symmetric form.
3. Find a parameterization of a line given information about
   (a) a point of the line and the direction of the line
   (b) two points contained in the line
4. Determine the direction of a line given its parameterization.

12.1 $\mathbb{R}^n$

The notation, $\mathbb{R}^n$ refers to the collection of ordered lists of n real numbers. More precisely, consider the following definition.

Definition 12.1.1 *Define*

$$\mathbb{R}^n \equiv \{(x_1, \cdots, x_n) : x_j \in \mathbb{R} \text{ for } j = 1, \cdots, n\}.$$

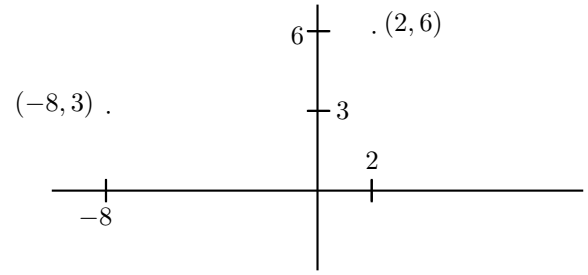
$(x_1, \cdots, x_n) = (y_1, \cdots, y_n)$ *if and only if for all* $j = 1, \cdots, n$, $x_j = y_j$. *When* $(x_1, \cdots, x_n) \in \mathbb{R}^n$, *it is conventional to denote* $(x_1, \cdots, x_n)$ *by the single bold face letter,* $\mathbf{x}$. *The numbers,* x_j *are called the* ***coordinates***. *The set*

$$\{(0, \cdots, 0, t, 0, \cdots, 0) : t \in \mathbb{R}\}$$

for t *in the* i^{th} *slot is called the* i^{th} *coordinate axis. The point* $\mathbf{0} \equiv (0, \cdots, 0)$ *is called the* ***origin***.

Thus $(1, 2, 4) \in \mathbb{R}^3$ and $(2, 1, 4) \in \mathbb{R}^3$ but $(1, 2, 4) \neq (2, 1, 4)$ because, even though the same numbers are involved, they don't match up. In particular, the first entries are not equal.

Why would anyone be interested in such a thing? 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 n dimensional space and its Cartesian coordinates.

[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 the book of Genesis scientifically. Descartes ended up dying in Sweden.

12.2 Algebra in $\mathbb{R}^n$

There are two algebraic operations done with elements of $\mathbb{R}^n$. One is addition and the other is multiplication by numbers, called scalars.

Definition 12.2.1 *If $\mathbf{x} \in \mathbb{R}^n$ and a is a number, also called a **scalar**. Then $a\mathbf{x} \in \mathbb{R}^n$ is defined by*

$$a\mathbf{x} = a(x_1, \cdots, x_n) \equiv (ax_1, \cdots, ax_n). \tag{12.1}$$

*This is known as **scalar multiplication**. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ then $\mathbf{x} + \mathbf{y} \in \mathbb{R}^n$ and is defined by*

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1, \cdots, x_n) + (y_1, \cdots, y_n) \\ &\equiv (x_1 + y_1, \cdots, x_n + y_n) \end{aligned} \tag{12.2}$$

With this definition, the algebraic properties satisfy the conclusions of the following theorem.

Theorem 12.2.2 *For $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ and α, β scalars, (real numbers), the following hold.*

$$\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}, \tag{12.3}$$

the commutative law of addition,

$$(\mathbf{v} + \mathbf{w}) + \mathbf{z} = \mathbf{v} + (\mathbf{w} + \mathbf{z}), \tag{12.4}$$

the associative law for addition,

$$\mathbf{v} + \mathbf{0} = \mathbf{v}, \tag{12.5}$$

the existence of an additive identity,

$$\mathbf{v} + (-\mathbf{v}) = \mathbf{0}, \tag{12.6}$$

the existence of an additive inverse, Also

$$\alpha(\mathbf{v} + \mathbf{w}) = \alpha\mathbf{v} + \alpha\mathbf{w}, \tag{12.7}$$

$$(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}, \tag{12.8}$$

$$\alpha(\beta\mathbf{v}) = \alpha\beta(\mathbf{v}), \tag{12.9}$$

$$1\mathbf{v} = \mathbf{v}. \tag{12.10}$$

In the above $\mathbf{0} = (0, \cdots, 0)$.

You should verify these properties all hold. For example, consider (12.7)

$$\begin{aligned} \alpha(\mathbf{v} + \mathbf{w}) &= \alpha(v_1 + w_1, \cdots, v_n + w_n) \\ &= (\alpha(v_1 + w_1), \cdots, \alpha(v_n + w_n)) \\ &= (\alpha v_1 + \alpha w_1, \cdots, \alpha v_n + \alpha w_n) \\ &= (\alpha v_1, \cdots, \alpha v_n) + (\alpha w_1, \cdots, \alpha w_n) \\ &= \alpha\mathbf{v} + \alpha\mathbf{w}. \end{aligned}$$

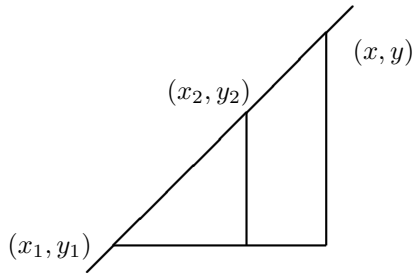
As usual subtraction is defined as $\mathbf{x} - \mathbf{y} \equiv \mathbf{x} + (-\mathbf{y})$.

12.3 Lines

To begin with consider the case $n = 1, 2$. In the case where $n = 1$, the only line is just $\mathbb{R}^1 = \mathbb{R}$. Therefore, if x_1 and x_2 are two different points in $\mathbb{R}$, consider

$$x = x_1 + t(x_2 - x_1)$$

where $t \in \mathbb{R}$ and the totality of all such points will give $\mathbb{R}$. You see that you can always solve the above equation for t, showing that every point on $\mathbb{R}$ is of this form. Now consider the plane. Does a similar formula hold? Let (x_1, y_1) and (x_2, y_2) be two different points in $\mathbb{R}^2$ which are contained in a line, l. Suppose that $x_1 \neq x_2$. Then if (x, y) is an arbitrary point on l,



Now by similar triangles,

$$m \equiv \frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1}$$

and so the point slope form of the line, l, is given as

$$y - y_1 = m(x - x_1).$$

If t is defined by

$$x = x_1 + t(x_2 - x_1),$$

you obtain this equation along with

$$\begin{aligned} y &= y_1 + mt(x_2 - x_1) \\ &= y_1 + t(y_2 - y_1). \end{aligned}$$

Therefore,

$$(x, y) = (x_1, y_1) + t(x_2 - x_1, y_2 - y_1).$$

If $x_1 = x_2$, then in place of the point slope form above, $x = x_1$. Since the two given points are different, $y_1 \neq y_2$ and so you still obtain the above formula for the line. Because of this, the following is the definition of a line in $\mathbb{R}^n$.

Definition 12.3.1 *A line in $\mathbb{R}^n$ containing the two different points, $\mathbf{x}^1$ and $\mathbf{x}^2$ is the collection of points of the form*

$$\mathbf{x} = \mathbf{x}^1 + t\left(\mathbf{x}^2 - \mathbf{x}^1\right)$$

*where $t \in \mathbb{R}$. This is known as a **parametric equation** and the variable t is called the **parameter**.*

Often t denotes time in applications to Physics. Note this definition agrees with the usual notion of a line in two dimensions and so this is consistent with earlier concepts.

Lemma 12.3.2 *Let* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ *with* $\mathbf{a} \neq \mathbf{0}$. *Then* $\mathbf{x} = t\mathbf{a} + \mathbf{b}$, $t \in \mathbb{R}$, *is a line.*

Proof: Let $\mathbf{x}^1 = \mathbf{b}$ and let $\mathbf{x}^2 - \mathbf{x}^1 = \mathbf{a}$ so that $\mathbf{x}^2 \neq \mathbf{x}^1$. Then $t\mathbf{a} + \mathbf{b} = \mathbf{x}^1 + t\left(\mathbf{x}^2 - \mathbf{x}^1\right)$ and so $\mathbf{x} = t\mathbf{a} + \mathbf{b}$ is a line containing the two different points, $\mathbf{x}^1$ and $\mathbf{x}^2$. This proves the lemma.

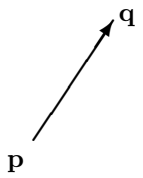
Definition 12.3.3 *The vector* $\mathbf{a}$ *in the above lemma is called a* ***direction vector*** *for the line.*

Definition 12.3.4 *Let* $\mathbf{p}$ *and* $\mathbf{q}$ *be two points in* $\mathbb{R}^n$, $\mathbf{p} \neq \mathbf{q}$. *The* ***directed line segment*** *from* $\mathbf{p}$ *to* $\mathbf{q}$, *denoted by* $\overrightarrow{\mathbf{pq}}$, *is defined to be the collection of points,*

$$\mathbf{x} = \mathbf{p} + t\left(\mathbf{q} - \mathbf{p}\right), \ t \in [0, 1]$$

with the direction corresponding to increasing t.

Think of $\overrightarrow{\mathbf{pq}}$ as an arrow whose point is on $\mathbf{q}$ and whose base is at $\mathbf{p}$ as shown in the following picture.



This line segment is a part of a line from the above Definition.

Example 12.3.5 *Find a parametric equation for the line through the points* $(1, 2, 0)$ *and* $(2, -4, 6)$.

Use the definition of a line given above to write

$$(x, y, z) = (1, 2, 0) + t\left(1, -6, 6\right), \ t \in \mathbb{R}.$$

The reason for the word, "a", rather than the word, "the" is there are infinitely many different parametric equations for the same line. To see this replace t with $3s$. Then you obtain a parametric equation for the same line because the same set of points are obtained. The difference is they are obtained from different values of the parameter. What happens is this: The line is a set of points but the parametric description gives more information than that. It tells us how the set of points are obtained. Obviously, there are many ways to trace out a given set of points and each of these ways corresponds to a different parametric equation for the line.

Example 12.3.6 *Find a parametric equation for the line which contains the point* $(1, 2, 0)$ *and has direction vector,* $(1, 2, 1)$.

From the above this is just

$$(x, y, z) = (1, 2, 0) + t\left(1, 2, 1\right), \ t \in \mathbb{R}. \tag{12.11}$$

Sometimes people elect to write a line like the above in the form

$$x = 1 + t,\ y = 2 + 2t,\ z = t,\ t \in \mathbb{R}. \tag{12.12}$$

This is a set of scalar parametric equations which amounts to the same thing as (12.11).

There is one other form for a line which is sometimes considered useful. It is the so called symmetric form. Consider the line of (12.12). You can solve for the parameter, t to write

$$t = x - 1, t = \frac{y-2}{2},\ t = z.$$

Therefore,

$$x - 1 = \frac{y-2}{2} = z.$$

This is the symmetric form of the line.

Example 12.3.7 *Suppose the* ***symmetric form of a line*** *is*

$$\frac{x-2}{3} = \frac{y-1}{2} = z + 3.$$

Find the line in parametric form.

Let $t = \frac{x-2}{3}, t = \frac{y-1}{2}$ and $t = z + 3$. Then solving for x, y, z, you get

$$x = 3t + 2,\ y = 2t + 1,\ z = t - 3,\ t \in \mathbb{R}.$$

Written in terms of vectors this is

$$(2, 1, -3) + t\,(3, 2, 1) = (x, y, z)\,,\ t \in \mathbb{R}.$$

12.4 Exercises

1. Verify all the properties (12.3)-(12.10).
2. Compute the following
   (a) $5\,(1, 2, 3, -2) + 6\,(2, 1, -2, 7)$
   (b) $5\,(1, 2, -2) - 6\,(2, 1, -2)$
   (c) $-3\,(1, 0, 3, -2) + (2, 0, -2, 1)$
   (d) $-3\,(1, -2, -3, -2) - 2\,(2, -1, -2, 7)$
   (e) $-\,(2, 2, -3, -2) + 2\,(2, 4, -2, 7)$
3. Find a parametric equation for the line through the points $(2, 3, 4)$ and $(-2, 3, 0)$.
4. Find a parametric equation for the line through the points $(2, 0, 4)$ and $(-2, 5, 0)$.
5. Find symmetric equations for the line through the points $(2, 2, 4)$ and $(-2, 3, 1)$.
6. Find symmetric equations for the line through the points $(1, 2, 4)$ and $(-2, 1, 1)$.
7. Symmetric equations for a line are

$$\frac{x+1}{3} = \frac{2x+3}{2} = z + 7.$$

   Find parametric equations for this line.

8. Symmetric equations for a line are

$$\frac{x-1}{3} = \frac{2x-3}{5} = z+5.$$

Find parametric equations for this line.

9. Parametric equations for a line are

$$x = 1+2t, y = 3-t, z = 5+3t.$$

Find symmetric equations for this line. Parametric equations for a line are

$$x = 1+4t, y = 3+t, z = 5-3t.$$

Find symmetric equations for this line.

10. Find the equation of the line through $(1,2,3)$ having direction vector, $(-3,2,-4)$.

11. Find the equation of the line through $(1,0,3)$ having direction vector, $(1,2,-4)$.

12. Parametric equations for a line are

$$x = 1+2t, y = 3-t, z = 5+3t.$$

What is a direction vector for this line?

13. Parametric equations for a line are

$$x = 1+8t, y = 3+t, z = 7+2t.$$

What is a direction vector for this line?

Systems Of Equations

13.0.1 Outcomes

1. Understand the geometric significance of a solution to a system of linear equations in simple cases.

2. Use row operations to find the complete solution to a system of equations.

1. Relate the types of solution sets of a system of two or three variables to the intersections of lines in a plane or the intersection of planes in three space.

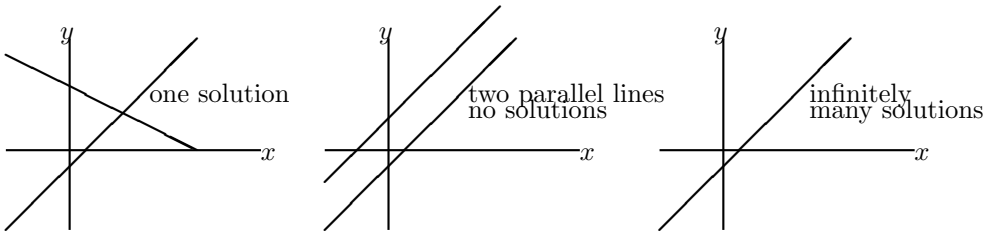
2. Determine whether a system of linear equations has no solution, a unique solution or an infinite number of solutions from its echelon form.

3. Solve a system of equations using Gauss elimination.

4. Model a physical system with linear equations and then solve

13.1 Geometric Interpretations

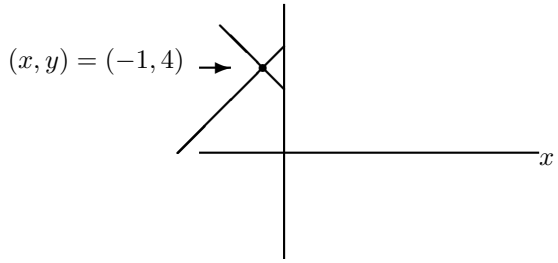
As you know from high school, equations like $2x + 3y = 6$ can be graphed as straight lines. To find the solution to two such equations, you could graph the two straight lines and the ordered pairs identifying the point (or points) of intersection would give the x and y values of the solution to the two equations because such an ordered pair satisfies both equations. The following picture illustrates what can occur with two equations involving two variables.



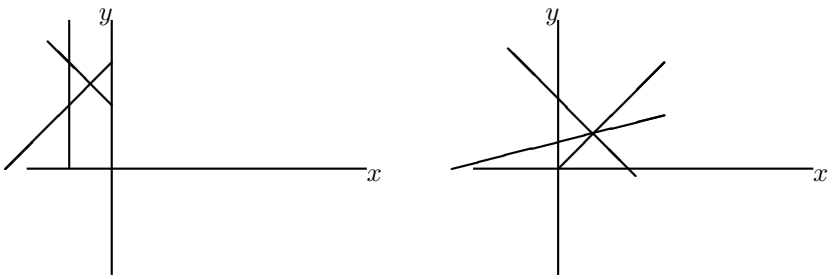
In the first example of the above picture, there is a unique point of intersection. In the second, there are no points of intersection. The other thing which can occur is that the two lines are really the same line. For example, $x + y = 1$ and $2x + 2y = 2$ are relations which when graphed yield the 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.

Example 13.1.1 *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 13.1.2 *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.

An equation like $2x+4y-5z = 8$ involving three variables is a plane in three dimensions. This will be discussed later. Therefore, the geometrical significance of solving systems of equations involving three variables, is to take the intersection of planes.

Example 13.1.3 *In the case of the intersection of planes, you can imagine that the intersection of two planes is a line and then if you have another plane, it could intersect this line in a single point or it could contain the line or be parallel to the line and not have any intersection with it.*

In higher dimensions it is customary to refer to such relations like $x + y - 2z + 4w = 8$ as a **hyper-plane**. Such pictures as above are useful in two or three dimensions for gaining insight into what can happen but they are not adequate for obtaining the exact solution set of the linear system. Furthermore, it is impossible to consider all possibilities through an attempt to draw pictures even in two or three dimensions and in higher dimensions, pictures are even less useful. The only rational and useful way to deal with this subject is through the use of algebra.

13.2 Systems Of Equations, Algebraic Procedures

13.2.1 Elementary Operations

Consider the following example.

Example 13.2.1 *Find x and y such that*

$$x + y = 7 \text{ and } 2x - y = 8. \tag{13.1}$$

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 13.2.2 ***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.*

Example 13.2.3 *To illustrate the third of these operations on this particular system, consider the following.*

$$\begin{array}{c} x + y = 7 \\ 2x - y = 8 \end{array}$$

The system has the same solution set as the system

$$\begin{array}{c} x + y = 7 \\ -3y = -6 \end{array}.$$

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 13.2.2 do not change the set of solutions to the system of linear equations.

Theorem 13.2.4 *Suppose you have two equations, involving the variables, $(x_1, \cdots, x_n)$*

$$E_1 = f_1, E_2 = f_2 \tag{13.2}$$

where E_1 and E_2 are expressions involving the variables. (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

$$E_1 = f_1,\ E_2 + aE_1 = f_2 + af_1. \tag{13.3}$$

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 $(x_1, \cdots, x_n)$ solves $E_1 = f_1, E_2 = f_2$ then it solves the first equation in $E_1 = f_1,\ E_2 + aE_1 = f_2 + af_1$. 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 $(x_1, \cdots, x_n)$ 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 $(x_1, \cdots, x_n)$ 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 $(x_1, \cdots, x_n)$ 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 13.2.5 *Find the solutions to the system,*

$$\begin{array}{c} x + 3y + 6z = 25 \\ 2x + 7y + 14z = 58 \\ 2y + 5z = 19 \end{array} \tag{13.4}$$

To solve this system replace the second equation by (-2) times the first equation added to the second. This yields the system

$$\begin{array}{c} x + 3y + 6z = 25 \\ y + 2z = 8 \\ 2y + 5z = 19 \end{array} \tag{13.5}$$

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

$$\begin{array}{c} x + 3y + 6z = 25 \\ y + 2z = 8 \\ z = 3 \end{array} \tag{13.6}$$

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 (13.6) 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

$$\begin{array}{c} x + 3y = 7 \\ y = 2 \\ z = 3 \end{array}$$

Now add (-3) times the second to the top. This yields

$$\begin{array}{c} x = 1 \\ y = 2 \\ z = 3 \end{array},$$

a system which has the same solution set as the original system. This avoided back substitution and led to the same solution set.

13.2.2 Gauss Elimination

A less cumbersome way to represent a linear system is to write it as an **augmented matrix.** For example the linear system, (13.4) can be written as

$$\left(\begin{array}{ccc|c} 1 & 3 & 6 & 25 \\ 2 & 7 & 14 & 58 \\ 0 & 2 & 5 & 19 \end{array}\right).$$

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 (13.4) would be to take (-2) times the first row of the augmented matrix above and add it to the second row,

$$\left(\begin{array}{ccc|c} 1 & 3 & 6 & 25 \\ 0 & 1 & 2 & 8 \\ 0 & 2 & 5 & 19 \end{array}\right).$$

Note how this corresponds to (13.5). Next take (-2) times the second row and add to the third,

$$\left(\begin{array}{ccc|c} 1 & 3 & 6 & 25 \\ 0 & 1 & 2 & 8 \\ 0 & 0 & 1 & 3 \end{array}\right)$$

This augmented matrix corresponds to the system

$$\begin{array}{c} x + 3y + 6z = 25 \\ y + 2z = 8 \\ z = 3 \end{array}$$

which is the same as (13.6). By back substitution you obtain the solution $x = 1, y = 6$, and $z = 3$.

In general a linear system is of the form

$$\begin{array}{c} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \end{array}, \tag{13.7}$$

where the x_i are variables and the a_{ij} and b_i are constants. This system can be represented by the augmented matrix,

$$\left(\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right). \tag{13.8}$$

Changes to the system of equations in (13.7) as a result of an elementary operations translate into changes of the augmented matrix resulting from a row operation. Note that Theorem 13.2.4 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 13.2.6 *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 13.2.7 *An augmented matrix is in **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.*

Definition 13.2.8 *An augmented matrix is in **echelon form** if all the above hold except that the entries above a leading entry are not required to equal zero.*

Example 13.2.9 *Here are some augmented matrices which are in reduced echelon form.*

$$\left(\begin{array}{ccccc|c} 1 & 0 & 0 & 5 & 8 & 2 \\ 0 & 0 & 1 & 2 & 7 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

Example 13.2.10 *Here are augmented matrices in echelon form which are not in reduced echelon form.*

$$\left(\begin{array}{ccccc|c} 1 & 0 & 6 & 5 & 8 & 2 \\ 0 & 0 & 1 & 2 & 7 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{ccc|c} 1 & 3 & 5 & 4 \\ 0 & 2 & 0 & 7 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Example 13.2.11 *Here are some augmented matrices which are not in echelon form.*

$$\left(\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 1 & 2 & 3 & 3 \\ 0 & 2 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{cc|c} 1 & 2 & 3 \\ 2 & 4 & -6 \\ 4 & 0 & 7 \end{array}\right), \left(\begin{array}{ccc|c} 0 & 2 & 3 & 3 \\ 1 & 5 & 0 & 2 \\ 7 & 5 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right).$$

Definition 13.2.12 *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 13.2.13 *Suppose*

$$A = \left(\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 6 \\ 4 & 4 & 4 & 10 \end{array}\right)$$

Where are the pivot positions and pivot columns?

Replace the second row by -3 times the first added to the second. This yields

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -6 \\ 4 & 4 & 4 & 10 \end{array}\right).$$

This is not in reduced echelon form so replace the bottom row by -4 times the top row added to the bottom. This yields

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -6 \\ 0 & -4 & -8 & -6 \end{array}\right).$$

This is still not in reduced echelon form. Replace the bottom row by -1 times the middle row added to the bottom. This yields

$$\left(\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -6 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

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:

$$\left(\begin{array}{ccc|c} \boxed{1} & 2 & 3 & 4 \\ 3 & \boxed{2} & 1 & 6 \\ 4 & 4 & 4 & 10 \end{array}\right)$$

Thus the pivot columns in the matrix are the first two columns.

The row reduction algorithm

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 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. Repeat steps 1 and 2 for the matrix obtained by ignoring the row containing the pivot and the rows above along with the pivot column and the columns to the left of the pivot column.
4. Continue till a matrix in echelon form has been obtained.
5. Finally, 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.

When applying the algorithm, it is best to not make explicit mention of the lines dividing the last column from the rest of the matrix.

Example 13.2.14 *Here is a matrix.*

$$\begin{pmatrix} 0 & 0 & 2 & 3 & 2 \\ 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \end{pmatrix}$$

Do row reductions till you obtain a matrix in echelon form. Then complete the process by producing one in 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.

$$\begin{pmatrix} 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \end{pmatrix}$$

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,

$$\begin{pmatrix} 2 & 3 & 2 \\ 1 & 2 & 2 \\ 0 & 0 & 0 \\ 0 & 2 & 1 \end{pmatrix}.$$

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

$$\begin{pmatrix} 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \end{pmatrix}.$$

The next matrix the steps in the algorithm are applied to is

$$\begin{pmatrix} -1 & -2 \\ 0 & 0 \\ 2 & 1 \end{pmatrix}.$$

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

$$\begin{pmatrix} 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -3 \end{pmatrix}.$$

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

$$\begin{pmatrix} 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

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.

$$\begin{pmatrix} 0 & 1 & 1 & 4 & 0 \\ 0 & 0 & 6 & 9 & 0 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Next work on the fourth column in the same way.

$$\begin{pmatrix} 0 & 3 & 3 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Finally, take $-1/2$ times the second row and add to the first.

$$\begin{pmatrix} 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

This is now in reduced echelon form. You can put in the dividing lines between the matrix and the last column if you desire.

$$\left(\begin{array}{cccc|c} 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

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 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 13.2.15 *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

$$\left(\begin{array}{ccc|c} 2 & 4 & -3 & -1 \\ 5 & 10 & -7 & -2 \\ 3 & 6 & 5 & 9 \end{array}\right)$$

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

$$\left(\begin{array}{ccc|c} 2 & 4 & -3 & -1 \\ 0 & 0 & 1 & 1 \\ 3 & 6 & 5 & 9 \end{array}\right)$$

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.

$$\left(\begin{array}{ccc|c} 2 & 4 & -3 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 21 \end{array}\right).$$

One more row operation, taking (-1) times the second row and adding to the bottom yields.

$$\left(\begin{array}{ccc|c} 2 & 4 & -3 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 20 \end{array}\right).$$

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 13.2.16 *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

$$\left(\begin{array}{ccc|c} 3 & -1 & -5 & 9 \\ 0 & 1 & -10 & 0 \\ -2 & 1 & 0 & -6 \end{array}\right)$$

Replace the last row with 2 times the top row added to 3 times the bottom row. This gives

$$\left(\begin{array}{ccc|c} 3 & -1 & -5 & 9 \\ 0 & 1 & -10 & 0 \\ 0 & 1 & -10 & 0 \end{array}\right).$$

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.

$$\left(\begin{array}{ccc|c} 3 & -1 & -5 & 9 \\ 0 & 1 & -10 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Take the middle row and add to the top and then divide the top row which results by 3.

$$\left(\begin{array}{ccc|c} 1 & 0 & -5 & 3 \\ 0 & 1 & -10 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

This is in reduced echelon form. The equations corresponding to this reduced echelon form are $y = 10z$ and $x = 3 + 5z$. Apparently z can equal any number. Lets call this number, t. [1]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 13.2.16 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 13.2.16 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 13.2.17 *Find the solution to the system*

$$\begin{array}{c} x + 2y - z + w = 3 \\ x + y - z + w = 1 \\ x + 3y - z + w = 5 \end{array}$$

[1]In this context t is called a **parameter.**

The augmented matrix is

$$\left(\begin{array}{cccc|c} 1 & 2 & -1 & 1 & 3 \\ 1 & 1 & -1 & 1 & 1 \\ 1 & 3 & -1 & 1 & 5 \end{array}\right).$$

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

$$\left(\begin{array}{cccc|c} 1 & 2 & -1 & 1 & 3 \\ 0 & -1 & 0 & 0 & -2 \\ 0 & 1 & 0 & 0 & 2 \end{array}\right)$$

Now add the second row to the bottom row

$$\left(\begin{array}{cccc|c} 1 & 2 & -1 & 1 & 3 \\ 0 & -1 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right) \tag{13.9}$$

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

$$\left(\begin{array}{c} x \\ y \\ z \\ w \end{array}\right) = \left(\begin{array}{c} -1 + s - t \\ 2 \\ s \\ t \end{array}\right). \tag{13.10}$$

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 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 (13.9). Instead we first place it in reduced echelon form as follows.

$$\left(\begin{array}{cccc|c} 1 & 0 & -1 & 1 & -1 \\ 0 & -1 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

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 (13.10).

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

$$\left(\begin{array}{cc|c} 1 & 2 & 3 \\ 0 & 4 & -2 \\ 0 & 0 & 1 \end{array}\right)$$

or there is a unique solution as in the case where row operations yield an echelon form like

$$\left(\begin{array}{ccc|c} 1 & 2 & 2 & 3 \\ 0 & 4 & 3 & -2 \\ 0 & 0 & 4 & 1 \end{array}\right).$$

Also, sometimes there are free variables and no solution as in the following:

$$\left(\begin{array}{ccc|c} 1 & 2 & 2 & 3 \\ 0 & 4 & 3 & -2 \\ 0 & 0 & 0 & 1 \end{array}\right).$$

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 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 reduced echelon form rather than just echelon form.

In summary,

Definition 13.2.18 *A **system of linear equations** is a list of equations,*

$$\begin{array}{c} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{array}$$

where a_{ij} are numbers, and b_j is a number. The above is a system of m equations in the n variables, $x_1, x_2 \cdots, x_n$. 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

$$\sum_{j=1}^{n} a_{ij}x_j = f_j,\ i = 1, 2, 3, \cdots, m$$

It is desired to find $(x_1, \cdots, x_n)$ 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.

13.3 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 $\blacksquare$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$\left(\begin{array}{ccccc|c} \blacksquare & * & * & * & * & * \\ 0 & \blacksquare & * & * & 0 & * \\ 0 & 0 & \blacksquare & * & * & * \\ 0 & 0 & 0 & 0 & \blacksquare & * \end{array}\right)$$

9. Here is an augmented matrix in which $*$ denotes an arbitrary number and $\blacksquare$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$\left(\begin{array}{ccc|c} \blacksquare & * & * & * \\ 0 & \blacksquare & * & * \\ 0 & 0 & \blacksquare & * \end{array}\right)$$

10. Here is an augmented matrix in which $*$ denotes an arbitrary number and $\blacksquare$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$\left(\begin{array}{ccccc|c} \blacksquare & * & * & * & * & * \\ 0 & \blacksquare & 0 & * & 0 & * \\ 0 & 0 & 0 & \blacksquare & * & * \\ 0 & 0 & 0 & 0 & \blacksquare & * \end{array}\right)$$

11. Here is an augmented matrix in which $*$ denotes an arbitrary number and $\blacksquare$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

$$\left(\begin{array}{ccccc|c} \blacksquare & * & * & * & * & * \\ 0 & \blacksquare & * & * & 0 & * \\ 0 & 0 & 0 & 0 & \blacksquare & 0 \\ 0 & 0 & 0 & 0 & * & \blacksquare \end{array}\right)$$

12. Find h such that

$$\left(\begin{array}{cc|c} 2 & h & 4 \\ 3 & 6 & 7 \end{array}\right)$$

is the augmented matrix of an inconsistent matrix.

13. Find h such that
$$\left(\begin{array}{cc|c} 1 & h & 3 \\ 2 & 4 & 6 \end{array}\right)$$
is the augmented matrix of a consistent matrix.

14. Find h such that
$$\left(\begin{array}{cc|c} 1 & 1 & 4 \\ 3 & h & 12 \end{array}\right)$$
is the augmented matrix of a consistent matrix.

15. 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.
$$\left(\begin{array}{cc|c} 1 & h & 2 \\ 2 & 4 & k \end{array}\right).$$

16. 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.
$$\left(\begin{array}{cc|c} 1 & 2 & 2 \\ 2 & h & k \end{array}\right).$$

17. Determine if the system is consistent.
$$\begin{array}{c} x+2y+z-w=2 \\ x-y+z+w=1 \\ 2x+y-z=1 \\ 4x+2y+z=5 \end{array}$$

18. Determine if the system is consistent.
$$\begin{array}{c} x+2y+z-w=2 \\ x-y+z+w=0 \\ 2x+y-z=1 \\ 4x+2y+z=3 \end{array}$$

19. Find the general solution of the system whose augmented matrix is
$$\left(\begin{array}{ccc|c} 1 & 2 & 0 & 2 \\ 1 & 3 & 4 & 2 \\ 1 & 0 & 2 & 1 \end{array}\right).$$

20. Find the general solution of the system whose augmented matrix is
$$\left(\begin{array}{ccc|c} 1 & 2 & 0 & 2 \\ 2 & 0 & 1 & 1 \\ 3 & 2 & 1 & 3 \end{array}\right).$$

21. Find the general solution of the system whose augmented matrix is
$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & 1 \\ 1 & 0 & 4 & 2 \end{array}\right).$$

22. Find the general solution of the system whose augmented matrix is

$$\left(\begin{array}{ccccc|c} 1 & 0 & 2 & 1 & 1 & 2 \\ 0 & 1 & 0 & 1 & 2 & 1 \\ 1 & 2 & 0 & 0 & 1 & 3 \\ 1 & 0 & 1 & 0 & 2 & 2 \end{array}\right).$$

23. Find the general solution of the system whose augmented matrix is

$$\left(\begin{array}{ccccc|c} 1 & 0 & 2 & 1 & 1 & 2 \\ 0 & 1 & 0 & 1 & 2 & 1 \\ 0 & 2 & 0 & 0 & 1 & 3 \\ 1 & -1 & 2 & 2 & 2 & 0 \end{array}\right).$$

24. 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.

25. Give the complete solution to the system of equations, $7x+14y+15z = 22$, $2x+4y+3z = 5$, and $3x+6y+10z = 13$.

26. 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.

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, $8x+2y+3z = -3, 8x+3y+3z = -1$, and $4x + y + 3z = -9$.

31. Give the complete solution to the system of equations, $-8x + 2y + 5z = 18, -8x + 3y + 5z = 13$, and $-4x + y + 5z = 19$.

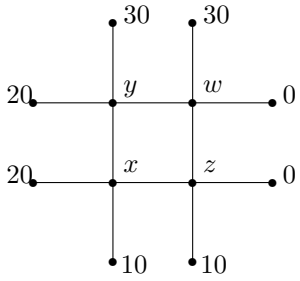
32. Give the complete solution to the system of equations, $3x - y - 2z = 3$, $y - 4z = 0$, and $-2x + y = -2$.

33. Give the complete solution to the system of equations, $-19x+8y = -108$, $-71x+30y = -404$, $-2x + y = -12$, $4x + z = 14$.

34. Four times the weight of Gaston is 150 pounds more than the weight of Ichabod. Four times the weight of Ichabod is 660 pounds less than seventeen times the weight of Gaston. 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.

35. 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 $z = \frac{1}{4}(10 + 0 + w + x)$.



36. Consider the system $-5x + 2y - z = 0$ 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?

37. Give the complete solution to the system of equations, $-9x+15y = 66$, $-11x+18y = 79$,$-x + y = 4$, and $z = 3$.

Matrices

14.0.1 Outcomes

1. Perform the basic matrix operations of matrix addition, scalar multiplication, transposition and matrix multiplication. Identify when these operations are not defined. Represent the basic operations in terms of double subscript notation.

2. Recall and prove algebraic properties for matrix addition, scalar multiplication, transposition, and matrix multiplication. Apply these properties to manipulate an algebraic expression involving matrices.

3. Evaluate the inverse of a matrix using Gauss Jordan elimination.

4. Recall the cancellation laws for matrix multiplication. Demonstrate when cancellation laws do not apply.

5. Recall and prove identities involving matrix inverses.

14.1 Matrix Arithmetic

14.1.1 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**. In these notes numbers will always be either real or complex numbers.

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 3×4 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 $\begin{pmatrix} 1 \\ 5 \\ 6 \end{pmatrix}$. 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 **Row**man **C**atholic. 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.

$$3\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 2 & 8 & 7 \\ 6 & -9 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 6 & 9 & 12 \\ 15 & 6 & 24 & 21 \\ 18 & -27 & 3 & 6 \end{pmatrix}.$$

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 areidentical. 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 **entries** of the matrix or the **components** of the matrix.

The above discussion stated for general matrices is given in the following definition.

Definition 14.1.1 *(Scalar Multiplication) If $A = (a_{ij})$ and k is a scalar, then $kA = (ka_{ij})$.*

Example 14.1.2 $7\begin{pmatrix} 2 & 0 \\ 1 & -4 \end{pmatrix} = \begin{pmatrix} 14 & 0 \\ 7 & -28 \end{pmatrix}$.

Definition 14.1.3 *(Addition) If $A = (a_{ij})$ and $B = (b_{ij})$ are two $m \times n$ matrices. Then $A + B = C$ where*

$$C = (c_{ij})$$

for $c_{ij} = a_{ij} + b_{ij}$.

Example 14.1.4

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 2 & 3 \\ -6 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 4 & 6 \\ -5 & 2 & 5 \end{pmatrix}$$

To save on notation, we will often use A_{ij} to refer to the ij^{th} entry of the matrix, A.

Definition 14.1.5 *(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.

Example 14.1.6 *The 2×3 zero matrix is* $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$.

Note there are 2×3 zero matrices, 3×4 zero matrices, etc. In fact there is a zero matrix for every size.

Definition 14.1.7 *(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.

$$A + B = B + A, \tag{14.1}$$

- Associative Law for Addition.

$$(A + B) + C = A + (B + C), \tag{14.2}$$

- Existence of an Additive Identity

$$A + 0 = A, \tag{14.3}$$

- Existence of an Additive Inverse

$$A + (-A) = 0, \tag{14.4}$$

Also for α, β scalars, the following additional properties hold.

- Distributive law over Matrix Addition.

$$\alpha (A + B) = \alpha A + \alpha B, \tag{14.5}$$

- Distributive law over Scalar Addition

$$(\alpha + \beta) A = \alpha A + \beta A, \tag{14.6}$$

- Associative law for Scalar Multiplication

$$\alpha (\beta A) = \alpha \beta (A), \tag{14.7}$$

- Rule for Multiplication by 1.

$$1A = A. \tag{14.8}$$

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.

14.1.2 Multiplication Of Matrices

Definition 14.1.8 *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*

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

*is also called a **column vector.** The $1 \times n$ matrix*

$$(x_1 \cdots x_n)$$

*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. We will illustrate the general definition by first considering 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.

$$\begin{pmatrix} \mathbf{7} & \mathbf{8} & \mathbf{9} \\ 1 & 2 & 3 \\ \mathbf{7} & \mathbf{8} & \mathbf{9} \\ 4 & 5 & 6 \end{pmatrix} \to \begin{pmatrix} 7 \times 1 + 8 \times 2 + 9 \times 3 \\ 7 \times 4 + 8 \times 5 + 9 \times 6 \end{pmatrix} = \begin{pmatrix} 50 \\ 122 \end{pmatrix}.$$

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,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{pmatrix}.$$

In general, here is the definition of how to multiply an $(m \times n)$ matrix times a $(n \times 1)$ matrix.

Definition 14.1.9 *Let $A = A_{ij}$ be an $m \times n$ matrix and let $\mathbf{v}$ be an $n \times 1$ matrix,*

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

Then $A\mathbf{v}$ is an $m \times 1$ matrix and the i^{th} component of this matrix is

$$(A\mathbf{v})_i = A_{i1}v_1 + A_{i2}v_2 + \cdots + A_{in}v_n = \sum_{j=1}^{n} A_{ij}v_j.$$

Thus

$$A\mathbf{v} = \begin{pmatrix} \sum_{j=1}^{n} A_{1j}v_j \\ \vdots \\ \sum_{j=1}^{n} A_{mj}v_j \end{pmatrix}. \tag{14.9}$$

In other words, if

$$A = (\mathbf{a}_1, \cdots, \mathbf{a}_n)$$

where the $\mathbf{a}_k$ *are the columns,*

$$A\mathbf{v} = \sum_{k=1}^{n} v_k \mathbf{a}_k$$

This follows from (14.9) and the observation that the j^{th} *column of* A *is*

$$\begin{pmatrix} A_{1j} \\ A_{2j} \\ \vdots \\ A_{mj} \end{pmatrix}$$

so (14.9) reduces to

$$v_1 \begin{pmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{m1} \end{pmatrix} + v_2 \begin{pmatrix} A_{12} \\ A_{22} \\ \vdots \\ A_{m2} \end{pmatrix} + \cdots + v_k \begin{pmatrix} A_{1n} \\ A_{2n} \\ \vdots \\ A_{mn} \end{pmatrix}$$

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 14.1.10 *Compute*

$$\begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 2 & 1 & -2 \\ 2 & 1 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 0 \\ 1 \end{pmatrix}.$$

First of all this is of the form $(3 \times 4)\,(4 \times 1)$ and so the result should be a (3×1). Note how the inside numbers cancel. To get the element in the second row and first and only column, compute

$$\begin{aligned} \sum_{k=1}^{4} a_{2k}v_k &= a_{21}v_1 + a_{22}v_2 + a_{23}v_3 + a_{24}v_4 \\ &= 0 \times 1 + 2 \times 2 + 1 \times 0 + (-2) \times 1 = 2. \end{aligned}$$

You should do the rest of the problem and verify

$$\begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 2 & 1 & -2 \\ 2 & 1 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 8 \\ 2 \\ 5 \end{pmatrix}.$$

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.

$$\overset{\text{these must match!}}{(m \times \widehat{\quad n)\,(n} \times p \quad}) = m \times p$$

Note the two outside numbers give the size of the product. Remember:

If the two middle numbers don't match, you can't multiply the matrices!

Definition 14.1.11 *When the number of columns of A equals the number of rows of B the two matrices are said to be* ***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*

$$B = (\mathbf{b}_1, \cdots, \mathbf{b}_p)$$

where $\mathbf{b}_k$ is an $n \times 1$ matrix or column vector. Then the $m \times p$ matrix, AB is defined as follows:

$$AB \equiv (A\mathbf{b}_1, \cdots, A\mathbf{b}_p) \tag{14.10}$$

where $A\mathbf{b}_k$ is an $m \times 1$ matrix or column vector which gives the k^{th} column of AB.

Example 14.1.12 *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 2×3 and the second matrix is a 3×3. Therefore, is it possible to multiply these matrices. According to the above discussion it should be a 2×3 matrix of the form

$$\left(\overbrace{\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}}^{\text{First column}}, \overbrace{\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}}^{\text{Second column}}, \overbrace{\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}}^{\text{Third column}} \right)$$

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 14.1.13 *Multiply the following.*

$$\begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 1 \\ -2 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \end{pmatrix}$$

First check if it is possible. This is of the form $(3\times 3)(2\times 3)$. 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.

Order Matters!

Matrix Multiplication Is Not Commutative!

This is very different than multiplication of numbers!

14.1.3 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 $A\mathbf{b}_j$. Now

$$\mathbf{b}_j = \begin{pmatrix} B_{1j} \\ \vdots \\ B_{nj} \end{pmatrix}$$

and from the above definition, the i^{th} entry is

$$\sum_{k=1}^{n} A_{ik}B_{kj}. \tag{14.11}$$

This shows the following definition for matrix multiplication in terms of the ij^{th} entries of the product coincides with Definition 14.1.11.

Definition 14.1.14 *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*

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}. \tag{14.12}$$

Example 14.1.15 *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 3×3 matrix. The answer is of the form

$$\left(\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}\begin{pmatrix} 2 \\ 7 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}\begin{pmatrix} 3 \\ 6 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix}\right)$$

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 3×3 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

$$\begin{aligned} \sum_j a_{3k} b_{k2} &= a_{31} b_{12} + a_{32} b_{22} \\ &= 2 \times 3 + 6 \times 6 = 42. \end{aligned}$$

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 14.1.16 *Multiply if possible* $\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \\ 0 & 0 & 0 \end{pmatrix}$.

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 14.1.17 *Multiply if possible* $\begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}$.

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 14.1.18 *Multiply if possible* $\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 & 0 \end{pmatrix}$.

In this case you are trying to do $(3 \times 1)(1 \times 4)$. The inside numbers match so you can do it. Verify

$$\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 4 & 2 & 0 \\ 1 & 2 & 1 & 0 \end{pmatrix}$$

14.1.4 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 14.1.19 *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 14.1.20 *If all multiplications and additions make sense, the following hold for matrices, A, B, C and a, b scalars.*

$$A(aB + bC) = a(AB) + b(AC) \tag{14.13}$$

$$(B + C)A = BA + CA \tag{14.14}$$

$$A(BC) = (AB)C \tag{14.15}$$

Proof: Using Definition 14.1.14,

$$\begin{aligned}
(A(aB + bC))_{ij} &= \sum_k A_{ik}(aB + bC)_{kj} \\
&= \sum_k A_{ik}(aB_{kj} + bC_{kj}) \\
&= a\sum_k A_{ik}B_{kj} + b\sum_k A_{ik}C_{kj} \\
&= a(AB)_{ij} + b(AC)_{ij} \\
&= (a(AB) + b(AC))_{ij}.
\end{aligned}$$

Thus $A(B + C) = AB + AC$ as claimed. Formula (14.14) is entirely similar.

Formula (14.15) is the associative law of multiplication. Using Definition 14.1.14,

$$\begin{aligned}
(A(BC))_{ij} &= \sum_k A_{ik}(BC)_{kj} \\
&= \sum_k A_{ik}\sum_l B_{kl}C_{lj} \\
&= \sum_l (AB)_{il}C_{lj} \\
&= ((AB)C)_{ij}.
\end{aligned}$$

This proves (14.15).

14.1.5 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.

$$\begin{pmatrix} 1 & 4 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}^T = \begin{pmatrix} 1 & 3 & 2 \\ 4 & 1 & 6 \end{pmatrix}$$

What happened? The first column became the first row and the second column became the second row. Thus the 3×2 matrix became a 2×3 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 14.1.21 *Let A be an $m \times n$ matrix. Then A^T denotes the $n \times m$ matrix which is defined as follows.*

$$\left(A^T\right)_{ij} = A_{ji}$$

Example 14.1.22

$$\begin{pmatrix} 1 & 2 & -6 \\ 3 & 5 & 4 \end{pmatrix}^T = \begin{pmatrix} 1 & 3 \\ 2 & 5 \\ -6 & 4 \end{pmatrix}.$$

The transpose of a matrix has the following important properties.

Lemma 14.1.23 *Let A be an $m \times n$ matrix and let B be a $n \times p$ matrix. Then*

$$(AB)^T = B^T A^T \tag{14.16}$$

and if α and β are scalars,

$$(\alpha A + \beta B)^T = \alpha A^T + \beta B^T \tag{14.17}$$

Proof: From the definition,

$$\begin{aligned} \left((AB)^T\right)_{ij} &= (AB)_{ji} \\ &= \sum_k A_{jk} B_{ki} \\ &= \sum_k \left(B^T\right)_{ik} \left(A^T\right)_{kj} \\ &= \left(B^T A^T\right)_{ij} \end{aligned}$$

The proof of Formula (14.17) is left as an exercise and this proves the lemma.

Definition 14.1.24 *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 14.1.25 *Let*

$$A = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 5 & -3 \\ 3 & -3 & 7 \end{pmatrix}.$$

Then A is symmetric.

Example 14.1.26 *Let*

$$A = \begin{pmatrix} 0 & 1 & 3 \\ -1 & 0 & 2 \\ -3 & -2 & 0 \end{pmatrix}$$

Then A is skew symmetric.

14.1.6 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 1×1 identity matrix, the second is the 2×2 identity matrix, the third is the 3×3 identity matrix, and the fourth is the 4×4 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 δ_{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 14.1.27 *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 $I_m A = A$.*

Proof:

$$\begin{aligned} (AI_n)_{ij} &= \sum_k A_{ik}\delta_{kj} \\ &= A_{ij} \end{aligned}$$

and so $AI_n = A$. The other case is left as an exercise for you.

Definition 14.1.28 *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 14.1.29 *Suppose A^{-1} exists and $AB = BA = I$. Then $B = A^{-1}$.*

Proof:

$$A^{-1} = A^{-1}I = A^{-1}(AB) = \left(A^{-1}A\right)B = IB = B.$$

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 14.1.30 *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

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = A^{-1}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) = A^{-1}\left(A\begin{pmatrix} -1 \\ 1 \end{pmatrix}\right) =$$

$$= \left(A^{-1}A\right)\begin{pmatrix} -1 \\ 1 \end{pmatrix} = I\begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

a contradiction. Thus the answer is that A does not have an inverse.

Example 14.1.31 *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.

14.1.7 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,

$$x + y = 1, x + 2y = 0$$

and

$$z + w = 0, z + 2w = 1.$$

Writing the augmented matrix for these two systems gives

$$\left(\begin{array}{cc|c} 1 & 1 & 1 \\ 1 & 2 & 0 \end{array}\right) \tag{14.18}$$

for the first system and

$$\left(\begin{array}{cc|c} 1 & 1 & 0 \\ 1 & 2 & 1 \end{array}\right) \tag{14.19}$$

for the second. Lets solve the first system. Take (-1) times the first row and add to the second to get

$$\left(\begin{array}{cc|c} 1 & 1 & 1 \\ 0 & 1 & -1 \end{array}\right)$$

Now take (-1) times the second row and add to the first to get

$$\left(\begin{array}{cc|c} 1 & 0 & 2 \\ 0 & 1 & -1 \end{array}\right).$$

Putting in the variables, this says $x = 2$ and $y = -1$.

Now solve the second system, (14.19) to find z and w. Take (-1) times the first row and add to the second to get

$$\left(\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 1 & 1 \end{array}\right).$$

Now take (-1) times the second row and add to the first to get

$$\left(\begin{array}{cc|c} 1 & 0 & -1 \\ 0 & 1 & 1 \end{array}\right).$$

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 $(I|\mathbf{v})$ where I is the identity and $\mathbf{v}$ gave a column of the inverse. In the above, $\begin{pmatrix} x \\ y \end{pmatrix}$, the first column of the inverse was obtained first and then the second column $\begin{pmatrix} z \\ w \end{pmatrix}$.

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 14.1.32 *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

$$(I|B). \tag{14.20}$$

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 14.1.33 *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

$$\left(\begin{array}{ccc|ccc} -7 & -14 & 0 & -6 & 5 & -2 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{array}\right).$$

Now take $(-7/5)$ times the second row and add to the top.

$$\left(\begin{array}{ccc|ccc} -7 & 0 & 0 & 1 & -2 & -2 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{array}\right).$$

Finally divide the top row by -7, the second row by -10 and the bottom row by 14 which yields

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ 0 & 1 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{array}\right).$$

Therefore, the inverse is

$$\left(\begin{array}{ccc} -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{array}\right)$$

Example 14.1.34 *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, $(A|I)$

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 & 1 & 0 \\ 2 & 2 & 4 & 0 & 0 & 1 \end{array}\right)$$

and proceed to do row operations attempting to obtain $\left(I|A^{-1}\right)$. Take (-1) times the top row and add to the second. Then take (-2) times the top row and add to the bottom.

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -2 & 0 & -1 & 1 & 0 \\ 0 & -2 & 0 & -2 & 0 & 1 \end{array}\right)$$

Next add (-1) times the second row to the bottom row.

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -2 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & -1 & 1 \end{array}\right)$$

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 14.1.35 *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,

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 & 1 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 \end{array}\right).$$

Now do row operations until the $n \times n$ matrix on the left becomes the identity matrix. This yields after some computations,

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 & -\frac{1}{2} & -\frac{1}{2} \end{array}\right)$$

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.

14.2 Exercises

1. Here are some matrices:

$$\begin{aligned} A &= \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 7 \end{pmatrix}, B = \begin{pmatrix} 3 & -1 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\ C &= \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}, D = \begin{pmatrix} -1 & 2 \\ 2 & -3 \end{pmatrix}, E = \begin{pmatrix} 2 \\ 3 \end{pmatrix}. \end{aligned}$$

   Find if possible $-3A, 3B - A, AC, CB, AE, EA$. If it is not possible explain why.

2. Here are some matrices:

$$\begin{aligned} A &= \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\ C &= \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 & 1 \\ 4 & -3 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}. \end{aligned}$$

   Find if possible $-3A, 3B - A, AC, CA, AE, EA, BE, DE$. If it is not possible explain why.

3. Here are some matrices:

$$\begin{aligned} A &= \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\ C &= \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 & 1 \\ 4 & -3 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}. \end{aligned}$$

Find if possible $-3A^T, 3B - A^T, AC, CA, AE, E^TB, BE, DE, EE^T, E^TE$. If it is not possible explain why.

4. Here are some matrices:

$$\begin{aligned} A &= \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\ C &= \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 \\ 4 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}. \end{aligned}$$

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 $A = \begin{pmatrix} 1 & 1 \\ -2 & -1 \\ 1 & 2 \end{pmatrix}$, $B = \begin{pmatrix} 1 & -1 & -2 \\ 2 & 1 & -2 \end{pmatrix}$, 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. 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?

7. 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?

8. Let $\mathbf{x} = (-1, -1, 1)$ and $\mathbf{y} = (0, 1, 2)$. Find $\mathbf{x}^T\mathbf{y}$ and $\mathbf{x}\mathbf{y}^T$ if possible.

9. 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.

10. 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) $(A+B)^2 = A^2 + AB + BA + B^2$

(e) $A^2B^2 = A(AB)B$

(f) $(A+B)^3 = A^3 + 3A^2B + 3AB^2 + B^3$

(g) $(A+B)(A-B) = A^2 - B^2$

11. Let $A = \begin{pmatrix} -1 & -1 \\ 3 & 3 \end{pmatrix}$. Find $\boxed{\underline{\textbf{all}}}$ 2×2 matrices, B such that $AB = 0$.

12. In (14.1) - (14.8) describe $-A$ and 0.

13. Let A be an $n \times n$ matrix. Show A equals the sum of a symmetric and a skew symmetric matrix.

14. 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.

15. Using only the properties (14.1) - (14.8) show $-A$ is unique.

16. Using only the properties (14.1) - (14.8) show 0 is unique.

17. Using only the properties (14.1) - (14.8) 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.

18. Using only the properties (14.1) - (14.8) and previous problems show $(-1)A = -A$.

19. Prove (14.17).

20. Prove that $I_m A = A$ where A is an $m \times n$ matrix.

21. 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.

22. 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.

23. 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.

24. 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.

25. Give an example of matrices, A, B, C such that $B \neq C$, $A \neq 0$, and yet $AB = AC$.

26. 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?

27. Find your own examples:

    (a) 2×2 matrices, A and B such that $A \neq 0, B \neq 0$ with $AB \neq BA$.

    (b) 2×2 matrices, A and B such that $A \neq 0, B \neq 0$, but $AB = 0$.

    (c) 2×2 matrices, A, D, and C such that $A \neq 0, C \neq D$, but $AC = AD$.

28. Explain why if $AB = AC$ and A^{-1} exists, then $B = C$.

29. Give an example of a matrix, A such that $A^2 = I$ and yet $A \neq I$ and $A \neq -I$.

30. Give an example of matrices, A, B such that neither A nor B equals zero and yet $AB = 0$.

31. Give another example other than the one given in this section of two square matrices, A and B such that $AB \neq BA$.

32. 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 $B = A^{-1}$.

33. Show $(AB)^{-1} = B^{-1}A^{-1}$.

34. Show that if A is an invertible $n \times n$ matrix, then so is A^T and $\left(A^T\right)^{-1} = \left(A^{-1}\right)^T$.

35. Show that if A is an $n \times n$ invertible matrix and $\mathbf{x}$ is a $n \times 1$ matrix such that $A\mathbf{x} = \mathbf{b}$ for $\mathbf{b}$ an $n \times 1$ matrix, then $\mathbf{x} = A^{-1}\mathbf{b}$.

36. Prove that if A^{-1} exists and $A\mathbf{x} = \mathbf{0}$ then $\mathbf{x} = \mathbf{0}$.

37. Show that $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ by verifying that $(ABC)\left(C^{-1}B^{-1}A^{-1}\right) = I$. Assume for now that if a matrix acts like the inverse on one side of a matrix, then it is the inverse and will work as such on the other side.

Determinants

15.0.1 Outcomes

1. Evaluate the determinant of a square matrix using by applying
   (a) the cofactor formula or
   (b) row operations.
2. Recall the general properties of determinants.
3. Recall that the determinant of a product of matrices is the product of the determinants. Recall that the determinant of a matrix is equal to the determinant of its transpose.
4. Apply Cramer's Rule to solve a 2×2 or a 3×3 linear system.
5. Use determinants to determine whether a matrix has an inverse.
6. Evaluate the inverse of a matrix using cofactors.

15.1 Basic Techniques And Properties

15.1.1 Cofactors And 2×2 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 2×2 matrix, this number is very easy to find.

Definition 15.1.1 *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

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix}.$$

Example 15.1.2 *Find* $\det \begin{pmatrix} 2 & 4 \\ -1 & 6 \end{pmatrix}$.

From the definition this is just $(2)(6) - (-1)(4) = 16$.

Having defined what is meant by the determinant of a 2×2 matrix, what about a 3×3 matrix?

Definition 15.1.3 *Suppose A is a 3×3 matrix. The ij^{th} **minor,** denoted as $\text{minor}(A)_{ij}$, is the determinant of the 2×2 matrix which results from deleting the i^{th} row and the j^{th} column.*

Example 15.1.4 *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 2×2 matrix which results when you delete the first row and the second column. This minor is therefore

$$\det \begin{pmatrix} 4 & 2 \\ 3 & 1 \end{pmatrix} = -2.$$

The $(2,3)$ minor is the determinant of the 2×2 matrix which results when you delete the second row and the third column. This minor is therefore

$$\det \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = -4.$$

Definition 15.1.5 *Suppose A is a 3×3 matrix. The ij^{th} **cofactor** is defined to be $(-1)^{i+j} \times \left(ij^{th} \text{ minor}\right)$. 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 15.1.6 *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 2×2 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 2×2 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 15.1.7 *The determinant of a 3×3 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 15.1.8 *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**".

$$\overbrace{1(-1)^{1+1}\begin{vmatrix} 3 & 2 \\ 2 & 1 \end{vmatrix}}^{\mathrm{cof}(A)_{11}} + \overbrace{4(-1)^{2+1}\begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix}}^{\mathrm{cof}(A)_{21}} + \overbrace{3(-1)^{3+1}\begin{vmatrix} 2 & 3 \\ 3 & 2 \end{vmatrix}}^{\mathrm{cof}(A)_{31}} = 0.$$

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.

$$\overbrace{4(-1)^{2+1}\begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix}}^{\mathrm{cof}(A)_{21}} + \overbrace{3(-1)^{2+2}\begin{vmatrix} 1 & 3 \\ 3 & 1 \end{vmatrix}}^{\mathrm{cof}(A)_{22}} + \overbrace{2(-1)^{2+3}\begin{vmatrix} 1 & 2 \\ 3 & 2 \end{vmatrix}}^{\mathrm{cof}(A)_{23}} = 0.$$

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 15.1.9 *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,** $\mathrm{cof}(A)_{ij}$ is defined to be $(-1)^{i+j} \times \left(ij^{th} \text{ minor}\right)$. In words, you multiply $(-1)^{i+j}$ times the ij^{th} minor to get the ij^{th} cofactor.*

Definition 15.1.10 *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 15.1.11 *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. Lets pick the third column. $\det(A) =$

$$3(-1)^{1+3}\begin{vmatrix} 5 & 4 & 3 \\ 1 & 3 & 5 \\ 3 & 4 & 2 \end{vmatrix} + 2(-1)^{2+3}\begin{vmatrix} 1 & 2 & 4 \\ 1 & 3 & 5 \\ 3 & 4 & 2 \end{vmatrix} +$$

$$4\left(-1\right)^{3+3}\begin{vmatrix}1&2&4\\5&4&3\\3&4&2\end{vmatrix}+3\left(-1\right)^{4+3}\begin{vmatrix}1&2&4\\5&4&3\\1&3&5\end{vmatrix}.$$

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 $4\times3\times2=24$ 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 The above examples motivate the following definitions, the second of which is incredible.

Definition 15.1.12 *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****, $\mathrm{cof}(A)$ is defined by $\mathrm{cof}(A)=(c_{ij})$ 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 $(-1)^{i+j}$. Thus $(-1)^{i+j}\times\left(\text{the } ij^{th} \text{ minor}\right)$ equals the ij^{th} cofactor. To make the formulas easier to remember, $\mathrm{cof}(A)_{ij}$ 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 15.1.13 *Let A be an $n\times n$ matrix where $n\geq2$ and suppose the determinant of an $(n-1)\times(n-1)$ has been defined. Then*

$$\det(A)=\sum_{j=1}^{n}a_{ij}\,\mathrm{cof}(A)_{ij}=\sum_{i=1}^{n}a_{ij}\,\mathrm{cof}(A)_{ij}. \tag{15.1}$$

The first formula consists of expanding the determinant along the i^{th} row and the second expands the determinant along the j^{th} column. This is called the method of Laplace expansion.

Theorem 15.1.14 *Expanding the $n\times n$ matrix along any row or column always gives the same answer so the above definition is a good definition.*

15.1.2 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 15.1.15 *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 15.1.16 *Let M be an upper (lower) triangular matrix. Then* $\det(M)$ *is obtained by taking the product of the entries on the main diagonal.*

Example 15.1.17 *Let*

$$A = \begin{pmatrix} 1 & 2 & 3 & 77 \\ 0 & 2 & 6 & 7 \\ 0 & 0 & 3 & 33.7 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

Find $\det(A)$.

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

$$1\begin{vmatrix} 2 & 6 & 7 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{vmatrix} + 0\,(-1)^{2+1}\begin{vmatrix} 2 & 3 & 77 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{vmatrix} +$$
$$0\,(-1)^{3+1}\begin{vmatrix} 2 & 3 & 77 \\ 2 & 6 & 7 \\ 0 & 0 & -1 \end{vmatrix} + 0\,(-1)^{4+1}\begin{vmatrix} 2 & 3 & 77 \\ 2 & 6 & 7 \\ 0 & 3 & 33.7 \end{vmatrix}$$

and the only nonzero term in the expansion is

$$1\begin{vmatrix} 2 & 6 & 7 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{vmatrix}.$$

Now expand this along the first column to obtain

$$1 \times \left(2 \times \begin{vmatrix} 3 & 33.7 \\ 0 & -1 \end{vmatrix} + 0\,(-1)^{2+1}\begin{vmatrix} 6 & 7 \\ 0 & -1 \end{vmatrix} + 0\,(-1)^{3+1}\begin{vmatrix} 6 & 7 \\ 3 & 33.7 \end{vmatrix}\right)$$
$$= 1 \times 2 \times \begin{vmatrix} 3 & 33.7 \\ 0 & -1 \end{vmatrix}$$

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.

15.1.3 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 15.1.18 *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 15.1.19 *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 15.1.20 *Let* $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, A_1 = \begin{pmatrix} 2 & 4 \\ 3 & 4 \end{pmatrix}$. $\det(A) = -2$, $\det(A_1) = -4$.

Theorem 15.1.21 *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 15.1.22 *Let* $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ *and let* $A_1 = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}$. $\det A = -2$, $\det(A_1) = 2$.

Theorem 15.1.23 *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 15.1.24 *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 15.1.25 *In Theorems 15.1.19 - 15.1.23 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 15.1.26 *Let A and B be two $n \times n$ matrices. Then*

$$\boxed{\det(AB) = \det(A)\det(B).}$$

Also,

$$\boxed{\det(A) = \det\left(A^T\right).}$$

Example 15.1.27 *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 $\det(A)\det(B) = 8 \times (-5) = -40$.

15.1.4 Finding Determinants Using Row Operations

Theorems 15.1.23 - 15.1.25 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 15.1.28 *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 15.1.23, 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 15.1.23 this didn't change the value of the determinant. Then the last row was multiplied by (-3). By Theorem 15.1.19 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 15.1.23. 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 3×3 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 15.1.29 *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 15.1.23 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 15.1.25 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 15.1.23 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.

15.2 Applications

15.2.1 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 14.1.28 on Page 377. Also recall the definition of the cofactor matrix given in Definition 15.1.12 on Page 388. 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 15.2.1 *A^{-1} exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = \left(a_{ij}^{-1}\right)$ where*

$$a_{ij}^{-1} = \det(A)^{-1} \operatorname{cof}(A)_{ji}$$

for $\operatorname{cof}(A)_{ij}$ the ij^{th} cofactor of A.

Example 15.2.2 *Find the inverse of the matrix,*

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 0 & 1 \\ 1 & 2 & 1 \end{pmatrix}$$

First find the determinant of this matrix. Using Theorems 15.1.23 - 15.1.25 on Page 390, 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 15.2.3 *Find the inverse of the matrix,*

$$A = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{6} & \frac{1}{3} & -\frac{1}{2} \\ -\frac{5}{6} & \frac{2}{3} & -\frac{1}{2} \end{pmatrix}$$

First find its determinant. This determinant is $\frac{1}{6}$. The inverse is therefore equal to

$$6\begin{pmatrix} \begin{vmatrix} \frac{1}{3} & -\frac{1}{2} \\ \frac{2}{3} & -\frac{1}{2} \end{vmatrix} & -\begin{vmatrix} -\frac{1}{6} & -\frac{1}{2} \\ -\frac{5}{6} & -\frac{1}{2} \end{vmatrix} & \begin{vmatrix} -\frac{1}{6} & \frac{1}{3} \\ -\frac{5}{6} & \frac{2}{3} \end{vmatrix} \\ -\begin{vmatrix} 0 & \frac{1}{2} \\ \frac{2}{3} & -\frac{1}{2} \end{vmatrix} & \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{5}{6} & -\frac{1}{2} \end{vmatrix} & -\begin{vmatrix} \frac{1}{2} & 0 \\ -\frac{5}{6} & \frac{2}{3} \end{vmatrix} \\ \begin{vmatrix} 0 & \frac{1}{2} \\ \frac{1}{3} & -\frac{1}{2} \end{vmatrix} & -\begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{6} & -\frac{1}{2} \end{vmatrix} & \begin{vmatrix} \frac{1}{2} & 0 \\ -\frac{1}{6} & \frac{1}{3} \end{vmatrix} \end{pmatrix}^T.$$

Expanding all the 2×2 determinants this yields

$$6\begin{pmatrix} \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{6} & -\frac{1}{3} \\ -\frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}^T = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix}$$

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 am interested in getting the right answer. A common mistake is to forget to take the transpose of the cofactor matrix.

Proof of Theorem 15.2.1: From the definition of the determinant in terms of expansion along a column, and letting $(a_{ir}) = A$, if $\det(A) \neq 0$,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ir} \det(A)^{-1} = \det(A) \det(A)^{-1} = 1.$$

Now consider

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1}$$

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 Theorem 15.1.21. However, expanding this matrix, B_k along the k^{th} column yields

$$0 = \det(B_k) \det(A)^{-1} = \sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1}$$

Summarizing,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1} = \delta_{rk} \equiv \begin{cases} 1 \text{ if } r = k \\ 0 \text{ if } r \neq k \end{cases}.$$

Now

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} = \sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)^{T}_{ki}$$

which is the kr^{th} entry of $\operatorname{cof}(A)^{T} A$. Therefore,

$$\frac{\operatorname{cof}(A)^{T}}{\det(A)} A = I. \tag{15.2}$$

Using the other formula in Definition 15.1.13, and similar reasoning,

$$\sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)_{kj} \det(A)^{-1} = \delta_{rk}$$

Now

$$\sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)_{kj} = \sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)^{T}_{jk}$$

which is the rk^{th} entry of $A \operatorname{cof}(A)^{T}$. Therefore,

$$A \frac{\operatorname{cof}(A)^{T}}{\det(A)} = I, \tag{15.3}$$

and it follows from (15.2) and (15.3) that $A^{-1} = \left(a_{ij}^{-1}\right)$, where

$$a_{ij}^{-1} = \operatorname{cof}(A)_{ji} \det(A)^{-1}.$$

In other words,

$$A^{-1} = \frac{\operatorname{cof}(A)^{T}}{\det(A)}.$$

Now suppose A^{-1} exists. Then by Theorem 15.1.26,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A)\det\left(A^{-1}\right)$$

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 15.2.4 *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

$$\frac{1}{e^t}\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t\cos t & e^t\sin t \\ 0 & -e^t\sin t & e^t\cos t \end{pmatrix}^T = \begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}.$$

15.2.2 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, $\mathbf{x}$, to a system of equations, $A\mathbf{x} = \mathbf{y}$ 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, $A\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$, it follows that if A^{-1} exists,

$$\mathbf{x} = \left(A^{-1}A\right)\mathbf{x} = A^{-1}(A\mathbf{x}) = A^{-1}\mathbf{y}$$

thus solving the system. Now in the case that A^{-1} exists, there is a formula for A^{-1} given above. Using this formula,

$$x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \sum_{j=1}^{n} \frac{1}{\det(A)} \operatorname{cof}(A)_{ji}\, y_j.$$

By the formula for the expansion of a determinant along a column,

$$x_i = \frac{1}{\det(A)} \det\begin{pmatrix} * & \cdots & y_1 & \cdots & * \\ \vdots & & \vdots & & \vdots \\ * & \cdots & y_n & \cdots & * \end{pmatrix},$$

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 15.2.5 *Suppose A is an $n \times n$ matrix and it is desired to solve the system $A\mathbf{x} = \mathbf{y}, \mathbf{y} = (y_1, \cdots, y_n)^T$ for $\mathbf{x} = (x_1, \cdots, x_n)^T$. 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 15.2.6 *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,

$$x = \frac{\begin{vmatrix} 1 & 2 & 1 \\ 2 & 2 & 1 \\ 3 & -3 & 2 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 1 \\ 3 & 2 & 1 \\ 2 & -3 & 2 \end{vmatrix}} = \frac{1}{2}$$

Now to find y,

$$y = \frac{\begin{vmatrix} 1 & 1 & 1 \\ 3 & 2 & 1 \\ 2 & 3 & 2 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 1 \\ 3 & 2 & 1 \\ 2 & -3 & 2 \end{vmatrix}} = -\frac{1}{7}$$

$$z = \frac{\begin{vmatrix} 1 & 2 & 1 \\ 3 & 2 & 2 \\ 2 & -3 & 3 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 1 \\ 3 & 2 & 1 \\ 2 & -3 & 2 \end{vmatrix}} = \frac{11}{14}$$

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 15.2.7 *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

$$z = \frac{\begin{vmatrix} 1 & 0 & 1 \\ 0 & e^t \cos t & t \\ 0 & -e^t \sin t & t^2 \end{vmatrix}}{\begin{vmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{vmatrix}} = t\left((\cos t)\, t + \sin t\right) e^{-t}.$$

You end up doing this sort of thing sometimes in ordinary differential equations in the method of variation of parameters.

15.3 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.

$$\begin{vmatrix} 1 & 4 & 1 & 2 \\ 3 & 2 & -2 & 3 \\ -1 & 0 & 3 & 3 \\ 2 & 1 & 2 & -2 \end{vmatrix}$$

10. 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.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

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.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} c & d \\ a & b \end{pmatrix}$$

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.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & b \\ a+c & b+d \end{pmatrix}$$

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.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & b \\ 2c & 2d \end{pmatrix}$$

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.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} b & a \\ d & c \end{pmatrix}$$

15. Tell whether the statement is true or false.

    (a) If A is a 3×3 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 $\det\left(A^{-1}\right) = \det(A)^{-1}$.

    (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 $\det\left(A^T A\right) \geq 0$.

    (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 $\det(A) = 0$.

16. Verify an example of each property of determinants found in Theorems 15.1.23 - 15.1.25 for 2×2 matrices.

17. 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?

18. Fill in the missing entries to make the matrix orthogonal as in Problem 17.

$$\begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & - & - \\ - & \frac{\sqrt{6}}{3} & - \end{pmatrix}.$$

19. If A^{-1} exist, what is the relationship between $\det(A)$ and $\det\left(A^{-1}\right)$. Explain your answer.

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. 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 $\det(A) = 0$.

22. Show $\det(aA) = a^n \det(A)$ where here A is an $n \times n$ matrix and a is a scalar.

23. 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?

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. Prove by doing computations that $\det(AB) = \det(A)\det(B)$ if A and B are 2×2 matrices.

28. Illustrate with an example of 2×2 matrices that the determinant of a product equals the product of the determinants.

29. 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 $\det(A)$?

30. Use Cramer's rule to find the solution to

$$\begin{aligned} x + 2y &= 1 \\ 2x - y &= 2 \end{aligned}$$

31. Use Cramer's rule to find the solution to

$$\begin{aligned} x + 2y + z &= 1 \\ 2x - y - z &= 2 \\ x + z &= 1 \end{aligned}$$

32. 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.

33. 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.

34. 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.

35. 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.

36. Here is a matrix,
$$\begin{pmatrix} e^t & \cosh t & \sinh t \\ e^t & \sinh t & \cosh t \\ e^t & \cosh t & \sinh t \end{pmatrix}$$
Does there exist a value of t for which this matrix fails to have an inverse? Explain.

37. 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 it is not possible to take the inverse, explain why.

38. Use the formula for the inverse in terms of the cofactor matrix to find the inverse of the matrix,
$$A = \begin{pmatrix} e^t & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & e^t \cos t - e^t \sin t & e^t \cos t + e^t \sin t \end{pmatrix}.$$

39. Find the inverse if it exists of the matrix,
$$\begin{pmatrix} e^t & \cos t & \sin t \\ e^t & -\sin t & \cos t \\ e^t & -\cos t & -\sin t \end{pmatrix}.$$

40. Let $F(t) = \det \begin{pmatrix} a(t) & b(t) \\ c(t) & d(t) \end{pmatrix}$. Verify
$$F'(t) = \det \begin{pmatrix} a'(t) & b'(t) \\ c(t) & d(t) \end{pmatrix} + \det \begin{pmatrix} a(t) & b(t) \\ c'(t) & d'(t) \end{pmatrix}.$$
Now suppose
$$F(t) = \det \begin{pmatrix} a(t) & b(t) & c(t) \\ d(t) & e(t) & f(t) \\ g(t) & h(t) & i(t) \end{pmatrix}.$$
Use Laplace expansion and the first part to verify $F'(t) =$
$$\det \begin{pmatrix} a'(t) & b'(t) & c'(t) \\ d(t) & e(t) & f(t) \\ g(t) & h(t) & i(t) \end{pmatrix} + \det \begin{pmatrix} a(t) & b(t) & c(t) \\ d'(t) & e'(t) & f'(t) \\ g(t) & h(t) & i(t) \end{pmatrix}$$
$$+ \det \begin{pmatrix} a(t) & b(t) & c(t) \\ d(t) & e(t) & f(t) \\ g'(t) & h'(t) & i'(t) \end{pmatrix}.$$
Conjecture a general result valid for $n \times n$ matrices and explain why it will be true. Can a similar thing be done with the columns?

41. Let $Ly = y^{(n)} + a_{n-1}(x) y^{(n-1)} + \cdots + a_1(x) y' + a_0(x) y$ where the a_i are given continuous functions defined on a closed interval, (a, b) and y is some function which

has n derivatives so it makes sense to write Ly. Suppose $Ly_k = 0$ for $k = 1, 2, \cdots, n$. The **Wronskian** of these functions, y_i is defined as

$$W(y_1, \cdots, y_n)(x) \equiv \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n-1)}(x) & \cdots & y_n^{(n-1)}(x) \end{pmatrix}$$

Show that for $W(x) = W(y_1, \cdots, y_n)(x)$ to save space,

$$W'(x) = \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n)}(x) & \cdots & y_n^{(n)}(x) \end{pmatrix}.$$

Now use the differential equation, $Ly = 0$ which is satisfied by each of these functions, y_i and properties of determinants presented above to verify that $W' + a_{n-1}(x) W = 0$. Give an explicit solution of this linear differential equation, **Abel's formula,** and use your answer to verify that the Wronskian of these solutions to the equation, $Ly = 0$ either vanishes identically on (a, b) or never. **Hint:** To solve the differential equation, let $A'(x) = a_{n-1}(x)$ and multiply both sides of the differential equation by $e^{A(x)}$ and then argue the left side is the derivative of something.

15.4 The Mathematical Theory Of Determinants

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 [23]. A recent book which also has a good introduction is Baker [4].

The following Lemma will be essential in the definition of the determinant.

Lemma 15.4.1 *There exists a unique function,* 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.*

$$\operatorname{sgn}_n(1, \cdots, n) = 1 \tag{15.4}$$

$$\operatorname{sgn}_n(i_1, \cdots, p, \cdots, q, \cdots, i_n) = -\operatorname{sgn}_n(i_1, \cdots, q, \cdots, p, \cdots, i_n) \tag{15.5}$$

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,* $(i_1, \cdots, i_n)$,

$$\operatorname{sgn}_n (i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) \equiv$$

$$(-1)^{n-\theta} \operatorname{sgn}_{n-1} (i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n) \tag{15.6}$$

where $n = i_\theta$ *in the ordered list,* $(i_1, \cdots, i_n)$.

Proof: To begin with, it is necessary to show the existence of such a function. This is clearly true if $n = 1$. Define $\operatorname{sgn}_1 (1) \equiv 1$ and observe that it works. No switching is possible. In the case where $n = 2$, it is also clearly true. Let $\operatorname{sgn}_2 (1, 2) = 1$ and $\operatorname{sgn}_2 (2, 1) = 0$ while $\operatorname{sgn}_2 (2, 2) = \operatorname{sgn}_2 (1, 1) = 0$ and verify it works. Assuming such a function exists for n, sgn_{n+1} will be defined in terms of sgn_n. If there are any repeated numbers in $(i_1, \cdots, i_{n+1})$, $\operatorname{sgn}_{n+1} (i_1, \cdots, i_{n+1}) \equiv 0$. If there are no repeats, then $n + 1$ appears somewhere in the ordered list. Let θ 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 (15.6) it must be that

$$\operatorname{sgn}_{n+1} (i_1, \cdots, i_{\theta-1}, n + 1, i_{\theta+1}, \cdots, i_{n+1}) \equiv$$

$$(-1)^{n+1-\theta} \operatorname{sgn}_n (i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_{n+1}).$$

It is necessary to verify this satisfies (15.4) and (15.5) with n replaced with $n + 1$. The first of these is obviously true because

$$\operatorname{sgn}_{n+1} (1, \cdots, n, n + 1) \equiv (-1)^{n+1-(n+1)} \operatorname{sgn}_n (1, \cdots, n) = 1.$$

If there are repeated numbers in $(i_1, \cdots, i_{n+1})$, then it is obvious (15.5) holds because both sides would equal zero from the above definition. It remains to verify (15.5) in the case where there are no numbers repeated in $(i_1, \cdots, i_{n+1})$. Consider

$$\operatorname{sgn}_{n+1} \left(i_1, \cdots, \overset{r}{p}, \cdots, \overset{s}{q}, \cdots, i_{n+1} \right),$$

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

$$\operatorname{sgn}_{n+1} \left(i_1, \cdots, \overset{r}{p}, \cdots, \overset{\theta}{n+1}, \cdots, \overset{s}{q}, \cdots, i_{n+1} \right) \equiv$$

$$(-1)^{n+1-\theta} \operatorname{sgn}_n \left(i_1, \cdots, \overset{r}{p}, \cdots, \overset{s-1}{q}, \cdots, i_{n+1} \right)$$

while

$$\operatorname{sgn}_{n+1} \left(i_1, \cdots, \overset{r}{q}, \cdots, \overset{\theta}{n+1}, \cdots, \overset{s}{p}, \cdots, i_{n+1} \right) =$$

$$(-1)^{n+1-\theta} \operatorname{sgn}_n \left(i_1, \cdots, \overset{r}{q}, \cdots, \overset{s-1}{p}, \cdots, i_{n+1} \right)$$

and so, by induction, a switch of p and q introduces a minus sign in the result. Similarly, if $\theta > s$ or if $\theta < r$ it also follows that (15.5) 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.

$$\operatorname{sgn}_{n+1} \left(i_1, \cdots, \overset{r}{n+1}, \cdots, \overset{s}{q}, \cdots, i_{n+1} \right) =$$

$$(-1)^{n+1-r}\operatorname{sgn}_n\left(i_1,\cdots,\overset{s-1}{q},\cdots,i_{n+1}\right) \tag{15.7}$$

while

$$\operatorname{sgn}_{n+1}\left(i_1,\cdots,\overset{r}{q},\cdots,n\overset{s}{+}1,\cdots,i_{n+1}\right)=$$
$$(-1)^{n+1-s}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right). \tag{15.8}$$

By making $s-1-r$ switches, move the q which is in the $s-1^{th}$ position in (15.7) to the r^{th} position in (15.8). By induction, each of these switches introduces a factor of -1 and so

$$\operatorname{sgn}_n\left(i_1,\cdots,\overset{s-1}{q},\cdots,i_{n+1}\right)=(-1)^{s-1-r}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right).$$

Therefore,

$$\operatorname{sgn}_{n+1}\left(i_1,\cdots,n\overset{r}{+}1,\cdots,\overset{s}{q},\cdots,i_{n+1}\right)=(-1)^{n+1-r}\operatorname{sgn}_n\left(i_1,\cdots,\overset{s-1}{q},\cdots,i_{n+1}\right)$$
$$=(-1)^{n+1-r}(-1)^{s-1-r}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right)$$
$$=(-1)^{n+s}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right)=(-1)^{2s-1}(-1)^{n+1-s}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right)$$
$$=-\operatorname{sgn}_{n+1}\left(i_1,\cdots,\overset{r}{q},\cdots,n\overset{s}{+}1,\cdots,i_{n+1}\right).$$

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 (15.4) and (15.5), 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 (15.5) gives both functions are equal to zero for that ordered list. This proves the lemma.

In what follows sgn will often be used rather than sgn_n because the context supplies the appropriate n.

Definition 15.4.2 *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 15.4.3 *Let $(a_{ij})=A$ denote an $n\times n$ matrix. The determinant of A, denoted by $\det(A)$ is defined by*

$$\det(A)\equiv\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_n)\,a_{1k_1}\cdots a_{nk_n}$$

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, $\operatorname{sgn}(k_1,\cdots,k_n)=0$ 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^{th} row is the r_k row of the matrix, A. Thus

$$\det(A(r_1, \cdots, r_n)) = \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n} \tag{15.9}$$

and

$$A(1, \cdots, n) = A.$$

Proposition 15.4.4 *Let*

$$(r_1, \cdots, r_n)$$

be an ordered list of numbers from $\{1, \cdots, n\}$. *Then*

$$\operatorname{sgn}(r_1, \cdots, r_n) \det(A)$$

$$= \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{r_1 k_1} \cdots a_{r_n k_n} \tag{15.10}$$

$$= \det(A(r_1, \cdots, r_n)). \tag{15.11}$$

Proof: Let $(1, \cdots, n) = (1, \cdots, r, \cdots s, \cdots, n)$ so $r < s$.

$$\det(A(1, \cdots, r, \cdots, s, \cdots, n)) = \tag{15.12}$$

$$\sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_r, \cdots, k_s, \cdots, k_n) a_{1k_1} \cdots a_{rk_r} \cdots a_{sk_s} \cdots a_{nk_n},$$

and renaming the variables, calling k_s, k_r and k_r, k_s, this equals

$$= \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_s, \cdots, k_r, \cdots, k_n) a_{1k_1} \cdots a_{rk_s} \cdots a_{sk_r} \cdots a_{nk_n}$$

$$= \sum_{(k_1, \cdots, k_n)} -\operatorname{sgn}\left(k_1, \cdots, \overbrace{k_r, \cdots, k_s}^{\text{These got switched}}, \cdots, k_n\right) a_{1k_1} \cdots a_{sk_r} \cdots a_{rk_s} \cdots a_{nk_n}$$

$$= -\det(A(1, \cdots, s, \cdots, r, \cdots, n)). \tag{15.13}$$

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 15.4.1, this implies

$$\det(A(r_1, \cdots, r_n)) = (-1)^p \det(A) = \operatorname{sgn}(r_1, \cdots, r_n) \det(A)$$

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 (15.12) -(15.13) shows that $A(r_1, \cdots, r_n) = 0$ and also $\operatorname{sgn}(r_1, \cdots, r_n) = 0$ so the formula holds in this case also.

Observation 15.4.5 *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 $n(n-1)$ 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.

With the above, it is possible to give a more symmetric description of the determinant from which it will follow that $\det(A)=\det\left(A^T\right)$.

Corollary 15.4.6 *The following formula for $\det(A)$ is valid.*

$$\det(A)=\frac{1}{n!}\cdot$$

$$\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\,a_{r_1k_1}\cdots a_{r_nk_n}. \tag{15.14}$$

And also $\det\left(A^T\right)=\det(A)$ where A^T is the transpose of A. (Recall that for $A^T=\left(a_{ij}^T\right)$, $a_{ij}^T=a_{ji}$.)

Proof: From Proposition 15.4.4, if the r_i are distinct,

$$\det(A)=\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\,a_{r_1k_1}\cdots a_{r_nk_n}.$$

Summing over all ordered lists, $(r_1,\cdots,r_n)$ where the r_i are distinct, (If the r_i are not distinct, $\operatorname{sgn}(r_1,\cdots,r_n)=0$ and so there is no contribution to the sum.)

$$n!\det(A)=$$

$$\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\,a_{r_1k_1}\cdots a_{r_nk_n}.$$

This proves the corollary since the formula gives the same number for A as it does for A^T.

Corollary 15.4.7 *If two rows or two columns in an $n\times n$ matrix, A, are switched, the determinant of the resulting matrix equals (-1) times the determinant of the original matrix. If A is an $n\times n$ matrix in which two rows are equal or two columns are equal then $\det(A)=0$. Suppose the i^{th} row of A equals $(xa_1+yb_1,\cdots,xa_n+yb_n)$. Then*

$$\det(A)=x\det(A_1)+y\det(A_2)$$

where the i^{th} row of A_1 is $(a_1,\cdots,a_n)$ and the i^{th} row of A_2 is $(b_1,\cdots,b_n)$, 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 15.4.4 when two rows are switched, the determinant of the resulting matrix is (-1) times the determinant of the original matrix. By Corollary 15.4.6 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,

$$\det(A)=\det\left(A^T\right)=-\det\left(A_1^T\right)=-\det(A_1).$$

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.

$$\begin{aligned}\det(A) &\equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots(xa_{k_i}+yb_{k_i})\cdots a_{nk_n}\\ &= x\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots a_{k_i}\cdots a_{nk_n}\\ &\quad +y\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots b_{k_i}\cdots a_{nk_n}\\ &\equiv x\det(A_1)+y\det(A_2).\end{aligned}$$

The same is true of columns because $\det\left(A^T\right)=\det(A)$ and the rows of A^T are the columns of A.

Definition 15.4.8 *A vector,* $\mathbf{w}$*, is a linear combination of the vectors* $\{\mathbf{v}_1,\cdots,\mathbf{v}_r\}$ *if there exists scalars,* $c_1,\cdots c_r$ *such that* $\mathbf{w}=\sum_{k=1}^r c_k\mathbf{v}_k$*. This is the same as saying* $\mathbf{w}\in \operatorname{span}\{\mathbf{v}_1,\cdots,\mathbf{v}_r\}$.

The following corollary is also of great use.

Corollary 15.4.9 *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 $A=\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{pmatrix}$ 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 15.4.7 you may rearrange the columns to have the n^{th} column a linear combination of the first r columns. Thus $\mathbf{a}_n=\sum_{k=1}^r c_k\mathbf{a}_k$ and so

$$\det(A)=\det\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \sum_{k=1}^r c_k\mathbf{a}_k\end{pmatrix}.$$

By Corollary 15.4.7

$$\det(A)=\sum_{k=1}^r c_k\det\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \mathbf{a}_k\end{pmatrix}=0.$$

The case for rows follows from the fact that $\det(A)=\det\left(A^T\right)$. This proves the corollary.

Recall the following definition of matrix multiplication.

Definition 15.4.10 *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 15.4.11 *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 15.4.4,

$$\det(AB) =$$

$$\begin{aligned}
&\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, c_{1k_1}\cdots c_{nk_n} \\
&= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n) \left(\sum_{r_1} a_{1r_1}b_{r_1k_1}\right)\cdots\left(\sum_{r_n} a_{nr_n}b_{r_nk_n}\right) \\
&= \sum_{(r_1\cdots,r_n)} \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, b_{r_1k_1}\cdots b_{r_nk_n}\left(a_{1r_1}\cdots a_{nr_n}\right) \\
&= \sum_{(r_1\cdots,r_n)} \operatorname{sgn}(r_1\cdots r_n)\, a_{1r_1}\cdots a_{nr_n}\det(B) = \det(A)\det(B).
\end{aligned}$$

This proves the theorem.

Lemma 15.4.12 *Suppose a matrix is of the form*

$$M = \begin{pmatrix} A & * \\ \mathbf{0} & a \end{pmatrix} \tag{15.15}$$

or

$$M = \begin{pmatrix} A & \mathbf{0} \\ * & a \end{pmatrix} \tag{15.16}$$

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 $\mathbf{0}$ 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, $m_{nn} = a$ and $m_{in} = 0$ if $i \neq n$. From the definition of the determinant,

$$\det(M) \equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}_n(k_1,\cdots,k_n)\, m_{1k_1}\cdots m_{nk_n}$$

Letting θ denote the position of n in the ordered list, $(k_1,\cdots,k_n)$ then using the earlier conventions used to prove Lemma 15.4.1, $\det(M)$ equals

$$\sum_{(k_1,\cdots,k_n)} (-1)^{n-\theta} \operatorname{sgn}_{n-1}\left(k_1,\cdots,k_{\theta-1},\overset{\theta}{k_{\theta+1}},\cdots,\overset{n-1}{k_n}\right) m_{1k_1}\cdots m_{nk_n}$$

Now suppose (15.16). 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

$$a \sum_{(k_1,\cdots,k_{n-1})} \operatorname{sgn}_{n-1}(k_1,\cdots k_{n-1})\, m_{1k_1}\cdots m_{(n-1)k_{n-1}} = a\det(A).$$

To get the assertion in the situation of (15.15) use Corollary 15.4.6 and (15.16) to write

$$\det(M) = \det\left(M^T\right) = \det\left(\begin{pmatrix} A^T & \mathbf{0} \\ * & a \end{pmatrix}\right) = a\det\left(A^T\right) = a\det(A).$$

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 15.4.13 *Let $A = (a_{ij})$ be an $n \times n$ matrix. Then a new matrix called the cofactor matrix,* $\operatorname{cof}(A)$ *is defined by* $\operatorname{cof}(A) = (c_{ij})$ *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,* $\operatorname{cof}(A)_{ij}$ *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 15.4.14 *Let A be an $n \times n$ matrix where $n \geq 2$. Then*

$$\det(A) = \sum_{j=1}^{n} a_{ij} \operatorname{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij} \operatorname{cof}(A)_{ij}.$$

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 $(a_{i1}, \cdots, a_{in})$ 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 15.4.7,

$$\det(A) = \sum_{j=1}^{n} \det(B_j)$$

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 $\operatorname{cof}(A)_{ij} \equiv (-1)^{i+j} \det\left(A^{ij}\right)$. At this point, recall that from Proposition 15.4.4, 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 15.4.12,

$$\begin{aligned} \det(B_j) &= (-1)^{n-j} (-1)^{n-i} \det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij} \end{pmatrix}\right) \\ &= (-1)^{i+j} \det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij} \end{pmatrix}\right) = a_{ij} \operatorname{cof}(A)_{ij}. \end{aligned}$$

Therefore,

$$\det(A) = \sum_{j=1}^{n} a_{ij} \operatorname{cof}(A)_{ij}$$

which is the formula for expanding $\det(A)$ along the i^{th} row. Also,

$$\begin{aligned} \det(A) &= \det\left(A^T\right) = \sum_{j=1}^{n} a_{ij}^T \operatorname{cof}\left(A^T\right)_{ij} \\ &= \sum_{j=1}^{n} a_{ji} \operatorname{cof}(A)_{ji} \end{aligned}$$

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 14.1.28 on Page 377.

Theorem 15.4.15 *A^{-1} exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = \left(a_{ij}^{-1}\right)$ where*

$$a_{ij}^{-1} = \det(A)^{-1} \operatorname{cof}(A)_{ji}$$

for $\operatorname{cof}(A)_{ij}$ the ij^{th} cofactor of A.

Proof: By Theorem 15.4.14 and letting $(a_{ir}) = A$, if $\det(A) \neq 0$,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ir} \det(A)^{-1} = \det(A) \det(A)^{-1} = 1.$$

Now consider

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1}$$

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 15.4.7. However, expanding this matrix along the k^{th} column yields

$$0 = \det(B_k) \det(A)^{-1} = \sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1}$$

Summarizing,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1} = \delta_{rk}.$$

Using the other formula in Theorem 15.4.14, and similar reasoning,

$$\sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)_{kj} \det(A)^{-1} = \delta_{rk}$$

This proves that if $\det(A) \neq 0$, then A^{-1} exists with $A^{-1} = \left(a_{ij}^{-1}\right)$, where

$$a_{ij}^{-1} = \operatorname{cof}(A)_{ji} \det(A)^{-1}.$$

Now suppose A^{-1} exists. Then by Theorem 15.4.11,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A) \det\left(A^{-1}\right)$$

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 15.4.16 *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 15.4.11 implies

$$\det B \det A = 1$$

and so $\det A \neq 0$. Therefore from Theorem 15.4.15, A^{-1} exists. Therefore,

$$A^{-1} = (BA) A^{-1} = B\left(AA^{-1}\right) = BI = B.$$

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 15.4.15 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.

In case you are solving a system of equations, $A\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$, it follows that if A^{-1} exists,

$$\mathbf{x} = \left(A^{-1}A\right)\mathbf{x} = A^{-1}\left(A\mathbf{x}\right) = A^{-1}\mathbf{y}$$

thus solving the system. Now in the case that A^{-1} exists, there is a formula for A^{-1} given above. Using this formula,

$$x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \sum_{j=1}^{n} \frac{1}{\det(A)} \operatorname{cof}(A)_{ji}\, y_j.$$

By the formula for the expansion of a determinant along a column,

$$x_i = \frac{1}{\det(A)} \det \begin{pmatrix} * & \cdots & y_1 & \cdots & * \\ \vdots & & \vdots & & \vdots \\ * & \cdots & y_n & \cdots & * \end{pmatrix},$$

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.

Definition 15.4.17 *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.

With this definition, here is a simple corollary of Theorem 15.4.14.

Corollary 15.4.18 *Let M be an upper (lower) triangular matrix. Then $\det(M)$ is obtained by taking the product of the entries on the main diagonal.*

Definition 15.4.19 *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. The **row rank** is defined to be the dimension of the span of the rows. The **column rank** is defined to be the dimension of the span of the columns.*

Theorem 15.4.20 *If A has determinant rank, r, then there exist r rows of the matrix such that every other row is a linear combination of these r rows.*

Proof: Suppose the determinant rank of $A = (a_{ij})$ equals r. If rows and columns are interchanged, the determinant rank of the modified matrix is unchanged. Thus rows and columns can be interchanged to produce an $r \times r$ matrix in the upper left corner of the matrix which has non zero determinant. Now consider the $r+1 \times r+1$ matrix, M,

$$\begin{pmatrix} a_{11} & \cdots & a_{1r} & a_{1p} \\ \vdots & & \vdots & \vdots \\ a_{r1} & \cdots & a_{rr} & a_{rp} \\ a_{l1} & \cdots & a_{lr} & a_{lp} \end{pmatrix}$$

where C will denote the $r \times r$ matrix in the upper left corner which has non zero determinant. I claim $\det(M) = 0$.

There are two cases to consider in verifying this claim. First, suppose $p > r$. Then the claim follows from the assumption that A has determinant rank r. On the other hand, if $p < r$, then the determinant is zero because there are two identical columns. Expand the determinant along the last column and divide by $\det(C)$ to obtain

$$a_{lp} = -\sum_{i=1}^{r} \frac{\operatorname{cof}(M)_{ip}}{\det(C)} a_{ip}.$$

Now note that $\operatorname{cof}(M)_{ip}$ does not depend on p. Therefore the above sum is of the form

$$a_{lp} = \sum_{i=1}^{r} m_i a_{ip}$$

which shows the l^{th} row is a linear combination of the first r rows of A. Since l is arbitrary, this proves the theorem.

Corollary 15.4.21 *The determinant rank equals the row rank.*

Proof: From Theorem 15.4.20, the row rank is no larger than the determinant rank. Could the row rank be smaller than the determinant rank? If so, there exist p rows for $p < r$ such that the span of these p rows equals the row space. But this implies that the $r \times r$ submatrix whose determinant is nonzero also has row rank no larger than p which is impossible if its determinant is to be nonzero because at least one row is a linear combination of the others.

Corollary 15.4.22 *If A has determinant rank, r, then there exist r columns of the matrix such that every other column is a linear combination of these r columns. Also the column rank equals the determinant rank.*

Proof: This follows from the above by considering A^T. The rows of A^T are the columns of A and the determinant rank of A^T and A are the same. Therefore, from Corollary 15.4.21, column rank of A = row rank of A^T = determinant rank of A^T = determinant rank of A.

The following theorem is of fundamental importance and ties together many of the ideas presented above.

Theorem 15.4.23 *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 15.4.20. In particular, it follows that for some m, the m^{th} column is a linear combination of all the others. Thus letting $A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_m & \cdots & \mathbf{a}_n \end{pmatrix}$ where the columns are denoted by $\mathbf{a}_i$, there exists scalars, α_i such that

$$\mathbf{a}_m = \sum_{k \neq m} \alpha_k \mathbf{a}_k.$$

Now consider the column vector, $\mathbf{x} \equiv \begin{pmatrix} \alpha_1 & \cdots & -1 & \cdots & \boldsymbol{\alpha}_n \end{pmatrix}^T$. Then

$$A\mathbf{x} = -\mathbf{a}_m + \sum_{k \neq m} \alpha_k \mathbf{a}_k = \mathbf{0}.$$

Since also $A\mathbf{0} = \mathbf{0}$, 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 $\mathbf{x} \neq \mathbf{0}$ such that

$$A^T\mathbf{x} = \mathbf{0}.$$

Taking the transpose of both sides yields

$$\mathbf{x}^T A = \mathbf{0}$$

where the $\mathbf{0}$ is a $1 \times n$ matrix or row vector. Now if $A\mathbf{y} = \mathbf{x}$, then

$$|\mathbf{x}|^2 = \mathbf{x}^T (A\mathbf{y}) = \left(\mathbf{x}^T A\right) \mathbf{y} = \mathbf{0}\mathbf{y} = 0$$

contrary to $\mathbf{x} \neq \mathbf{0}$. Consequently there can be no $\mathbf{y}$ such that $A\mathbf{y} = \mathbf{x}$ 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 15.4.15 A^{-1} exists and so for every $\mathbf{y} \in \mathbb{F}^n$ there exists a unique $\mathbf{x} \in \mathbb{F}^n$ such that $A\mathbf{x} = \mathbf{y}$. In fact $\mathbf{x} = A^{-1}\mathbf{y}$. Thus A would be onto contrary to 3.). This shows 3.) implies 1.) and proves the theorem.

Corollary 15.4.24 *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.

15.5 The Cayley Hamilton Theorem

Definition 15.5.1 *Let A be an $n \times n$ matrix. The **characteristic polynomial** is defined as*

$$p_A(t) \equiv \det(tI - A).$$

For A a matrix and $p(t) = t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0$, denote by $p(A)$ the matrix defined by

$$p(A) \equiv A^n + a_{n-1}A^{n-1} + \cdots + a_1 A + a_0 I.$$

The explanation for the last term is that A^0 is interpreted as I, the identity matrix.

The Cayley Hamilton theorem states that every matrix satisfies its characteristic equation, that equation defined by $P_A(t) = 0$. It is one of the most important theorems in linear algebra. The following lemma will help with its proof.

Lemma 15.5.2 *Suppose for all $|\lambda|$ large enough,*

$$A_0 + A_1\lambda + \cdots + A_m\lambda^m = 0,$$

where the A_i are $n \times n$ matrices. Then each $A_i = 0$.

Proof: Multiply by λ^{-m} to obtain

$$A_0\lambda^{-m} + A_1\lambda^{-m+1} + \cdots + A_{m-1}\lambda^{-1} + A_m = 0.$$

Now let $|\lambda| \to \infty$ to obtain $A_m = 0$. With this, multiply by λ to obtain

$$A_0\lambda^{-m+1} + A_1\lambda^{-m+2} + \cdots + A_{m-1} = 0.$$

Now let $|\lambda| \to \infty$ to obtain $A_{m-1} = 0$. Continue multiplying by λ and letting $\lambda \to \infty$ to obtain that all the $A_i = 0$. This proves the lemma.

With the lemma, here is a simple corollary.

Corollary 15.5.3 *Let A_i and B_i be $n \times n$ matrices and suppose*

$$A_0 + A_1\lambda + \cdots + A_m\lambda^m = B_0 + B_1\lambda + \cdots + B_m\lambda^m$$

for all $|\lambda|$ large enough. Then $A_i = B_i$ for all i. Consequently if λ is replaced by any $n \times n$ matrix, the two sides will be equal. That is, for C any $n \times n$ matrix,

$$A_0 + A_1C + \cdots + A_mC^m = B_0 + B_1C + \cdots + B_mC^m.$$

Proof: Subtract and use the result of the lemma.

With this preparation, here is a relatively easy proof of the Cayley Hamilton theorem.

Theorem 15.5.4 *Let A be an $n \times n$ matrix and let $p(\lambda) \equiv \det(\lambda I - A)$ be the characteristic polynomial. Then $p(A) = 0$.*

Proof: Let $C(\lambda)$ equal the transpose of the cofactor matrix of $(\lambda I - A)$ for $|\lambda|$ large. (If $|\lambda|$ is large enough, then λ cannot be in the finite list of eigenvalues of A and so for such λ, $(\lambda I - A)^{-1}$ exists.) Therefore, by Theorem 15.4.15

$$C(\lambda) = p(\lambda)(\lambda I - A)^{-1}.$$

Note that each entry in $C(\lambda)$ is a polynomial in λ having degree no more than $n - 1$. Therefore, collecting the terms,

$$C(\lambda) = C_0 + C_1\lambda + \cdots + C_{n-1}\lambda^{n-1}$$

for C_j some $n \times n$ matrix. It follows that for all $|\lambda|$ large enough,

$$(A - \lambda I)\left(C_0 + C_1\lambda + \cdots + C_{n-1}\lambda^{n-1}\right) = p(\lambda) I$$

and so Corollary 15.5.3 may be used. It follows the matrix coefficients corresponding to equal powers of λ are equal on both sides of this equation. Therefore, if λ is replaced with A, the two sides will be equal. Thus

$$0 = (A - A)\left(C_0 + C_1A + \cdots + C_{n-1}A^{n-1}\right) = p(A) I = p(A).$$

This proves the Cayley Hamilton theorem.

15.6 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 $A_1\mathbf{x}$ is the vector,

$$A_1\mathbf{x} = \begin{pmatrix} A\mathbf{x} \\ \mathbf{0} \end{pmatrix}$$

which equals zero if and only if $A\mathbf{x} = \mathbf{0}$.

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 15.4.15 it was claimed that $\det(I) = 1$. Here $I = (\delta_{ij})$. Prove this assertion. Also prove Corollary 15.4.18.

5. Let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ be vectors in $\mathbb{F}^n$ and let $M(\mathbf{v}_1, \cdots, \mathbf{v}_n)$ denote the matrix whose i^{th} column equals $\mathbf{v}_i$. Define

$$d(\mathbf{v}_1, \cdots, \mathbf{v}_n) \equiv \det(M(\mathbf{v}_1, \cdots, \mathbf{v}_n)).$$

Prove that d is linear in each variable, (multilinear), that

$$d(\mathbf{v}_1, \cdots, \mathbf{v}_i, \cdots, \mathbf{v}_j, \cdots, \mathbf{v}_n) = -d(\mathbf{v}_1, \cdots, \mathbf{v}_j, \cdots, \mathbf{v}_i, \cdots, \mathbf{v}_n), \tag{15.17}$$

and

$$d(\mathbf{e}_1, \cdots, \mathbf{e}_n) = 1 \tag{15.18}$$

where here $\mathbf{e}_j$ 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 $f : \mathbb{F}^n \times \cdots \times \mathbb{F}^n \to \mathbb{F}$ satisfies (15.17) and (15.18) 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 $\det(A) = \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n)\, a_{k_1 1} \cdots a_{k_n n}$.

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. Let $Ly = y^{(n)} + a_{n-1}(x)\, y^{(n-1)} + \cdots + a_1(x)\, y' + a_0(x)\, y$ where the a_i are given continuous functions defined on a closed interval, (a,b) and y is some function which has n derivatives so it makes sense to write Ly. Suppose $Ly_k = 0$ for $k = 1, 2, \cdots, n$. The Wronskian of these functions, y_i is defined as

$$W(y_1, \cdots, y_n)(x) \equiv \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n-1)}(x) & \cdots & y_n^{(n-1)}(x) \end{pmatrix}$$

Show that for $W(x) = W(y_1, \cdots, y_n)(x)$ to save space,

$$W'(x) = \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n)}(x) & \cdots & y_n^{(n)}(x) \end{pmatrix}.$$

Now use the differential equation, $Ly = 0$ which is satisfied by each of these functions, y_i and properties of determinants presented above to verify that $W' + a_{n-1}(x)\, W = 0$. Give an explicit solution of this linear differential equation, Abel's formula, and use your answer to verify that the Wronskian of these solutions to the equation, $Ly = 0$ either vanishes identically on (a,b) or never.

12. 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.

13. Suppose the characteristic polynomial of an $n \times n$ matrix, A is of the form

$$t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0$$

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 $a_0 \neq 0$.

14. In constitutive modeling of the stress and strain tensors, one sometimes considers sums of the form $\sum_{k=0}^{\infty} a_k A^k$ where A is a 3×3 matrix. Show using the Cayley Hamilton theorem that if such a thing makes any sense, you can always obtain it as a finite sum having no more than n terms.

Vector Spaces

16.0.1 Outcomes

1. Define vector space.
2. Define the span of a set of vectors. Recall that a span of vectors in a vector space is a subspace.
3. Determine whether a set of vectors is a subspace.
4. Define linear independence.
5. Determine whether a set of vectors is linearly independent or linearly dependent.
6. Determine a basis and the dimension of a vector space.

16.1 Vector Spaces

The symbol, $\mathbb{R}^n$ denotes the set of $n \times 1$ matrices which have all real entries. Recall that these are also called column vectors or just vectors for short. The symbol, $\mathbb{C}^n$ denotes the set of $n \times 1$ matrices which have complex entries. Thus an example of something in $\mathbb{C}^3$ is

$$\begin{pmatrix} 1+i \\ 2 \\ 3-2i \end{pmatrix}.$$

Since every real number may be considered a complex number, it follows that every vector in $\mathbb{R}^n$ is a vector in $\mathbb{C}^n$. You will have to use $\mathbb{C}^n$ when you study differential equations. These two examples will be sufficient for many applications but not for all. It turns out that all algebraic considerations are the same for either $\mathbb{R}^n$ or $\mathbb{C}^n$ and so to avoid fussing with unenlightening details, we will denote by $\mathbb{F}^n$ the set of $n \times 1$ matrices. Here $\mathbb{F}$ will stand for either $\mathbb{R}$ or $\mathbb{C}$.

16.1.1 Vector Space Axioms

The concept of a vector space turns out to be a significant unifying idea in many different subjects. For example, $\mathbb{F}^n$, the $n \times 1$ matrices (column vectors) will turn out to be a vector space. So is the collection of all $m \times n$ matrices. In differential equations, you will see another example of a vector space consisting of certain collections of functions. Solutions to linear systems in any context turn out to involve vector spaces.

Definition 16.1.1 *A **vector space**, V is a nonempty set of objects on which are defined two operations, **addition** and **multiplication by scalars**, real or complex numbers*[1]*, satisfying the properties listed below.*

1. *Closure under addition.*
$$\text{If } \mathbf{u}, \mathbf{v} \in V, \text{ then } \mathbf{u} + \mathbf{v} \in V.$$

2. *Closure under scalar multiplication.*
$$\text{If } \alpha \text{ is a scalar and } \mathbf{u} \in V, \text{ then } \alpha\mathbf{u} \in V.$$

3. *The following eight **vector space axioms.***

- ***Commutative Law Of Addition.*** *For $\mathbf{a}, \mathbf{b} \in V$*
$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}, \tag{16.1}$$

- ***Associative Law for Addition.*** *For $\mathbf{a}, \mathbf{b}, \mathbf{c} \in V$,*
$$(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c}), \tag{16.2}$$

- ***Existence of an Additive Identity.*** *There exists $\mathbf{0} \in V$ such that for all $\mathbf{a} \in V$,*
$$\mathbf{a} + \mathbf{0} = \mathbf{a}, \tag{16.3}$$

- ***Existence of an Additive Inverse.*** *For each $\mathbf{a} \in V$, there exists $-\mathbf{a} \in V$ such that*
$$\mathbf{a} + (-\mathbf{a}) = 0. \tag{16.4}$$

- ***Distributive law over Vector Addition.*** *For any scalar, α and any two vectors, $\mathbf{a}, \mathbf{b} \in V$,*
$$\alpha(\mathbf{a} + \mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}, \tag{16.5}$$

- ***Distributive law over Scalar Addition.*** *For any two scalars, α, β and any vector $\mathbf{a} \in V$,*
$$(\alpha + \beta)\mathbf{a} = \alpha\mathbf{a} + \beta\mathbf{a}, \tag{16.6}$$

- ***Associative law for Scalar Multiplication.*** *For any two scalars, α, β and any vector $\mathbf{a} \in V$,*
$$\alpha(\beta\mathbf{a}) = \alpha\beta(\mathbf{a}), \tag{16.7}$$

- ***Rule for Multiplication by*** 1. *For any vector $\mathbf{a} \in V$,*
$$1\mathbf{a} = \mathbf{a}. \tag{16.8}$$

Theorem 16.1.2 *Let X be any vector space.*

1. *Then the additive identity is unique.*

2. *For each $\mathbf{u} \in X$, the additive inverse $-\mathbf{u}$ is unique.*

[1]For this book, the field of scalars will always be either $\mathbb{R}$ or $\mathbb{C}$ but there are many other fields used. Fields are simply algebraic objects which have the same algebraic properties as the real or complex numbers.

3. *If* $\mathbf{u} \in X$, *then* $0\mathbf{u} = \mathbf{0}$. *That is, the scalar 0 times the vector* $\mathbf{u}$ *gives the vector* $\mathbf{0}$.

4. *For any* $\mathbf{u} \in X, -\mathbf{u} = (-1)\,\mathbf{u}$.

Proof: To see the additive identity is unique, suppose $\mathbf{0}'$ is another one. Then from the properties satisfied by the additive identities,

$$\mathbf{0}' = \mathbf{0}' + \mathbf{0} = \mathbf{0}.$$

This verifies the first claim.

Suppose $\mathbf{v}$ is another additive inverse for $\mathbf{u}$. Then

$$-\mathbf{u} + \mathbf{u} = \mathbf{0},\ \mathbf{v} + \mathbf{u} = \mathbf{0}$$

and so

$$-\mathbf{u} + \mathbf{u} = \mathbf{v} + \mathbf{u}.$$

Add $-\mathbf{u}$ to both sides as follows

$$(-\mathbf{u} + \mathbf{u}) + (-\mathbf{u}) = (\mathbf{v} + \mathbf{u}) + (-\mathbf{u}).$$

By the associative law of addition,

$$-\mathbf{u} + (\mathbf{u} + (-\mathbf{u})) = \mathbf{v} + (\mathbf{u} + (-\mathbf{u}))$$

and so

$$-\mathbf{u} = -\mathbf{u} + \mathbf{0} = \mathbf{v} + \mathbf{0} = \mathbf{v}.$$

This verifies the second assertion.

By the distributive law,

$$0\mathbf{u} = (0 + 0)\,\mathbf{u} = 0\mathbf{u} + 0\mathbf{u}.$$

Now by the existence of the additive inverse, there exists an additive inverse to $0\mathbf{u}$. Add it to both sides and then use the associative law for addition to write

$$\begin{aligned} \mathbf{0} &= (-(0\mathbf{u})) + 0\mathbf{u} = (-(0\mathbf{u})) + (0\mathbf{u} + 0\mathbf{u}) \\ &= (-(0\mathbf{u}) + 0\mathbf{u}) + 0\mathbf{u} = \mathbf{0} + 0\mathbf{u} = 0\mathbf{u}. \end{aligned}$$

This verifies the third assertion.

To verify the fourth, use the distributive law and the third assertion which was just proved to write

$$\mathbf{u} + (-1)\,\mathbf{u} = (1 + (-1))\,\mathbf{u} = 0\mathbf{u} = \mathbf{0}.$$

Therefore, $(-1)\,\mathbf{u}$ acts like the additive inverse of $\mathbf{u}$. By the second assertion which was just proved, it follows $(-1)\,\mathbf{u} = -\mathbf{u}$.

What follows is a list of examples. You will benefit greatly from verifying the claim that each is a vector space.

Example 16.1.3 $\mathbb{F}^n$ *is a vector space with respect to the scalar multiplication defined earlier. Recall the properties of matrix addition and multiplication by scalars.*

Example 16.1.4 *The set of* $m \times n$ *matrices is a vector space.*

Example 16.1.5 *Consider the space of real valued functions defined on some set, D. If f, g are two such functions then $f+g$ is the function defined by $(f+g)(x) = f(x)+g(x)$. Also for α a real number, αf is the function defined by $\alpha f(x) = \alpha(f(x))$. With this understanding of addition and scalar multiplication, this set of functions is a vector space, the field of scalars being $\mathbb{R}$.*

Example 16.1.6 *Consider the space of complex valued functions defined on some set, D. If f, g are two such functions then $f+g$ is the function defined by $(f+g)(x) = f(x)+g(x)$. Also for α a complex number, αf is the function defined by $\alpha f(x) = \alpha(f(x))$. With this understanding of addition and scalar multiplication, this set of functions is a vector space. Note that in this case the field of scalars is $\mathbb{C}$.*

Example 16.1.7 *Consider the space of functions defined on $\mathbb{R}$ which have a continuous derivative. This is a vector space.*

Example 16.1.8 *Consider the space of functions defined on $[0,1]$ which have a Riemann integral. This is a vector space.*

Example 16.1.9 *The space of functions, f defined on an interval, $[a,b]$ which satisfy a condition of the form*

$$|f(x) - f(y)| \leq C|x-y|^{\alpha}$$

for $\alpha \in (0,1)$ is a very important example of a vector space of functions.

Example 16.1.10 *Consider the polynomials defined on $[0,1]$ having degree no larger than 3. If you add two such functions, you get another such function and if you multiply one by a scalar, you get another one also. Also all the vector space axioms hold for this set of functions so this is also a vector space.*

Example 16.1.11 *Consider functions of the form $a\sin x + b\cos x$ where a, b are real numbers. This is a vector space also. Things like this will be important in beginning courses in differential equations.*

Example 16.1.12 *This example is a little more exotic than the above and we won't need it in what follows but if you are interested, let your field of scalars be the rational numbers and let your vectors be numbers of the form $a + b\sqrt{5}$ where a, b are rational numbers. You can verify this is also a vector space.*

16.1.2 Spans

An important concept in the applications of vector spaces is that of linear combinations and spans.

Definition 16.1.13 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ be vectors in a vector space, X. A **linear combination** is any expression of the form*

$$c_1\mathbf{x}_1 + \cdots + c_p\mathbf{x}_p = \sum_{i=1}^{p} c_i\mathbf{x}_i$$

where the c_i are scalars.

Example 16.1.14 *The vector,* $\begin{pmatrix} 7 \\ 4 \end{pmatrix}$ *is a linear combination of the vectors* $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ *and* $\begin{pmatrix} 5 \\ 0 \end{pmatrix}$ *because*

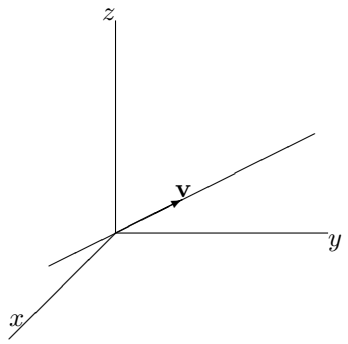
$$2\begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 5 \\ 0 \end{pmatrix} = \begin{pmatrix} 7 \\ 4 \end{pmatrix}.$$

Definition 16.1.15 *The set of all linear combinations of* $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ *is called the* ***span*** *and is written as*

$$\text{span}\left(\mathbf{x}_1, \cdots, \mathbf{x}_n\right).$$

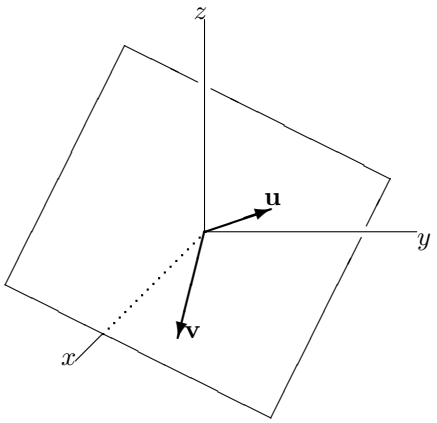
You can consider the span of any number of vectors.

Example 16.1.16 *Consider the span of one vector in* $\mathbb{R}^3$.

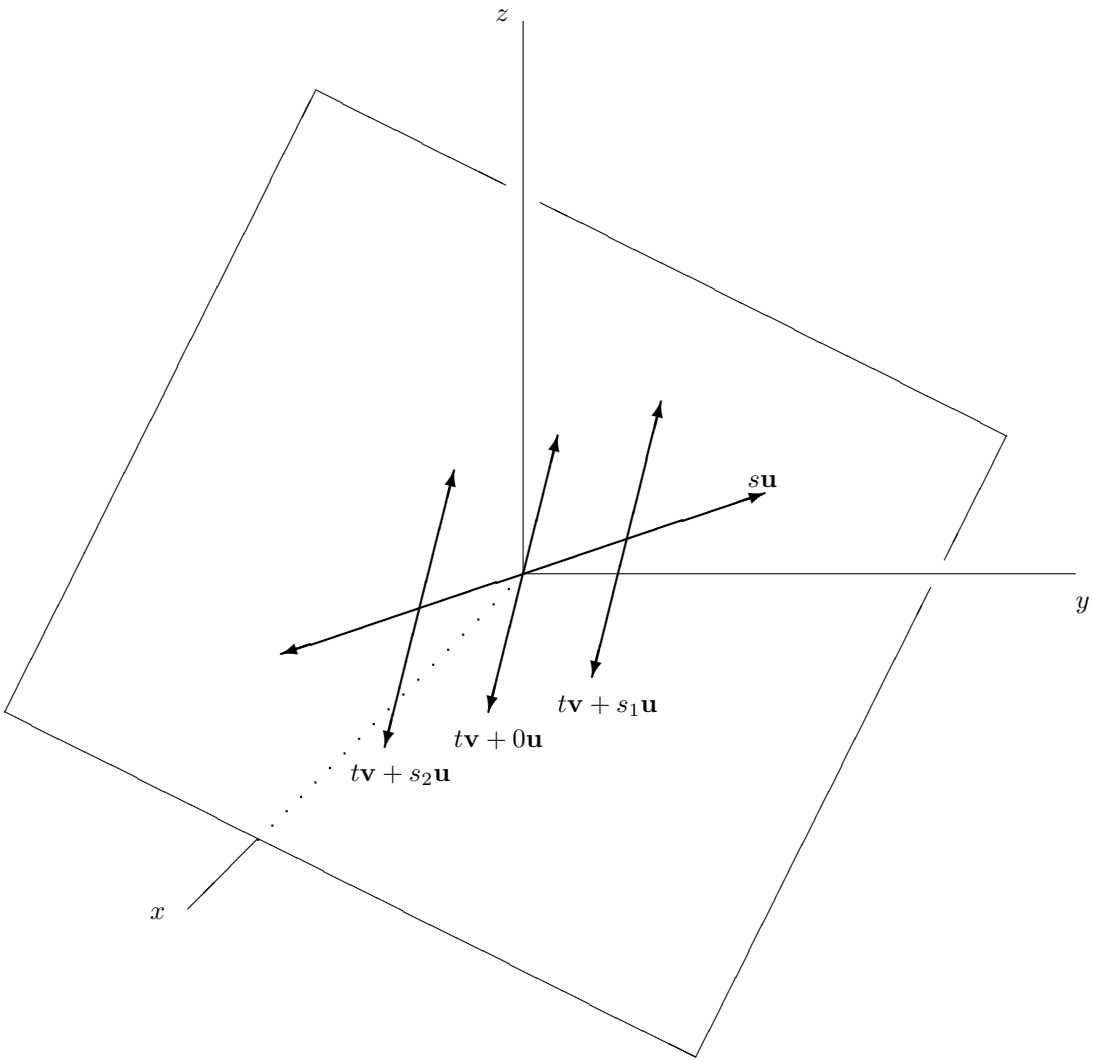


You see there is a vector, $\mathbf{v}$ and the span of this single vector, $\{t\mathbf{v}$ such that $t \in \mathbb{R}\}$ gives the indicated line which goes through the origin, $(0,0,0)$ having $\mathbf{v}$ as a direction vector.

Example 16.1.17 *You can get an idea of the appearance of the span of two vectors in* $\mathbb{R}^3$. *These are just planes which pass through the origin. Here is a picture.*



Lets consider why the displayed plane really is the span of the two vectors which lie in this plane as shown.



As indicated in the above picture, a typical thing in the span of these two vectors is of the form $s\mathbf{u}+t\mathbf{v}$ where s and t are real numbers. By specifying s, you determine a point on the line through the origin, $(0,0,0)$ having direction vector, $\mathbf{u}$. Then through this point, there is a line having direction vector, $\mathbf{v}$. We have drawn three such lines in the above picture, one for $s=0$, s_1, and s_2. The totality of all such lines yields the span of the two vectors, $\mathbf{u}$ and $\mathbf{v}$ and you see from geometric considerations it is just a plane.

It is important that neither of the two vectors $\mathbf{u}$ and $\mathbf{v}$ be a multiple of the other in order for a plane like the one shown to be obtained. If $\mathbf{v}$ had pointed in the same direction or opposite direction as $\mathbf{u}$, you see that the span of these two vectors would reduce to nothing more than the line $s\mathbf{u}$ where $s \in \mathbb{R}$. Also if one of these is a multiple of the other, then the vector, $\mathbf{n}$ given above would equal $\mathbf{0}$ and so the set of vectors perpendicular to this vector $\mathbf{n}$ would not equal a plane.

Some examples have absolutely nothing to do with geometry.

Example 16.1.18 *Let X denote the set of functions defined on an interval, (a,b). Let $f_0(x)=1, f_1(x)=x$, and $f_2(x)=x^2$. Find* $\operatorname{span}(f_0,f_1,f_2)$.

Something in the span is of the form $af_0+bf_1+cf_2$ and $(af_0+bf_1+cf_2)(x)=a+bx+cx^2$ so you see that the span of these vectors consists polynomials of degree no more than 2.

16.1.3 Subspaces

Any time you are dealing with a vector space, there is a concept of **subspace.**

Definition 16.1.19 *Let X be a vector space and let V be a collection of vectors of X. Then V is called a **subspace** of X if V is a vector space contained in X with respect to the same vector addition and scalar multiplication.*

Proposition 16.1.20 *Let X be a vector space and let V be a collection of vectors of X. Then V is a subspace of X if and only if whenever α, β are scalars and $\mathbf{u}$ and $\mathbf{v}$ are vectors of V, it follows $\alpha\mathbf{u}+\beta\mathbf{v} \in V$. That is, V is "**closed** under the algebraic operations of vector addition and scalar multiplication".*

Proof: Suppose first that V is a subspace of X. This means that V is itself a vector space. Therefore, if α, β are scalars and $\mathbf{u}, \mathbf{v}$ are vectors in V it follows that $\alpha\mathbf{u} + \beta\mathbf{v}$ is a vector of V from the axioms of a vector space given in Definition 16.1.1 on Page 420.

Next suppose the condition involving V being closed with respect to the vector space operations holds. Then it follows V must be a vector space because the other eight vector space axioms all are valid because they are valid for all vectors from X and so in particular they hold for vectors in V because every vector in V is given to be in X. Therefore, V is a vector space. This proves the proposition.

Example 16.1.21 *For example, in $\mathbb{R}^3$, the subspaces are the zero vector, $\{\mathbf{0}\}$, any line through the origin, or any plane through the origin or all of $\mathbb{R}^3$. It follows from some of the theorems presented below that these are the only examples of subspaces in $\mathbb{R}^3$.*

Proposition 16.1.22 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ be vectors in a vector space, X. Then $\operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is a subspace.*

Proof: It is suffices to verify the condition of closure with respect to the vector space operations found in Proposition 16.1.20. Let $\mathbf{u}, \mathbf{v} \in \operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_p)$. This means there exist scalars, $\alpha_1, \cdots, \alpha_p$ and $\beta_1, \cdots, \beta_p$ such that

$$\mathbf{u} = \alpha_1\mathbf{x}_1 + \cdots + \alpha_p\mathbf{x}_p \equiv \sum_{i=1}^{p} \alpha_i\mathbf{x}_i, \ \mathbf{v} = \sum_{i=1}^{p} \beta_i\mathbf{x}_i.$$

If a, b are two scalars, we need to verify that $a\mathbf{u} + b\mathbf{v} \in \operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_p)$. But

$$\begin{aligned} a\mathbf{u} + b\mathbf{v} &= a\sum_{i=1}^{p} \alpha_i\mathbf{x}_i + b\sum_{i=1}^{p} \beta_i\mathbf{x}_i \\ &= \sum_{i=1}^{p} (a\alpha_i + b\beta_i)\mathbf{x}_i \\ &\in \operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_p). \end{aligned}$$

This verifies the assertion of this proposition.

You should note that there are only finitely many vectors in $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ but there will likely be infinitely many vectors in $\operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_p)$ so don't confuse the two.

When S is any subset of a vector space, X, the term $\operatorname{span}(S)$ denotes the set of finite linear combinations of vectors of S. Suppose V is a subspace of X. Then $\operatorname{span}(V) = V$ from the conclusion of Proposition 16.1.20. Therefore, every subspace is the span of some set of vectors. However, it is much more desirable to find a minimal set of vectors, hopefully a finite set, which also spans the subspace. We will address this issue in Section 16.1.5.

Definition 16.1.23 *When* $V = \text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_p)$, *the set of vectors,* $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ *is called a* ***spanning set*** *for* V.

Example 16.1.24 *Let* $X = \mathbb{F}^n$ *and let* $V = \{\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{F}^n : x_n = 0\}$. *Is* V *a subspace?*

You have to verify that if a and b are scalars and $\mathbf{x}, \mathbf{y}$ vectors in V, then the linear combination $a\mathbf{x} + b\mathbf{y}$ is in V. Since $\mathbf{x}, \mathbf{y}$ are in V, it follows $\mathbf{x} = (x_1, \cdots, x_{n-1}, 0)$ and $\mathbf{y} = (y_1, \cdots, y_{n-1}, 0)$. Therefore,

$$\begin{aligned} a\mathbf{x} + b\mathbf{y} &= a(x_1, \cdots, x_{n-1}, 0) + b(y_1, \cdots, y_{n-1}, 0) \\ &= (ax_1 + by_1, \cdots, ax_{n-1} + by_{n-1}, 0) \end{aligned}$$

which is a vector of V. Therefore, V is a subspace.

Example 16.1.25 *Let* $X = \mathbb{F}^n$ *and let* $V = \{\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{F}^n : x_n \geq 0\}$. *Is* V *a subspace?*

In this case, V is not a subspace because $(x_1, \cdots, x_{n-1}, 1)$ is a vector of V but

$$(-1)(x_1, \cdots, x_{n-1}, 1) = (-x_1, \cdots, -x_{n-1}, -1)$$

and this vector is not in V because the number in the n^{th} slot is negative, not nonnegative. Thus V is not closed with respect to scalar multiplication so it cannot be a subspace.

Example 16.1.26 *Let* X *be the functions defined on* $[0, 1]$ *and let* V *be those functions which are polynomials of degree 3 or less. Is* V *a subspace?*

The answer is yes because if you add two polynomials of degree 3 or less, you get a polynomial of degree three or less. If you multiply such a polynomial by a scalar, you get another such polynomial.

16.1.4 Linear Independence

Probably the most important concept in linear algebra is that of linear independence.

Definition 16.1.27 *A set of vectors* $\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ *if* ***linearly independent*** *if the vector equation,*

$$c_1\mathbf{v}_1 + \cdots + c_p\mathbf{v}_p = \mathbf{0}$$

has only the ***trivial solution****,*

$$c_1 = c_2 = \cdots = c_p = 0.$$

Otherwise the set of vectors is said to be ***dependent****.*

Theorem 16.1.28 *A set of vectors* $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ *is linearly independent if and only if none of the vectors can be obtained as a linear combination of the others.*

Proof: Suppose first that $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ is linearly independent. If $\mathbf{x}_k = \sum_{j \neq k} c_j \mathbf{x}_j$, then

$$\mathbf{0} = 1\mathbf{x}_k + \sum_{j \neq k} (-c_j)\mathbf{x}_j,$$

a nontrivial linear combination, contrary to assumption. This shows that if the set is linearly independent, then none of the vectors is a linear combination of the others.

Now suppose no vector is a linear combination of the others. It is desired to show that $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ is linearly independent so suppose it is not. Then there exist scalars, c_i, not all zero such that

$$\sum_{i=1}^{p} c_i \mathbf{x}_i = \mathbf{0}.$$

Say $c_k \neq 0$. Then you can solve for $\mathbf{x}_k$ as

$$\mathbf{x}_k = \sum_{j \neq k} \left(-\frac{c_j}{c_k} \right) \mathbf{x}_j$$

contrary to assumption. This proves the lemma.

Restating this theorem in terms of dependent sets is useful.

Corollary 16.1.29 *A set of vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ is linearly dependent if and only if one of the vectors can be obtained as a linear combination of the others.*

You can consider this vector which is a linear combination of the other vectors as dependent on the others. However, the precise meaning of dependent and independent pertains to a set of vectors, not an individual vector.

Example 16.1.30 *Let X denote the vector space of all functions defined on $[0,1]$. Let $f_0(x) = 1, f_1(x) = x$, and $f_2(x) = x^2$. Is the set of vectors, $\{f_0, f_1, f_2\}$ independent or dependent?*

Suppose $af_0 + bf_1 + cf_2 = 0$. Here the 0 refers to the 0 function, that function which sends every x to 0. Thus

$$a + bx + cx^2 = 0$$

for all $x \in [0,1]$. Then since this holds for all such x, you could assign the value 0 to x and conclude that $a = 0$. Thus

$$bx + cx^2 = 0.$$

Now since this holds for every $x \in [0,1]$, you could take the derivative of both sides and obtain

$$b + 2cx = 0.$$

Now assign x the value 0 and conclude that $b = 0$. Hence $2cx = 0$. Let $x = 1$ and conclude $c = 0$ also. Therefore, $\{f_0, f_1, f_2\}$ is linearly independent.

Example 16.1.31 *Consider the three vectors $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$. Are the vectors dependent or independent?*

To find whether they are independent, you must determine whether there are non zero solutions, x, y, and z to

$$x \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + y \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{16.9}$$

This is equivalent to finding whether there are solutions to the system

$$\begin{array}{c} x + 2y = 0 \\ y + z = 0 \\ x + z = 0 \end{array}.$$

The augmented matrix for this system of equations is

$$\left(\begin{array}{ccc|c} 1 & 2 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{array}\right).$$

Taking -1 times the top row and adding to the bottom,

$$\left(\begin{array}{ccc|c} 1 & 2 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -2 & 1 & 0 \end{array}\right).$$

Now taking 2 times the second row and adding to the bottom yields

$$\left(\begin{array}{ccc|c} 1 & 2 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & -1 & 0 \end{array}\right)$$

and so you can see at this point that the only solution to the system (16.9) is the solution, $x = y = z = 0$. Thus the vectors are linearly independent.

The following is called the **exchange theorem**. It is the fundamental idea upon which every significant idea in linear algebra is based. In particular it is essential for the next section. The proof is presented later on Page 433. Here is the statement.

Theorem 16.1.32 *(Exchange Theorem) Let* $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ *be a linearly independent set of vectors such that each* $\mathbf{x}_i$ *is in span*$(\mathbf{y}_1, \cdots, \mathbf{y}_s)$. *Then* $r \leq s$.

In words and with slightly less precision it says that a spanning set has at least as many vectors as a linearly independent set.

16.1.5 Basis And Dimension

Example 16.1.33 *Consider the three vectors,* $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}$ *in* $\mathbb{R}^3$. *Are they dependent or independent?*

These three vectors are dependent because the third is the sum of the first two. The significance of this is that you can throw out the third of these vectors and obtain a smaller list of vectors which has the same span as follows.

$$\begin{aligned} & a\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + b\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + c\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} \\ = \; & a\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + b\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + c\left(\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}\right) \\ = \; & (a+c)\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + (b+c)\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \end{aligned}$$

which is in the span of the first two vectors.

This is like the situation with the span of two vectors in which one was a multiple of the other. The span of the two was the same as the span of one of them. Geometrically, this yielded a line rather than a plane even though two vectors were listed.

When you have a dependent set of vectors, you can always throw out some, obtaining a smaller list which has the same span as the original list of vectors. The concept of a basis is related to this.

Definition 16.1.34 *Let V be a vector space. A finite set of vectors, $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is a **basis** for V if $\mathrm{span}(\mathbf{x}_1, \cdots, \mathbf{x}_r) = V$ and $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is linearly independent.*

Example 16.1.35 *Here are three vectors. Determine whether they are a basis for $\mathbb{R}^3$.*

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}$$

First check for linear independence. If

$$c_1 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} + c_3 \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

does it follow that c_1, c_2, c_3 are all equal to zero? In other words, is the only solution to the following system of equations the zero solution?

$$\begin{aligned} c_1 + 2c_2 + 3c_3 &= 0 \\ c_1 + 0c_2 + c_3 &= 0 \\ c_1 + c_2 + 0c_3 &= 0 \end{aligned}$$

You know how to solve such systems by now. When you do so, you find the only solution is $c_2 = 0, c_1 = 0, c_3 = 0$. Therefore, these vectors are linearly independent.

Next check whether the vectors span $\mathbb{R}^3$. In other words, for any choice of x, y, z, there must be constants, c_1, c_2, c_3 such that

$$c_1 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} + c_3 \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

This is equivalent to determining whether there exists a solution to the system of equations,

$$\begin{pmatrix} c_1 + 2c_2 + 3c_3 = x \\ c_1 + 0c_2 + c_3 = y \\ c_1 + c_2 + 0c_3 = z \end{pmatrix}$$

for any choice of x, y, z. In terms of matrices, this is equivalent to finding a solution to

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

This can always be done in this case because the matrix on the left has an inverse. You know how to find the inverse of a matrix now. Its inverse is

$$\begin{pmatrix} -\frac{1}{4} & \frac{3}{4} & \frac{1}{2} \\ \frac{1}{4} & -\frac{3}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{2} \end{pmatrix}.$$

Multiplying both sides by this inverse matrix you find the solution to above system of equations is

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} -\frac{1}{4} & \frac{3}{4} & \frac{1}{2} \\ \frac{1}{4} & -\frac{3}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\frac{1}{4}x + \frac{3}{4}y + \frac{1}{2}z \\ \frac{1}{4}x - \frac{3}{4}y + \frac{1}{2}z \\ \frac{1}{4}x + \frac{1}{4}y - \frac{1}{2}z \end{pmatrix}.$$

Since there exists such a solution, it follows the span of these vectors is the whole space and they are therefore a basis.

Corollary 16.1.36 *Let* $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ *and* $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$ *be two bases*[2] *of* $\mathbb{F}^n$. *Then* $r = s = n$.

Proof: From the exchange theorem, $r \leq s$ and $s \leq r$. Now note the vectors,

$$\mathbf{e}_i = \overbrace{(0, \cdots, 0, 1, 0 \cdots, 0)^T}^{1 \text{ is in the } i^{th} \text{ slot}}$$

for $i = 1, 2, \cdots, n$ are a basis for $\mathbb{F}^n$. This proves the corollary.

There are many bases for $\mathbb{F}^n$.

Example 16.1.37 *The vectors,*

$$\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}$$

form a basis for $\mathbb{R}^3$. *So do the vectors*

$$\begin{pmatrix} 1 \\ 7 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}.$$

You can verify this as in Example 16.1.35.

The following definition is actually included in the earlier definition of a basis but we list it here for the sake of emphasis.

Definition 16.1.38 *A finite set of vectors,* $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ *is a basis for a subspace,* V *of* $\mathbb{F}^n$ *if* $\operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_r) = V$ *and* $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ *is linearly independent.*

Corollary 16.1.39 *Let* $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ *and* $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$ *be two bases for* V. *Then* $r = s$.

Proof: From the exchange theorem, $r \leq s$ and $s \leq r$. Therefore, this proves the corollary.

Definition 16.1.40 *Let* V *be a subspace of* $\mathbb{F}^n$. *Then* $\dim(V)$ *read as the dimension of* V *is the number of vectors in a basis.*

Of course you should wonder right now whether an arbitrary subspace of $\mathbb{F}^n$ even has a basis. In fact it does and this is in the next theorem. First, here is an important lemma.

[2]This is the plural form of basis. We could say basiss but it would involve an inordinate amount of hissing as in "The sixth shiek's sixth sheep is sick". This is the reason that bases is used instead of basiss.

Lemma 16.1.41 *Suppose* $\mathbf{v} \notin \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ *and* $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ *is linearly independent. Then* $\{\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{v}\}$ *is also linearly independent.*

Proof: Suppose $\sum_{i=1}^{k} c_i \mathbf{u}_i + d\mathbf{v} = \mathbf{0}$. It is required to verify that each $c_i = 0$ and that $d = 0$. But if $d \neq 0$, then you can solve for $\mathbf{v}$ as a linear combination of the vectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$,

$$\mathbf{v} = -\sum_{i=1}^{k} \left(\frac{c_i}{d}\right) \mathbf{u}_i$$

contrary to assumption. Therefore, $d = 0$. But then $\sum_{i=1}^{k} c_i \mathbf{u}_i = 0$ and the linear independence of $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ implies each $c_i = 0$ also. This proves the lemma.

Theorem 16.1.42 *Let* V *be a nonzero subspace of* $\mathbb{F}^n$. *Then* V *has a basis.*

Proof: Let $\mathbf{v}_1 \in V$ where $\mathbf{v}_1 \neq 0$. If $\operatorname{span}\{\mathbf{v}_1\} = V$, stop. $\{\mathbf{v}_1\}$ is a basis for V. Otherwise, there exists $\mathbf{v}_2 \in V$ which is not in $\operatorname{span}\{\mathbf{v}_1\}$. By Lemma 16.1.41 $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a linearly independent set of vectors. If $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\} = V$ stop, $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a basis for V. If $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\} \neq V$, then there exists $\mathbf{v}_3 \notin \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a larger linearly independent set of vectors. Continuing this way, the process must stop before $n+1$ steps because if not, it would be possible to obtain $n+1$ linearly independent vectors contrary to the exchange theorem. This proves the theorem.

Example 16.1.43 *Consider the plane* $2x + 3y + z = 0$. *Show this is a subspace and find a basis for the subspace.*

When we write the equation, $2x + 3y + z = 0$, we mean the set of all vectors, (x, y, z) such that $2x + 3y + z = 0$. Why is this a subspace? Suppose α, β are scalars and (x_1, y_1, z_1) and (x_2, y_2, z_2) are two vectors satisfying the condition determined by the equation. Does $\alpha(x_1, y_1, z_1) + \beta(x_2, y_2, z_2)$ also satisfy the condition defined by the equation?

$$\alpha(x_1, y_1, z_1) + \beta(x_2, y_2, z_2) = (\alpha x_1 + \beta x_2, \alpha y_1 + \beta y_2, \alpha z_1 + \beta z_2).$$

$$\begin{aligned} & 2(\alpha x_1 + \beta x_2) + 3(\alpha y_1 + \beta y_2) + (\alpha z_1 + \beta z_2) \\ = \quad & \alpha(2x_1 + 3y_1 + z_1) + \beta(2x_2 + 3y_2 + z_2) = \alpha 0 + \beta 0 = 0. \end{aligned}$$

Therefore, this does specify a subspace. It remains to find a basis for it.

From the equation, $z = -2x - 3y$ and so the vectors which satisfy the equation are of the form

$$\begin{pmatrix} x \\ y \\ -2x - 3y \end{pmatrix} = x \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ -3 \end{pmatrix}. \tag{16.10}$$

Therefore, a spanning set for this subspace is

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -3 \end{pmatrix} \right\}.$$

This will be a basis if it is linearly independent. Suppose

$$c_1 \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} + c_2 \begin{pmatrix} 0 \\ 1 \\ -3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This is equivalent to the system of equations,

$$\begin{array}{c} c_1 + 0c_2 = 0 \\ 0c_1 + c_2 = 0 \\ -2c_1 - 3c_2 = 0 \end{array}$$

having augmented matrix,

$$\left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & -3 & 0 \end{array}\right)$$

and you see the only solution to this is $c_1 = c_2 = 0$ from the top two lines of the augmented matrix. Therefore, these vectors form a basis for the subspace.

Are there any other bases? The answer is that there are infinitely many bases for this or any subspace. A simple way to see this is to replace one of the vectors by any nonzero multiple of itself. For example, the same subspace is obtained as the span of the two vectors,

$$\left\{\begin{pmatrix} 2 \\ 0 \\ -4 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -3 \end{pmatrix}\right\}.$$

We just replaced the first vector by 2 times the first vector. A more interesting example is

$$\left\{\begin{pmatrix} 1 \\ 1 \\ -5 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -3 \end{pmatrix}\right\}.$$

You can verify this is also a basis. We got it by replacing the first vector by the sum of the two. Another way to get a basis would be solve the equation for x rather than z. Thus

$$x = \frac{1}{2}\left(-z - 3y\right)$$

and then vectors in the subspace are of the form

$$\begin{pmatrix} \frac{-1}{2}z - \frac{3}{2}y \\ y \\ z \end{pmatrix}, y, z \in \mathbb{R}.$$

Thus, in the same way as above, a basis is

$$\left\{\begin{pmatrix} \frac{-1}{2} \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} \frac{3}{2} \\ 1 \\ 0 \end{pmatrix}\right\}$$

or if you like,

$$\left\{\begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix}\right\}$$

where we simply multiplied both vectors by 2.

In words the following corollary states that any linearly independent set of vectors can be enlarged to form a basis.

Corollary 16.1.44 *Let V be a subspace of $\mathbb{F}^n$ and let $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ be a linearly independent set of vectors in V. Then either it is a basis for V or there exist vectors, $\mathbf{v}_{r+1}, \cdots, \mathbf{v}_s$ such that $\{\mathbf{v}_1, \cdots, \mathbf{v}_r, \mathbf{v}_{r+1}, \cdots, \mathbf{v}_s\}$ is a basis for V.*

Proof: This follows immediately from the proof of Theorem 16.1.42. You do exactly the same argument except you start with $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ rather than $\{\mathbf{v}_1\}$.

It is also true that any spanning set of vectors can be restricted to obtain a basis.

Theorem 16.1.45 *Let V be a subspace of $\mathbb{F}^n$ and suppose $\operatorname{span}(\mathbf{u}_1 \cdots, \mathbf{u}_p) = V$ where the $\mathbf{u}_i$ are nonzero vectors. Then there exist vectors, $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ such that $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\} \subseteq \{\mathbf{u}_1 \cdots, \mathbf{u}_p\}$ and $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ is a basis for V.*

Proof: Let r be the smallest positive integer with the property that for some set, $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\} \subseteq \{\mathbf{u}_1 \cdots, \mathbf{u}_p\}$,

$$\operatorname{span}(\mathbf{v}_1 \cdots, \mathbf{v}_r) = V.$$

Then $r \leq p$ and it must be the case that $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ is linearly independent because if it were not so, one of the vectors, say $\mathbf{v}_k$ would be a linear combination of the others. But then you could delete this vector from $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ and the resulting list of $r-1$ vectors would still span V contrary to the definition of r. This proves the theorem.

16.1.6 Proof Of Exchange Theorem

Theorem 16.1.46 *(Exchange Theorem) Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ be a linearly independent set of vectors such that each $\mathbf{x}_i$ is in span$(\mathbf{y}_1, \cdots, \mathbf{y}_s)$. Then $r \leq s$.*

Proof: Define $\operatorname{span}\{\mathbf{y}_1, \cdots, \mathbf{y}_s\} \equiv V$, it follows there exist scalars, $c_1, \cdots, c_s$ such that

$$\mathbf{x}_1 = \sum_{i=1}^{s} c_i \mathbf{y}_i. \tag{16.11}$$

Not all of these scalars can equal zero because if this were the case, it would follow that $\mathbf{x}_1 = 0$ and so $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ would not be linearly independent. Indeed, if $\mathbf{x}_1 = \mathbf{0}$, $1\mathbf{x}_1 + \sum_{i=2}^{r} 0\mathbf{x}_i = \mathbf{x}_1 = \mathbf{0}$ and so there would exist a nontrivial linear combination of the vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ which equals zero.

Say $c_k \neq 0$. Then solve ((16.11)) for $\mathbf{y}_k$ and obtain

$$\mathbf{y}_k \in \operatorname{span}\left(\mathbf{x}_1, \overbrace{\mathbf{y}_1, \cdots, \mathbf{y}_{k-1}, \mathbf{y}_{k+1}, \cdots, \mathbf{y}_s}^{\text{s-1 vectors here}}\right).$$

Define $\{\mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$ by

$$\{\mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\} \equiv \{\mathbf{y}_1, \cdots, \mathbf{y}_{k-1}, \mathbf{y}_{k+1}, \cdots, \mathbf{y}_s\}$$

Therefore, $\operatorname{span}\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\} = V$ because if $\mathbf{v} \in V$, there exist constants $c_1, \cdots, c_s$ such that

$$\mathbf{v} = \sum_{i=1}^{s-1} c_i \mathbf{z}_i + c_s \mathbf{y}_k.$$

Now replace the $\mathbf{y}_k$ in the above with a linear combination of the vectors, $\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$ to obtain $\mathbf{v} \in \operatorname{span}\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$. The vector $\mathbf{y}_k$, in the list $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$, has now been replaced with the vector $\mathbf{x}_1$ and the resulting modified list of vectors has the same span as the original list of vectors, $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$.

Now suppose that $r > s$ and that $\operatorname{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{z}_1, \cdots, \mathbf{z}_p\} = V$ where the vectors, $\mathbf{z}_1, \cdots, \mathbf{z}_p$ are each taken from the set, $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$ and $l + p = s$. This has now been done for $l = 1$ above. Then since $r > s$, it follows that $l \leq s < r$ and so $l + 1 \leq r$. Therefore,

$\mathbf{x}_{l+1}$ is a vector not in the list, $\{\mathbf{x}_1, \cdots, \mathbf{x}_l\}$ and since $\text{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{z}_1, \cdots, \mathbf{z}_p\} = V$, there exist scalars, c_i and d_j such that

$$\mathbf{x}_{l+1} = \sum_{i=1}^{l} c_i \mathbf{x}_i + \sum_{j=1}^{p} d_j \mathbf{z}_j. \tag{16.12}$$

Now not all the d_j can equal zero because if this were so, it would follow that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ would be a linearly dependent set because one of the vectors would equal a linear combination of the others. Therefore, ((16.12)) can be solved for one of the $\mathbf{z}_i$, say $\mathbf{z}_k$, in terms of $\mathbf{x}_{l+1}$ and the other $\mathbf{z}_i$ and just as in the above argument, replace that $\mathbf{z}_i$ with $\mathbf{x}_{l+1}$ to obtain

$$\text{span}\left\{\mathbf{x}_1, \cdots \mathbf{x}_l, \mathbf{x}_{l+1}, \overbrace{\mathbf{z}_1, \cdots \mathbf{z}_{k-1}, \mathbf{z}_{k+1}, \cdots, \mathbf{z}_p}^{\text{p-1 vectors here}}\right\} = V.$$

Continue this way, eventually obtaining

$$\text{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_s\} = V.$$

But then $\mathbf{x}_r \in \text{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_s\}$ contrary to the assumption that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is linearly independent. Therefore, $r \leq s$ as claimed.

16.2 Exercises

1. Let V denote the 2×2 matrices which are of the form $\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ where a and b are numbers. Define addition and scalar multiplication in the usual way. Determine whether V is a vector space.

2. Let V denote the 2×2 matrices which are of the form $\begin{pmatrix} a & 1 \\ 0 & b \end{pmatrix}$ where a and b are numbers. Define addition and scalar multiplication in the usual way. Determine whether V is a vector space.

3. Suppose you define addition of vectors in $\mathbb{R}^2$ in the following funny way. $(x, y) + (z, w) = (x + 2z, y + w)$ and scalar multiplication in the usual way, $k(x, y) = (kx, ky)$. With these operations, is $\mathbb{R}^2$ a vector space? Explain.

4. Suppose you define addition of vectors in $\mathbb{R}^2$ in the usual way. $(x, y) + (z, w) = (x + z, y + w)$ but scalar multiplication in the following strange way: $k(x, y) = (x, y)$. With these operations, is $\mathbb{R}^2$ a vector space? Explain.

5. Here are three vectors in $\mathbb{R}^3$: $(1, 2, 1), (0, 1, 1), (1, 4, 3)$. Find a simple description of the span of these three vectors. Is the span of these three vectors a subspace?

6. Let $M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 = u_1 = 0\}$. Is M a subspace? Explain.

7. Let $M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 \geq u_1\}$. Is M a subspace? Explain.

8. Let $\mathbf{w} \in \mathbb{R}^4$ and let $M = \left\{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \sum_{i=1}^{4} u_i w_i = 0\right\}$. Is M a subspace? Explain.

9. Let $M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_i \geq 0 \text{ for each } i = 1, 2, 3, 4\}$. Is M a subspace? Explain.

10. Let $\mathbf{w}, \mathbf{w}_1$ be given vectors in $\mathbb{R}^4$ and define

$$M = \left\{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \sum_{i=1}^{4} u_i w_i = 0 \text{ and } \sum_{i=1}^{4} u_i w_{1i} = 0 \right\}.$$

Is M a subspace? Explain.

11. Let $M = \left\{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : |u_1| \leq 4 \right\}$. Is M a subspace? Explain.

12. Let $M = \left\{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \sin(u_1) = 1 \right\}$. Is M a subspace? Explain.

13. Here are three vectors. Determine whether they are linearly independent or linearly dependent.

$$\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}$$

14. Verify that the set of real valued functions defined on D for some set, D is a vector space. See the examples listed after the definition of vector space.

15. Here are three vectors. Determine whether they are linearly independent or linearly dependent.

$$\begin{pmatrix} 4 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}$$

16. Here are three vectors. Determine whether they are linearly independent or linearly dependent.

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 5 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}$$

17. Here are four vectors. Determine whether they span $\mathbb{R}^3$. Are these vectors linearly independent?

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 3 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$$

18. Here are four vectors. Determine whether they span $\mathbb{R}^3$. Are these vectors linearly independent?

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 3 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$$

19. Determine whether the following vectors are a basis for $\mathbb{R}^3$. If they are, explain why they are and if they are not, give a reason and tell whether they span $\mathbb{R}^3$.

$$\begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 3 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix}$$

20. Determine whether the following vectors are a basis for $\mathbb{R}^3$. If they are, explain why they are and if they are not, give a reason and tell whether they span $\mathbb{R}^3$.

$$\begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}$$

21. Determine whether the following vectors are a basis for $\mathbb{R}^3$. If they are, explain why they are and if they are not, give a reason and tell whether they span $\mathbb{R}^3$.

$$\begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

22. Determine whether the following vectors are a basis for $\mathbb{R}^3$. If they are, explain why they are and if they are not, give a reason and tell whether they span $\mathbb{R}^3$.

$$\begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

23. If you have 5 vectors in $\mathbb{F}^5$ and the vectors are linearly independent, can it always be concluded they span $\mathbb{F}^5$? Explain.

24. If you have 6 vectors in $\mathbb{F}^5$, is it possible they are linearly independent? Explain.

25. Consider the vectors of the form

$$\left\{ \begin{pmatrix} 2t+3s \\ s-t \\ t+s \end{pmatrix} : s,t \in \mathbb{R} \right\}.$$

Is this set of vectors a subspace of $\mathbb{R}^3$? If so, explain why, give a basis for the subspace and find its dimension.

26. Consider the vectors of the form

$$\left\{ \begin{pmatrix} 2t+3s+u \\ s-t \\ t+s \\ u \end{pmatrix} : s,t,u \in \mathbb{R} \right\}.$$

Is this set of vectors a subspace of $\mathbb{R}^4$? If so, explain why, give a basis for the subspace and find its dimension.

27. Consider the vectors of the form

$$\left\{ \begin{pmatrix} 2t+u \\ t+3u \\ t+s+v \\ u \end{pmatrix} : s,t,u,v \in \mathbb{R} \right\}.$$

Is this set of vectors a subspace of $\mathbb{R}^4$? If so, explain why, give a basis for the subspace and find its dimension.

28. In any vector space, show that if $\mathbf{x}+\mathbf{y}=\mathbf{0}$, then $\mathbf{y}=-\mathbf{x}$.

29. Show that in any vector space, $0\mathbf{x}=\mathbf{0}$. That is, the scalar 0 times the vector $\mathbf{x}$ gives the vector $\mathbf{0}$.

30. Show that in any vector space, $(-1)\,\mathbf{x}=-\mathbf{x}$.

31. Let X be a vector space and suppose $\{\mathbf{x}_1,\cdots,\mathbf{x}_k\}$ is a set of vectors from X. Show that $\mathbf{0}$ is in $\operatorname{span}(\mathbf{x}_1,\cdots,\mathbf{x}_k)$.

32. Let the vectors be polynomials of degree no more than 3. Show that with the usual definitions of scalar multiplication wherein for $p(x)$ a polynomial, $(\alpha p)(x)=\alpha p(x)$ this is a vector space.

33. In the previous problem show that a basis for the vector space is $\left\{1,x,x^2,x^3\right\}$.

34. Suppose A is an $m\times n$ matrix. $A(\mathbb{F}^n)$ is defined to be the set of vector which are equal to $A\mathbf{x}$ for some $\mathbf{x}\in\mathbb{F}^n$. Show $A(\mathbb{F}^n)$ is a subspace of $\mathbb{F}^m$. If $\{\mathbf{y}_1,\cdots,\mathbf{y}_p\}$ is a basis for $A(\mathbb{F}^n)$, such that $\mathbf{y}_i=A\mathbf{x}_i$ where $\mathbf{x}_i\in\mathbb{F}^n$, show $\{\mathbf{x}_1,\cdots,\mathbf{x}_p\}$ is linearly independent. Would the same conclusion hold if you only knew $\{\mathbf{y}_1,\cdots,\mathbf{y}_p\}$ is a linearly independent set?

35. Suppose A is an $m\times n$ matrix and $\{\mathbf{x}_1,\cdots,\mathbf{x}_p\}$ is a linearly independent set in $\mathbb{F}^n$, when can you conclude $\{A\mathbf{x}_1,\cdots,A\mathbf{x}_p\}$ is a linearly independent set in $\mathbb{F}^m$?

Part IV

Vectors In $\mathbb{R}^n$

Vectors And Points In $\mathbb{R}^n$

17.0.1 Outcomes

1. Evaluate the distance between two points in $\mathbb{R}^n$.
2. Be able to represent a vector in each of the following ways for $n = 2, 3$
   (a) as a directed arrow in n space
   (b) as an ordered n tuple
   (c) as a linear combination of unit coordinate vectors
3. Carry out the vector operations:
   (a) addition
   (b) scalar multiplication
   (c) find magnitude (norm or length)
   (d) Find the vector of unit length in the direction of a given vector.
4. Represent the operations of vector addition, scalar multiplication and norm geometrically.
5. Recall and apply the basic properties of vector addition, scalar multiplication and norm.
6. Model and solve application problems using vectors.
7. Describe an open ball in $\mathbb{R}^n$.
8. Determine whether a set in $\mathbb{R}^n$ is open, closed, or neither.

17.1 Distance in $\mathbb{R}^n$

How is distance between two points in $\mathbb{R}^n$ defined?

Definition 17.1.1 *Let* $\mathbf{x} = (x_1, \cdots, x_n)$ *and* $\mathbf{y} = (y_1, \cdots, y_n)$ *be two points in* $\mathbb{R}^n$. *Then* $|\mathbf{x} - \mathbf{y}|$ *to indicates the distance between these points and is defined as*

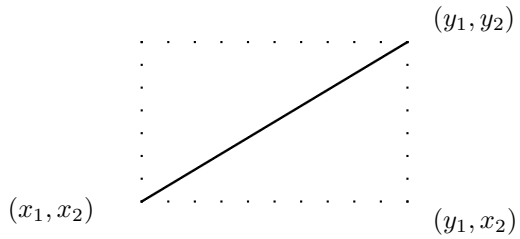
$$\text{distance between } \mathbf{x} \text{ and } \mathbf{y} \equiv |\mathbf{x} - \mathbf{y}| \equiv \left(\sum_{k=1}^{n} |x_k - y_k|^2 \right)^{1/2}.$$

This is called the ***distance formula****. 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* $\mathbf{a}$*. It gives all the points in* $\mathbb{R}^n$ *which are closer to* $\mathbf{a}$ *than* r*.*

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. Consider the following picture in the case that $n = 2$.

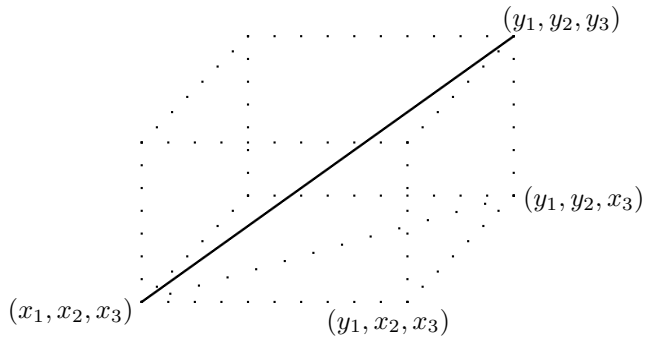


There are two points in the plane whose Cartesian coordinates are (x_1, x_2) and (y_1, y_2) respectively. Then the solid line joining these two points is the hypotenuse of a right triangle which is half of the rectangle shown in dotted lines. What is its length? Note the lengths of the sides of this 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.

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

$$\left((y_1 - x_1)^2 + (y_2 - x_2)^2\right)^{1/2}$$

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

$$\begin{aligned}&\left\{\left[\left((y_1 - x_1)^2 + (y_2 - x_2)^2\right)^{1/2}\right]^2 + (y_3 - x_3)^2\right\}^{1/2}\\ &= \left((y_1 - x_1)^2 + (y_2 - x_2)^2 + (y_3 - x_3)^2\right)^{1/2},\end{aligned}$$

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 not problem with the formula for distance in any number of dimensions. Here is an example.

Example 17.1.2 *Find the distance between the points in* $\mathbb{R}^4$, $\mathbf{a} = (1, 2, -4, 6)$ *and* $\mathbf{b} = (2, 3, -1, 0)$

Use the distance formula and write

$$|\mathbf{a} - \mathbf{b}|^2 = (1 - 2)^2 + (2 - 3)^2 + (-4 - (-1))^2 + (6 - 0)^2 = 47$$

Therefore, $|\mathbf{a} - \mathbf{b}| = \sqrt{47}$.

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 17.1.3 *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

$$2x + 2y + 2z = -9. \tag{17.1}$$

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 (17.1) holds.

The following lemma is fundamental. It is a form of the Cauchy Schwarz inequality.

Lemma 17.1.4 *Let* $\mathbf{x} = (x_1, \cdots, x_n)$ *and* $\mathbf{y} = (y_1, \cdots, y_n)$ *be two points in* $\mathbb{R}^n$. *Then*

$$\left| \sum_{i=1}^{n} x_i y_i \right| \leq |\mathbf{x}| \, |\mathbf{y}| . \tag{17.2}$$

Proof: Let θ be either 1 or -1 such that

$$\theta \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i \left(\theta y_i\right) = \left| \sum_{i=1}^{n} x_i y_i \right|$$

and consider $p(t) \equiv \sum_{i=1}^{n} \left(x_i + t\theta y_i\right)^2$. Then for all $t \in \mathbb{R}$,

$$\begin{aligned} 0 &\leq p(t) = \sum_{i=1}^{n} x_i^2 + 2t \sum_{i=1}^{n} x_i \theta y_i + t^2 \sum_{i=1}^{n} y_i^2 \\ &= |\mathbf{x}|^2 + 2t \sum_{i=1}^{n} x_i \theta y_i + t^2 |\mathbf{y}|^2 \end{aligned}$$

If $|\mathbf{y}| = 0$ then (17.2) is obviously true because both sides equal zero. Therefore, assume $|\mathbf{y}| \neq 0$ and then $p(t)$ is a polynomial of degree two whose graph opens up. Therefore, it either has no zeroes, two zeros or one repeated zero. If it has two zeros, the above inequality must be violated because in this case the graph must dip below the x axis. Therefore, it either has no zeros or exactly one. From the quadratic formula this happens exactly when

$$4 \left(\sum_{i=1}^{n} x_i \theta y_i \right)^2 - 4 |\mathbf{x}|^2 |\mathbf{y}|^2 \leq 0$$

and so

$$\sum_{i=1}^{n} x_i \theta y_i = \left| \sum_{i=1}^{n} x_i y_i \right| \leq |\mathbf{x}| \, |\mathbf{y}|$$

as claimed. This proves the inequality.

There are certain properties of the distance which are obvious. Two of them which follow directly from the definition are

$$|\mathbf{x} - \mathbf{y}| = |\mathbf{y} - \mathbf{x}| ,$$

$$|\mathbf{x} - \mathbf{y}| \geq 0 \text{ and equals 0 only if } \mathbf{y} = \mathbf{x}.$$

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. The following corollary is equivalent to this simple statement.

Corollary 17.1.5 *Let* $\mathbf{x}, \mathbf{y}$ *be points of* $\mathbb{R}^n$. *Then*

$$|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}| .$$

Proof: Using the Cauchy Schwarz inequality, Lemma 17.1.4,

$$\begin{aligned} |\mathbf{x} + \mathbf{y}|^2 &\equiv \sum_{i=1}^{n} \left(x_i + y_i\right)^2 \\ &= \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 \\ &\leq |\mathbf{x}|^2 + 2 |\mathbf{x}| \, |\mathbf{y}| \leq |\mathbf{y}|^2 \\ &= \left(|\mathbf{x}| + |\mathbf{y}|\right)^2 \end{aligned}$$

and so upon taking square roots of both sides,

$$|\mathbf{x}+\mathbf{y}| \leq |\mathbf{x}|+|\mathbf{y}|$$

and this proves the corollary.

17.2 Open And Closed Sets

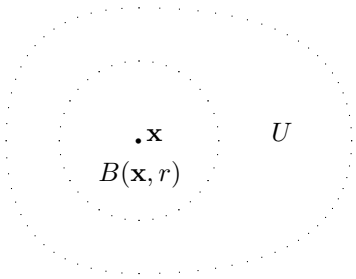
Eventually, one must consider functions which are defined on subsets of $\mathbb{R}^n$ and their properties. The next definition will end up being quite important. It describe a type of subset of $\mathbb{R}^n$ with the property that if $\mathbf{x}$ is in this set, then so is $\mathbf{y}$ whenever $\mathbf{y}$ is close enough to $\mathbf{x}$.

Definition 17.2.1 *Let $U \subseteq \mathbb{R}^n$. U is an **open set** if whenever $\mathbf{x} \in U$, there exists $r > 0$ such that $B(\mathbf{x}, r) \subseteq U$. More generally, if U is any subset of $\mathbb{R}^n$, $\mathbf{x} \in U$ is an **interior point** of U if there exists $r > 0$ such that $\mathbf{x} \in B(\mathbf{x}, r) \subseteq U$. In other words U is an open set exactly when every point of U is an interior point of U.*

If there is something called an open set, surely there should be something called a closed set and here is the definition of one.

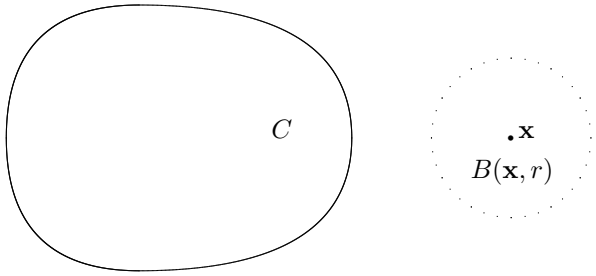
Definition 17.2.2 *A subset, C, of $\mathbb{R}^n$ is called a **closed set** if $\mathbb{R}^n \setminus C$ is an open set. They symbol, $\mathbb{R}^n \setminus C$ denotes everything in $\mathbb{R}^n$ which is not in C. It is also called the **complement** of C. The symbol, S^C is a short way of writing $\mathbb{R}^n \setminus S$.*

To illustrate this definition, consider the following picture.



You see in this picture how the edges are dotted. This is because an open set, can not include the edges or the set would fail to be open. For example, consider what would happen if you picked a point out on the edge of U in the above picture. Every open ball centered at that point would have in it some points which are outside U. Therefore, such a point would violate the above definition. You also see the edges of $B(\mathbf{x}, r)$ dotted suggesting that $B(\mathbf{x}, r)$ ought to be an open set. This is intuitively clear but does require a proof. This will be done in the next theorem and will give examples of open sets. Also, you can see that if $\mathbf{x}$ is close to the edge of U, you might have to take r to be very small.

It is roughly the case that open sets don't have their skins while closed sets do. Here is a picture of a closed set, C.



Note that $\mathbf{x} \notin C$ and since $\mathbb{R}^n \setminus C$ is open, there exists a ball, $B(\mathbf{x}, r)$ contained entirely in $\mathbb{R}^n \setminus C$. If you look at $\mathbb{R}^n \setminus C$, what would be its skin? It can't be in $\mathbb{R}^n \setminus C$ and so it must be in C. This is a rough heuristic explanation of what is going on with these definitions. Also note that $\mathbb{R}^n$ and $\emptyset$ are both open and closed. Here is why. If $\mathbf{x} \in \emptyset$, then there must be a ball centered at $\mathbf{x}$ which is also contained in $\emptyset$. This must be considered to be true because there is nothing in $\emptyset$ so there can be no example to show it false[1]. Therefore, from the definition, it follows $\emptyset$ is open. It is also closed because if $\mathbf{x} \notin \emptyset$, then $B(\mathbf{x}, 1)$ is also contained in $\mathbb{R}^n \setminus \emptyset = \mathbb{R}^n$. Therefore, $\emptyset$ is both open and closed. From this, it follows $\mathbb{R}^n$ is also both open and closed.

Theorem 17.2.3 *Let $\mathbf{x} \in \mathbb{R}^n$ and let $r \geq 0$. Then $B(\mathbf{x}, r)$ is an open set. Also,*

$$D(\mathbf{x}, r) \equiv \{\mathbf{y} \in \mathbb{R}^n : |\mathbf{y} - \mathbf{x}| \leq r\}$$

is a closed set.

Proof: Suppose $\mathbf{y} \in B(\mathbf{x}, r)$. It is necessary to show there exists $r_1 > 0$ such that $B(\mathbf{y}, r_1) \subseteq B(\mathbf{x}, r)$. Define $r_1 \equiv r - |\mathbf{x} - \mathbf{y}|$. Then if $|\mathbf{z} - \mathbf{y}| < r_1$, it follows from the above triangle inequality that

$$\begin{aligned} |\mathbf{z} - \mathbf{x}| &= |\mathbf{z} - \mathbf{y} + \mathbf{y} - \mathbf{x}| \\ &\leq |\mathbf{z} - \mathbf{y}| + |\mathbf{y} - \mathbf{x}| \\ &< r_1 + |\mathbf{y} - \mathbf{x}| = r - |\mathbf{x} - \mathbf{y}| + |\mathbf{y} - \mathbf{x}| = r. \end{aligned}$$

Note that if $r = 0$ then $B(\mathbf{x}, r) = \emptyset$, the empty set. This is because if $\mathbf{y} \in \mathbb{R}^n$, $|\mathbf{x} - \mathbf{y}| \geq 0$ and so $\mathbf{y} \notin B(\mathbf{x}, 0)$. Since $\emptyset$ has no points in it, it must be open because every point in it, (There are none.) satisfies the desired property of being an interior point.

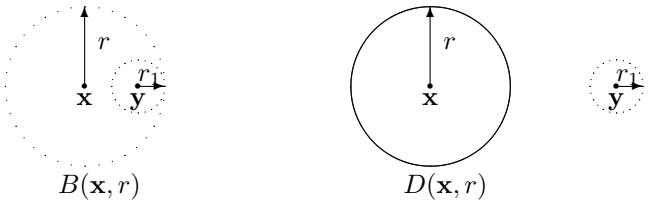
Now suppose $\mathbf{y} \notin D(\mathbf{x}, r)$. Then $|\mathbf{x} - \mathbf{y}| > r$ and defining $\delta \equiv |\mathbf{x} - \mathbf{y}| - r$, it follows that if $\mathbf{z} \in B(\mathbf{y}, \delta)$, then by the triangle inequality,

$$\begin{aligned} |\mathbf{x} - \mathbf{z}| &\geq |\mathbf{x} - \mathbf{y}| - |\mathbf{y} - \mathbf{z}| > |\mathbf{x} - \mathbf{y}| - \delta \\ &= |\mathbf{x} - \mathbf{y}| - (|\mathbf{x} - \mathbf{y}| - r) = r \end{aligned}$$

and this shows that $B(\mathbf{y}, \delta) \subseteq \mathbb{R}^n \setminus D(\mathbf{x}, r)$. Since $\mathbf{y}$ was an arbitrary point in $\mathbb{R}^n \setminus D(\mathbf{x}, r)$, it follows $\mathbb{R}^n \setminus D(\mathbf{x}, r)$ is an open set which shows from the definition that $D(\mathbf{x}, r)$ is a closed set as claimed.

[1] To a mathematician, the statment: Whenever a pig is born with wings it can fly must be taken as true. We do not consider biological or aerodynamic considerations in such statements. There is no such thing as a winged pig and therefore, all winged pigs must be superb flyers since there can be no example of one which is not. On the other hand we would also consider the statement: Whenever a pig is born with wings it can't possibly fly, as equally true. The point is, you can say anything you want about the elements of the empty set and no one can gainsay your statement. Therefore, such statements are considered as true by default. You may say this is a very strange way of thinking about truth and ultimately this is because mathematics is not about truth. It is more about consistency and logic.

A picture which is descriptive of the conclusion of the above theorem which also implies the manner of proof is the following.



Recall $\mathbb{R}^2$ consists of ordered pairs, (x, y) such that $x \in \mathbb{R}$ and $y \in \mathbb{R}$. $\mathbb{R}^2$ is also written as $\mathbb{R} \times \mathbb{R}$. In general, the following definition holds.

Definition 17.2.4 *The **Cartesian product** of two sets, $A \times B$, means* $\{(a, b) : a \in A,\ b \in B\}$. *If you have n sets, $A_1, A_2, \cdots, A_n$*

$$\prod_{i=1}^{n} A_i = \{(x_1, x_2, \cdots, x_n) : \text{ each } x_i \in A_i\}.$$

Now suppose $A \subseteq \mathbb{R}^m$ and $B \subseteq \mathbb{R}^n$. Then if $(\mathbf{x}, \mathbf{y}) \in A \times B$, $\mathbf{x} = (x_1, \cdots, x_m)$ and $\mathbf{y} = (y_1, \cdots, y_n)$, the following identification will be made.

$$(\mathbf{x}, \mathbf{y}) = (x_1, \cdots, x_m, y_1, \cdots, y_n) \in \mathbb{R}^{n+m}.$$

Similarly, starting with something in $\mathbb{R}^{n+m}$, you can write it in the form $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$. The following theorem has to do with the Cartesian product of two closed sets or two open sets. Also here is an important definition.

Definition 17.2.5 *A set, $A \subseteq \mathbb{R}^n$ is said to be **bounded** if there exist finite intervals, $[a_i, b_i]$ such that*

$$A \subseteq \prod_{i=1}^{n} [a_i, b_i].$$

Theorem 17.2.6 *Let U be an open set in $\mathbb{R}^m$ and let V be an open set in $\mathbb{R}^n$. Then $U \times V$ is an open set in $\mathbb{R}^{n+m}$. If C is a closed set in $\mathbb{R}^m$ and H is a closed set in $\mathbb{R}^n$, then $C \times H$ is a closed set in $\mathbb{R}^{n+m}$. If C and H are bounded, then so is $C \times H$.*

Proof: Let $(\mathbf{x}, \mathbf{y}) \in U \times V$. Since U is open, there exists $r_1 > 0$ such that $B(\mathbf{x}, r_1) \subseteq U$. Similarly, there exists $r_2 > 0$ such that $B(\mathbf{y}, r_2) \subseteq V$. Now

$$B((\mathbf{x}, \mathbf{y}), \delta) \equiv$$

$$\left\{ (\mathbf{s}, \mathbf{t}) \in \mathbb{R}^{n+m} : \sum_{k=1}^{m} |x_k - s_k|^2 + \sum_{j=1}^{n} |y_j - t_j|^2 < \delta^2 \right\}$$

Therefore, if $\delta \equiv \min(r_1, r_2)$ and $(\mathbf{s}, \mathbf{t}) \in B((\mathbf{x}, \mathbf{y}), \delta)$, then it follows that $\mathbf{s} \in B(\mathbf{x}, r_1) \subseteq U$ and that $\mathbf{t} \in B(\mathbf{y}, r_2) \subseteq V$ which shows that $B((\mathbf{x}, \mathbf{y}), \delta) \subseteq U \times V$. Hence $U \times V$ is open as claimed.

Next suppose $(\mathbf{x}, \mathbf{y}) \notin C \times H$. It is necessary to show there exists $\delta > 0$ such that $B((\mathbf{x}, \mathbf{y}), \delta) \subseteq \mathbb{R}^{n+m} \setminus (C \times H)$. Either $\mathbf{x} \notin C$ or $\mathbf{y} \notin H$ since otherwise $(\mathbf{x}, \mathbf{y})$ would be a point of $C \times H$. Suppose therefore, that $\mathbf{x} \notin C$. Since C is closed, there exists $r > 0$ such that

$B(\mathbf{x},r) \subseteq \mathbb{R}^m \setminus C$. Consider $B((\mathbf{x},\mathbf{y}),r)$. If $(\mathbf{s},\mathbf{t}) \in B((\mathbf{x},\mathbf{y}),r)$, it follows that $\mathbf{s} \in B(\mathbf{x},r)$ which is contained in $\mathbb{R}^m \setminus C$. Therefore, $B((\mathbf{x},\mathbf{y}),r) \subseteq \mathbb{R}^{n+m} \setminus (C \times H)$ showing $C \times H$ is closed. A similar argument holds if $\mathbf{y} \notin H$.

If C is bounded, there exist $[a_i,b_i]$ such that $C \subseteq \prod_{i=1}^m [a_i,b_i]$ and if H is bounded, $H \subseteq \prod_{i=m+1}^{m+n} [a_i,b_i]$ for intervals $[a_{m+1},b_{m+1}],\cdots,[a_{m+n},b_{m+n}]$. Therefore, $C \times H \subseteq \prod_{i=1}^{m+n} [a_i,b_i]$ and this establishes the last part of this theorem.

17.3 Exercises

1. 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) $(2,-5)$

2. Does it make sense to write $(1,2)+(2,3,1)$? Explain.

3. 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) $(-2,3,-2)$

4. You are given two points in $\mathbb{R}^3, (4,5,-4)$ and $(2,3,0)$. Show the distance from the point, $(3,4,-2)$ to the first of these points is the same as the distance from this point to the second of the original pair of points. Note that $3 = \frac{4+2}{2}, 4 = \frac{5+3}{2}$. Obtain a theorem which will be valid for general pairs of points, (x,y,z) and (x_1,y_1,z_1) and prove your theorem using the distance formula.

5. A sphere is the set of all points which are at a given distance from a single given point. Find an equation for the sphere which is the set of all points that are at a distance of 4 from the point $(1,2,3)$ in $\mathbb{R}^3$.

6. A parabola is the set of all points (x,y) in the plane such that the distance from the point (x,y) to a given point, (x_0,y_0) equals the distance from (x,y) to a given line. The point, (x_0,y_0) is called the **focus** and the line is called the **directrix**. Find the equation of the parabola which results from the line $y=l$ and (x_0,y_0) a given focus with $y_0 < l$. Repeat for $y_0 > l$.

7. A sphere centered at the point $(x_0,y_0,z_0) \in \mathbb{R}^3$ having radius r consists of all points, (x,y,z) whose distance to (x_0,y_0,z_0) equals r. Write an equation for this sphere in $\mathbb{R}^3$.

8. Suppose the distance between (x,y) and (x',y') were defined to equal the larger of the two numbers $|x-x'|$ and $|y-y'|$. Draw a picture of the sphere centered at the point, $(0,0)$ if this notion of distance is used.

9. Repeat the same problem except this time let the distance between the two points be $|x-x'|+|y-y'|$.

10. If (x_1, y_1, z_1) and (x_2, y_2, z_2) are two points such that $|(x_i, y_i, z_i)| = 1$ for $i = 1, 2$, show that in terms of the usual distance, $\left|\left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}, \frac{z_1+z_2}{2}\right)\right| < 1$. What would happen if you used the way of measuring distance given in Problem 8 ($|(x, y, z)|$ = maximum of $|z|, |x|, |y|$.)?

11. Give a simple description using the distance formula of the set of points which are at an equal distance between the two points (x_1, y_1, z_1) and (x_2, y_2, z_2).

12. Suppose you are given two points, $(-a, 0)$ and $(a, 0)$ in $\mathbb{R}^2$ and a number, $r > 2a$. The set of points described by

$$\left\{(x, y) \in \mathbb{R}^2 : |(x, y) - (-a, 0)| + |(x, y) - (a, 0)| = r\right\}$$

is known as an ellipse. The two given points are known as the **focus points** of the ellipse. Simplify this to the form $\left(\frac{x-A}{\alpha}\right)^2 + \left(\frac{y}{\beta}\right)^2 = 1$. This is a nice exercise in messy algebra.

13. Suppose you are given two points, $(-a, 0)$ and $(a, 0)$ in $\mathbb{R}^2$ and a number, $r > 2a$. The set of points described by

$$\left\{(x, y) \in \mathbb{R}^2 : |(x, y) - (-a, 0)| - |(x, y) - (a, 0)| = r\right\}$$

is known as **hyperbola**. The two given points are known as the **focus points** of the hyperbola. Simplify this to the form $\left(\frac{x-A}{\alpha}\right)^2 - \left(\frac{y}{\beta}\right)^2 = 1$. This is a nice exercise in messy algebra.

14. Let (x_1, y_1) and (x_2, y_2) be two points in $\mathbb{R}^2$. Give a simple description using the distance formula of the perpendicular bisector of the line segment joining these two points. Thus you want all points, (x, y) such that $|(x, y) - (x_1, y_1)| = |(x, y) - (x_2, y_2)|$.

15. Let $U = \{(x, y, z) \text{ such that } z > 0\}$. Determine whether U is open, closed or neither.

16. Let $U = \{(x, y, z) \text{ such that } z \geq 0\}$. Determine whether U is open, closed or neither.

17. Let $U = \left\{(x, y, z) \text{ such that } \sqrt{x^2 + y^2 + z^2} < 1\right\}$. Determine whether U is open, closed or neither.

18. Let $U = \left\{(x, y, z) \text{ such that } \sqrt{x^2 + y^2 + z^2} \leq 1\right\}$. Determine whether U is open, closed or neither.

19. Show carefully that $\mathbb{R}^n$ is both open and closed.

20. Show that every open set in $\mathbb{R}^n$ is the union of open balls contained in it.

21. Show the intersection of any two open sets is an open set.

22. If S is a nonempty subset of $\mathbb{R}^p$, a point, $\mathbf{x}$ is said to be a **limit point** of S if $B(\mathbf{x}, r)$ contains infinitely many points of S for each $r > 0$. Show this is equivalent to saying that $B(\mathbf{x}, r)$ contains a point of S different than $\mathbf{x}$ for each $r > 0$.

23. Closed sets were defined to be those sets which are complements of open sets. Show that a set is closed if and only if it contains all its limit points.

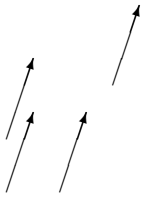
17.4 Physical Vectors

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 17.4.1 ***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.*

Of course this is a little vague and will be left a little vague until the presentation of Newton's second law later.

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).



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 $\overrightarrow{\mathbf{pq}}$ be a directed line segment or vector. Then from Definition 12.3.4 it follows that $\overrightarrow{\mathbf{pq}}$ consists of the points of the form

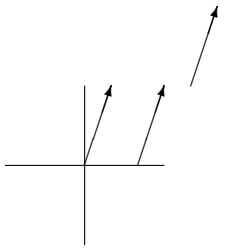
$$\mathbf{p} + t\,(\mathbf{q} - \mathbf{p})$$

where $t \in [0,1]$. Subtract $\mathbf{p}$ from all these points to obtain the directed line segment consisting of the points

$$\mathbf{0} + t\,(\mathbf{q} - \mathbf{p}),\ t \in [0,1].$$

The point in $\mathbb{R}^n, \mathbf{q} - \mathbf{p}$, will represent the vector.

Geometrically, the arrow, $\overrightarrow{\mathbf{pq}}$, 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 elements of $\mathbb{R}^n$.

The magnitude of a vector determined by a directed line segment $\overrightarrow{\mathbf{pq}}$ is just the distance between the point $\mathbf{p}$ and the point $\mathbf{q}$. By the distance formula this equals

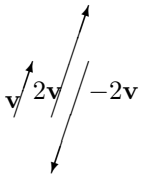
$$\left(\sum_{k=1}^{n} (q_k - p_k)^2\right)^{1/2} = |\mathbf{p} - \mathbf{q}|$$

and for $\mathbf{v}$ any vector in $\mathbb{R}^n$ the magnitude of $\mathbf{v}$ equals $\left(\sum_{k=1}^n v_k^2\right)^{1/2} = |\mathbf{v}|$.

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 $r\mathbf{v}$?

$$|r\mathbf{v}| = \left(\sum_{k=1}^n (ra_i)^2\right)^{1/2} = \left(\sum_{k=1}^n r^2 (a_i)^2\right)^{1/2}$$
$$= \left(r^2\right)^{1/2} \left(\sum_{k=1}^n a_i^2\right)^{1/2} = |r|\,|\mathbf{v}|.$$

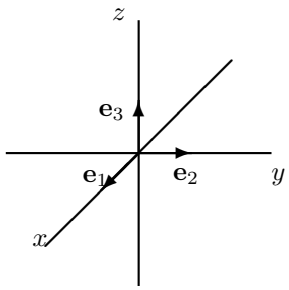
Thus the magnitude of $r\mathbf{v}$ equals $|r|$ times the magnitude of $\mathbf{v}$. If r is positive, then the vector represented by $r\mathbf{v}$ has the same direction as the vector, $\mathbf{v}$ because multiplying by the scalar, r, only has the effect of scaling all the distances. Thus the unit 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

$$\mathbf{e}_i \equiv (0, \cdots, 0, 1, 0, \cdots, 0)$$

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 $\mathbf{e}_i$ is referred to as the i^{th} direction. Given a vector, $\mathbf{v} = (a_1, \cdots, a_n)$, $a_i\mathbf{e}_i$ is the i^{th} component of the vector. Thus $a_i\mathbf{e}_i = (0, \cdots, 0, a_i, 0, \cdots, 0)$ 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 $\mathbf{v} = (a_1, \cdots, a_n)$,

$$\mathbf{v} = \sum_{k=1}^n a_i\mathbf{e}_i.$$

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 $\mathbf{a} = \sum_{k=1}^{n} a_i \mathbf{e}_i$ and $\mathbf{b} = \sum_{k=1}^{n} b_i \mathbf{e}_i$. Then the vector, $\mathbf{a}$ involves a component in the i^{th} direction, $a_i \mathbf{e}_i$ while the component in the i^{th} direction of $\mathbf{b}$ is $b_i \mathbf{e}_i$. Then it seems physically reasonable that the resultant vector should have a component in the i^{th} direction equal to $(a_i + b_i)\,\mathbf{e}_i$. This is exactly what is obtained when the vectors, $\mathbf{a}$ and $\mathbf{b}$ are added.

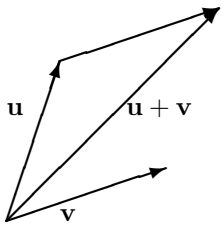
$$\begin{aligned} \mathbf{a} + \mathbf{b} &= (a_1 + b_1, \cdots, a_n + b_n)\,. \\ &= \sum_{i=1}^{n} (a_i + b_i)\,\mathbf{e}_i. \end{aligned}$$

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 $\mathbf{u}, \mathbf{v}$ are vectors,

$$\mathbf{u} = (u_1, \cdots, u_n)\,, \mathbf{v} = (v_1, \cdots, v_n)$$

Then $\mathbf{u} + \mathbf{v} = (u_1 + v_1, \cdots, u_n + v_n)$. How can one obtain this geometrically? Consider the directed line segment, $\overrightarrow{\mathbf{0u}}$ and then, starting at the end of this directed line segment, follow the directed line segment $\overrightarrow{\mathbf{u}\,(\mathbf{u}+\mathbf{v})}$ to its end, $\mathbf{u} + \mathbf{v}$. In other words, place the vector $\mathbf{u}$ in standard position with its base at the origin and then slide the vector $\mathbf{v}$ till its base coincides with the point of $\mathbf{u}$. The point of this slid vector, determines $\mathbf{u} + \mathbf{v}$. To illustrate, see the following picture



Note the vector $\mathbf{u} + \mathbf{v}$ is the diagonal of a parallelogram determined from the two vectors $\mathbf{u}$ and $\mathbf{v}$ and that identifying $\mathbf{u} + \mathbf{v}$ with the directed diagonal of the parallelogram determined by the vectors $\mathbf{u}$ and $\mathbf{v}$ 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 $\mathbf{i}$ for $\mathbf{e}_1, \mathbf{j}$ for $\mathbf{e}_2$, and $\mathbf{k}$ for $\mathbf{e}_3$. Now here are some applications of vector addition to some problems.

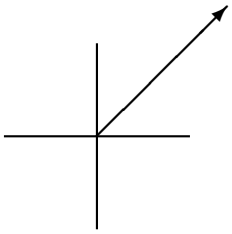
Example 17.4.2 *There are three ropes attached to a car and three people pull on these ropes. The first exerts a force of* $2\mathbf{i}+3\mathbf{j}-2\mathbf{k}$ *Newtons, the second exerts a force of* $3\mathbf{i}+5\mathbf{j}+\mathbf{k}$ *Newtons and the third exerts a force of* $5\mathbf{i}-\mathbf{j}+2\mathbf{k}$. *Newtons. Find the total force in the direction of* $\mathbf{i}$.

To find the total force add the vectors as described above. This gives $10\mathbf{i}+7\mathbf{j}+\mathbf{k}$ Newtons. Therefore, the force in the $\mathbf{i}$ direction is 10 Newtons.

As mentioned earlier, the Newton is a unit of force like pounds.

Example 17.4.3 *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 $100/\sqrt{2}\mathbf{i} + 100/\sqrt{2}\mathbf{j}$.

This example also motivates the concept of **velocity**.

Definition 17.4.4 *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 specifing the direction.*

Thus the velocity vector in the above example is $100/\sqrt{2}\mathbf{i} + 100/\sqrt{2}\mathbf{j}$.

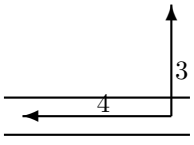
Example 17.4.5 *The velocity of an airplane is* $100\mathbf{i}+\mathbf{j}+\mathbf{k}$ *measured in kilometers per hour and at a certain instant of time its position is* $(1,2,1)$. *Here imagine a Cartesian coordinate system in which the third component is altitude and the first and second components 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 17.4.6 *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

$$\sqrt{(5/6)^2-(1/2)^2}=\frac{2}{3}\text{ miles.}$$

17.5 Exercises

1. The wind blows from West to East at a speed of 50 kilometers per hour and an airplane is heading North West at a speed of 300 Kilometers per hour. 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?

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 $(100,200)$ where distances are in miles. An airplane flies at 300 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 20 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.

Vector Products

18.0.1 Outcomes

1. Evaluate a dot product from the angle formula or the coordinate formula.
2. Interpret the dot product geometrically.
3. Evaluate the following using the dot product:
   (a) the angle between two vectors
   (b) the magnitude of a vector
   (c) the work done by a constant force on an object
4. Evaluate a cross product from the angle formula or the coordinate formula.
5. Interpret the cross product geometrically.
6. Evaluate the following using the cross product:
   (a) the area of a parallelogram
   (b) the area of a triangle
   (c) physical quantities such as the torque and angular velocity.
7. Find the volume of a parallelepiped using the box product.
8. Recall, apply and derive the algebraic properties of the dot and cross products.

18.1 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 18.1.1 *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, α and β will denote scalars and $\mathbf{a}, \mathbf{b}, \mathbf{c}$ will denote vectors.

Proposition 18.1.2 *The dot product satisfies the following properties.*

$$\mathbf{a}\cdot\mathbf{b}=\mathbf{b}\cdot\mathbf{a} \tag{18.1}$$

$$\mathbf{a}\cdot\mathbf{a}\geq 0 \text{ and equals zero if and only if } \mathbf{a}=\mathbf{0} \tag{18.2}$$

$$(\alpha\mathbf{a}+\beta\mathbf{b})\cdot\mathbf{c}=\alpha(\mathbf{a}\cdot\mathbf{c})+\beta(\mathbf{b}\cdot\mathbf{c}) \tag{18.3}$$

$$\mathbf{c}\cdot(\alpha\mathbf{a}+\beta\mathbf{b})=\alpha(\mathbf{c}\cdot\mathbf{a})+\beta(\mathbf{c}\cdot\mathbf{b}) \tag{18.4}$$

$$|\mathbf{a}|^2=\mathbf{a}\cdot\mathbf{a} \tag{18.5}$$

You should verify these properties. Also be sure you understand that (18.4) follows from the first three and is therefore redundant. It is listed here for the sake of convenience.

Example 18.1.3 *Find* $(1,2,0,-1)\cdot(0,1,2,3)$.

This equals $0+2+0+-3=-1$.

Example 18.1.4 *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 Schwartz inequality.**

Theorem 18.1.5 *The dot product satisfies the inequality*

$$|\mathbf{a}\cdot\mathbf{b}|\leq|\mathbf{a}||\mathbf{b}|. \tag{18.6}$$

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 (18.6) 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 $t\in\mathbb{R}$

$$f(t)=(\mathbf{a}+t\mathbf{b})\cdot(\mathbf{a}+t\mathbf{b}).$$

Then by (18.2), $f(t)\geq 0$ for all $t\in\mathbb{R}$. Also from (18.3),(18.4),(18.1), and (18.5)

$$\begin{aligned}f(t)&=\mathbf{a}\cdot(\mathbf{a}+t\mathbf{b})+t\mathbf{b}\cdot(\mathbf{a}+t\mathbf{b})\\&=\mathbf{a}\cdot\mathbf{a}+t(\mathbf{a}\cdot\mathbf{b})+t\mathbf{b}\cdot\mathbf{a}+t^2\mathbf{b}\cdot\mathbf{b}\\&=|\mathbf{a}|^2+2t(\mathbf{a}\cdot\mathbf{b})+|\mathbf{b}|^2t^2.\end{aligned}$$

Now

$$\begin{aligned}f(t)&=|\mathbf{b}|^2\left(t^2+2t\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|^2}+\frac{|\mathbf{a}|^2}{|\mathbf{b}|^2}\right)\\&=|\mathbf{b}|^2\left(t^2+2t\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|^2}+\left(\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|^2}\right)^2-\left(\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|^2}\right)^2+\frac{|\mathbf{a}|^2}{|\mathbf{b}|^2}\right)\\&=|\mathbf{b}|^2\left(\left(t+\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|^2}\right)^2+\left(\frac{|\mathbf{a}|^2}{|\mathbf{b}|^2}-\left(\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|^2}\right)^2\right)\right)\geq 0\end{aligned}$$

for all $t \in \mathbb{R}$. In particular $f(t) \geq 0$ when $t = -\left(\mathbf{a}\cdot\mathbf{b}/|\mathbf{b}|^2\right)$ which implies

$$\frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} - \left(\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|^2}\right)^2 \geq 0. \tag{18.7}$$

Multiplying both sides by $|\mathbf{b}|^4$,

$$|\mathbf{a}|^2 |\mathbf{b}|^2 \geq (\mathbf{a}\cdot\mathbf{b})^2$$

which yields (18.6).

From Theorem 18.1.5, equality holds in (18.6) whenever one of the vectors is a scalar multiple of the other. It only remains to verify this is the only way equality can occur. If either vector equals zero, then equality is obtained in (18.6) so it can be assumed both vectors are non zero and that equality is obtained in (18.7). This implies that $f(t) = 0$ when $t = -\left(\mathbf{a}\cdot\mathbf{b}/|\mathbf{b}|^2\right)$ and so from (18.2), it follows that for this value of t, $\mathbf{a}+t\mathbf{b} = \mathbf{0}$ showing $\mathbf{a} = -t\mathbf{b}$. This proves the theorem.

You should note that the entire argument was based only on the properties of the dot product listed in (18.1) - (18.5). 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 18.1.6 *(Triangle inequality) For* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$

$$|\mathbf{a}+\mathbf{b}| \leq |\mathbf{a}| + |\mathbf{b}| \tag{18.8}$$

and equality holds if and only if one of the vectors is a nonnegative scalar multiple of the other. Also

$$||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a}-\mathbf{b}| \tag{18.9}$$

Proof: By properties of the dot product and the Cauchy Schwartz inequality,

$$\begin{aligned}
|\mathbf{a}+\mathbf{b}|^2 &= (\mathbf{a}+\mathbf{b})\cdot(\mathbf{a}+\mathbf{b}) \\
&= (\mathbf{a}\cdot\mathbf{a}) + (\mathbf{a}\cdot\mathbf{b}) + (\mathbf{b}\cdot\mathbf{a}) + (\mathbf{b}\cdot\mathbf{b}) \\
&= |\mathbf{a}|^2 + 2(\mathbf{a}\cdot\mathbf{b}) + |\mathbf{b}|^2 \\
&\leq |\mathbf{a}|^2 + 2|\mathbf{a}\cdot\mathbf{b}| + |\mathbf{b}|^2 \\
&\leq |\mathbf{a}|^2 + 2|\mathbf{a}||\mathbf{b}| + |\mathbf{b}|^2 \\
&= (|\mathbf{a}| + |\mathbf{b}|)^2.
\end{aligned}$$

Taking square roots of both sides you obtain (18.8).

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 18.1.5 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

$$\begin{aligned} |\mathbf{a}| &= |\mathbf{a} - \mathbf{b} + \mathbf{b}| \\ &\leq |\mathbf{a} - \mathbf{b}| + |\mathbf{b}| . \end{aligned}$$

Therefore,

$$|\mathbf{a}| - |\mathbf{b}| \leq |\mathbf{a} - \mathbf{b}| \tag{18.10}$$

Similarly,

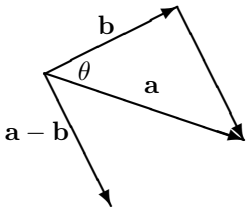
$$|\mathbf{b}| - |\mathbf{a}| \leq |\mathbf{b} - \mathbf{a}| = |\mathbf{a} - \mathbf{b}| . \tag{18.11}$$

It follows from (18.10) and (18.11) that (18.9) holds. This is because $||\mathbf{a}| - |\mathbf{b}||$ equals the left side of either (18.10) or (18.11) and either way, $||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a} - \mathbf{b}|$. This proves the theorem.

18.2 The Geometric Significance Of The Dot Product

18.2.1 The Angle Between Two Vectors

Given two vectors, **a** and **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.



By the law of cosines,

$$|\mathbf{a} - \mathbf{b}|^2 = |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2 |\mathbf{a}| |\mathbf{b}| \cos\theta.$$

Also from the properties of the dot product,

$$\begin{aligned} |\mathbf{a} - \mathbf{b}|^2 &= (\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b}) \\ &= |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2\mathbf{a} \cdot \mathbf{b} \end{aligned}$$

and so comparing the above two formulas,

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos\theta. \tag{18.12}$$

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 18.2.1 *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 (18.12) the cosine of the included angle equals

$$\cos\theta = \frac{9}{\sqrt{26}\sqrt{6}} = .720\,58$$

Now the cosine is known, the angle can be determines by solving the equation, $\cos\theta = .720\,58$. This will involve using a calculator or a table of trigonometric functions. The answer is $\theta = .766\,16$ radians or in terms of degrees, $\theta = .766\,16 \times \frac{360}{2\pi} = 43.898°$. Recall how this last computation is done. Set up a proportion, $\frac{x}{.766\,16} = \frac{360}{2\pi}$ because $360°$ corresponds to 2π 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 18.2.2 *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° and the angle between* $\mathbf{v}$ *and the positive* x *axis is -30°. Find* $\mathbf{u}\cdot\mathbf{v}$.

From the geometric description of the dot product in (18.12)

$$\mathbf{u}\cdot\mathbf{v} = 3\times 4\times\cos(60°) = 3\times 4\times 1/2 = 6.$$

Observation 18.2.3 *Two vectors are said to be* ***perpendicular*** *if the included angle is* $\pi/2$ *radians (90°). You can tell if two nonzero vectors are perpendicular by simply taking their dot product. If the answer is zero, this means they are are perpendicular because* $\cos\theta = 0$.

Example 18.2.4 *Determine whether the two vectors,* $2\mathbf{i}+\mathbf{j}-\mathbf{k}$ *and* $1\mathbf{i}+3\mathbf{j}+5\mathbf{k}$ *are perpendicular.*

When you take this dot product you get $2+3-5=0$ and so these two are indeed perpendicular.

Definition 18.2.5 *When two lines intersect, the angle between the two lines is the smaller of the two angles determined.*

Example 18.2.6 *Find the angle between the two lines,* $(1,2,0)+t(1,2,3)$ *and* $(0,4,-3)+t(-1,2,-3)$.

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

$$\cos\theta = \frac{-6}{14} = \frac{-3}{7}. \tag{18.13}$$

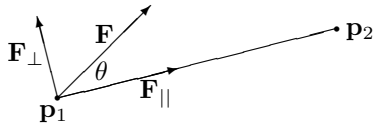
We don't want this angle because it is obtuse. The angle desired is the acute angle given by

$$\cos\theta = \frac{3}{7}.$$

It is obtained by using replacing one of the direction vectors with -1 times it.

18.2.2 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. 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, $\mathbf{F}_{||}$ and $\mathbf{F}_{\perp}$ and the picture is intended to indicate that when you add these two vectors you get $\mathbf{F}$ while $\mathbf{F}_{||}$ acts in the direction of motion and $\mathbf{F}_{\perp}$ acts perpendicular to the direction of motion. Only $\mathbf{F}_{||}$ contributes to the work done by $\mathbf{F}$ on the object as it moves from $\mathbf{p}_1$ to $\mathbf{p}_2$. From trigonometry, you see the magnitude of $\mathbf{F}_{||}$ should equal $|\mathbf{F}|\,|\cos\theta|$. Thus, since $\mathbf{F}_{||}$ points in the direction of the vector from $\mathbf{p}_1$ to $\mathbf{p}_2$, the total work done should equal

$$|\mathbf{F}|\left|\overrightarrow{\mathbf{p}_1\mathbf{p}_2}\right|\cos\theta = |\mathbf{F}|\,|\mathbf{p}_2-\mathbf{p}_1|\cos\theta$$

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 18.2.7 *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 18.2.8 *Let* $\mathbf{F}$ *and* $\mathbf{D}$ *be nonzero vectors. Then there exist unique vectors* $\mathbf{F}_{||}$ *and* $\mathbf{F}_{\perp}$ *such that*

$$\mathbf{F} = \mathbf{F}_{||} + \mathbf{F}_{\perp} \tag{18.14}$$

where $\mathbf{F}_{||}$ *is a scalar multiple of* $\mathbf{D}$*, also referred to as*

$$\mathrm{proj}_{\mathbf{D}}(\mathbf{F}),$$

and $\mathbf{F}_{\perp} \cdot \mathbf{D} = 0$.

Proof: Suppose (18.14) and $\mathbf{F}_{||} = \alpha\mathbf{D}$. 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

$$\begin{aligned} \mathbf{F}_{\perp} \cdot \mathbf{D} &= \mathbf{F} \cdot \mathbf{D} - \frac{\mathbf{F} \cdot \mathbf{D}}{|\mathbf{D}|^2}\mathbf{D} \cdot \mathbf{D} \\ &= \mathbf{F} \cdot \mathbf{D} - \mathbf{F} \cdot \mathbf{D} = 0. \end{aligned}$$

This proves the theorem.

Example 18.2.9 *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 joinging these points where distances are measured in meters.*

According to the definition, this work is

$$\begin{aligned} (2\mathbf{i} + 7\mathbf{j} - 3\mathbf{k}) \cdot (-10\mathbf{i} - 5\mathbf{j} + \mathbf{k}) &= -20 + (-35) + (-3) \\ &= -58 \text{ Newton meters.} \end{aligned}$$

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 18.2.10 *Find* $\mathrm{proj}_{\mathbf{u}}(\mathbf{v})$ *if* $\mathbf{u} = 2\mathbf{i} + 3\mathbf{j} - 4\mathbf{k}$ *and* $\mathbf{v} = \mathbf{i} - 2\mathbf{j} + \mathbf{k}$.

From the above discussion in Theorem 18.2.8, this is just

$$\begin{aligned} & \frac{1}{4+9+16}\left(\mathbf{i}-2\mathbf{j}+\mathbf{k}\right)\cdot\left(2\mathbf{i}+3\mathbf{j}-4\mathbf{k}\right)\left(2\mathbf{i}+3\mathbf{j}-4\mathbf{k}\right) \\ = \quad & \frac{-8}{29}\left(2\mathbf{i}+3\mathbf{j}-4\mathbf{k}\right)=-\frac{16}{29}\mathbf{i}-\frac{24}{29}\mathbf{j}+\frac{32}{29}\mathbf{k}. \end{aligned}$$

Example 18.2.11 *Suppose* $\mathbf{a}$, *and* $\mathbf{b}$ *are vectors and* $\mathbf{b}_\perp = \mathbf{b} - \mathrm{proj}_{\mathbf{a}}(\mathbf{b})$. *What is the magnitude of* $\mathbf{b}_\perp$ *in terms of the included angle*?

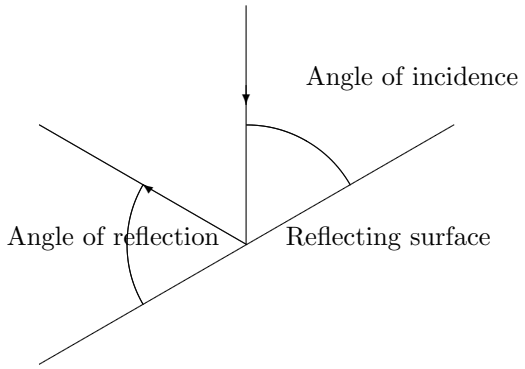
$$\begin{aligned} |\mathbf{b}_\perp|^2 &= (\mathbf{b}-\mathrm{proj}_{\mathbf{a}}(\mathbf{b}))\cdot(\mathbf{b}-\mathrm{proj}_{\mathbf{a}}(\mathbf{b})) \\ &= \left(\mathbf{b}-\frac{\mathbf{b}\cdot\mathbf{a}}{|\mathbf{a}|^2}\mathbf{a}\right)\cdot\left(\mathbf{b}-\frac{\mathbf{b}\cdot\mathbf{a}}{|\mathbf{a}|^2}\mathbf{a}\right) \\ &= |\mathbf{b}|^2-2\frac{(\mathbf{b}\cdot\mathbf{a})^2}{|\mathbf{a}|^2}+\left(\frac{\mathbf{b}\cdot\mathbf{a}}{|\mathbf{a}|^2}\right)^2|\mathbf{a}|^2 \\ &= |\mathbf{b}|^2\left(1-\frac{(\mathbf{b}\cdot\mathbf{a})^2}{|\mathbf{a}|^2|\mathbf{b}|^2}\right) \\ &= |\mathbf{b}|^2\left(1-\cos^2\theta\right)=|\mathbf{b}|^2\sin^2(\theta) \end{aligned}$$

where θ is the included angle between $\mathbf{a}$ and $\mathbf{b}$ which is less than π radians. Therefore, taking square roots,

$$|\mathbf{b}_\perp| = |\mathbf{b}|\sin\theta.$$

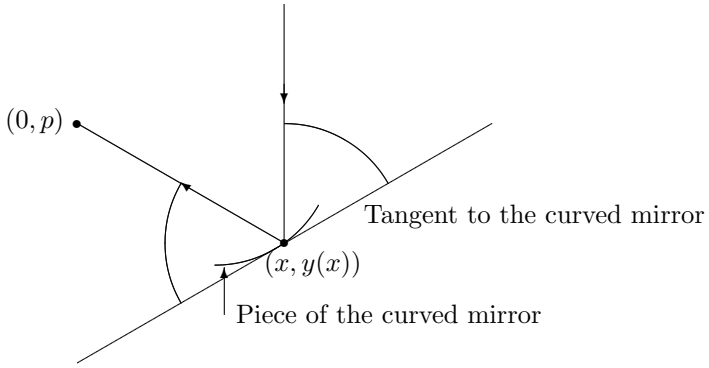
18.2.3 The Parabolic Mirror, An Application

When light is reflected the angle of incidence is always equal to the angle of reflection. This is illustrated in the following picture in which a ray of light reflects off something like a mirror.



An interesting problem is to design a curved mirror which has the property that it will direct all rays of light coming from a long distance away (essentially parallel rays of light) to a single point. You might be interested in a reflecting telescope for example or some sort of scheme for achieving high temperatures by reflecting the rays of the sun to a small area. Turning things around, you could place a source of light at the single point and desire to have the mirror reflect this in a beam of light consisting of parallel rays. How can you design such a mirror?

It turns out this is pretty easy given the above techniques for finding the angle between vectors. Consider the following picture.



It suffices to consider this in a plane for $x > 0$ and then let the mirror be obtained as a surface of revolution. In the above picture, let $(0,p)$ be the special point at which all the parallel rays of light will be directed. This is set up so the rays of light are parallel to the y axis. The two indicated angles will be equal and the equation of the indicated curve will be $y = y(x)$ while the reflection is taking place at the point $(x, y(x))$ as shown. To say the two angles are equal is to say their cosines are equal. Thus from the above,

$$\frac{(0,1)\cdot(1,y'(x))}{\sqrt{1+y'(x)^2}} = \frac{(-x,p-y)\cdot(-1,-y'(x))}{\sqrt{x^2+(y-p)^2}\sqrt{1+y'(x)^2}}.$$

This follows because the vectors forming the sides of one of the angles are $(0,1)$ and $(1,y'(x))$ while the vectors forming the other angle are $(-x,p-y)$ and $(-1,-y'(x))$. Therefore, this yields the differential equation,

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

which is written more simply as

$$\left(\sqrt{x^2+(y-p)^2}+(p-y)\right)y' = x$$

Now let $y - p = xv$ so that $y' = xv' + v$. Then in terms of v the differential equation is

$$xv' = \frac{1}{\sqrt{1+v^2}-v} - v.$$

This reduces to

$$\left(\frac{1}{\sqrt{1+v^2}-v} - v\right)\frac{dv}{dx} = \frac{1}{x}.$$

If $G \in \int \left(\frac{1}{\sqrt{1+v^2}-v} - v\right) dv$, then a solution to the differential equation is of the form

$$G(v) - \ln x = C$$

where C is a constant. This is because if you differentiate both sides with respect to x,

$$G'(v)\frac{dv}{dx} - \frac{1}{x} = \left(\frac{1}{\sqrt{1+v^2}-v} - v\right)\frac{dv}{dx} - \frac{1}{x} = 0.$$

To find $G \in \int \left(\frac{1}{\sqrt{1+v^2}-v} - v\right) dv$, use a trig. substitution, $v = \tan\theta$. Then in terms of θ, the antiderivative becomes

$$\begin{aligned}\int \left(\frac{1}{\sec\theta - \tan\theta} - \tan\theta\right) \sec^2\theta \, d\theta &= \int \sec\theta \, d\theta \\ &= \ln|\sec\theta + \tan\theta| + C.\end{aligned}$$

Now in terms of v, this is

$$\ln\left(v + \sqrt{1+v^2}\right) = \ln x + c.$$

There is no loss of generality in letting $c = \ln C$ because ln maps onto $\mathbb{R}$. Therefore, from laws of logarithms,

$$\begin{aligned}\ln\left|v + \sqrt{1+v^2}\right| &= \ln x + c = \ln x + \ln C \\ &= \ln Cx.\end{aligned}$$

Therefore,

$$v + \sqrt{1+v^2} = Cx$$

and so

$$\sqrt{1+v^2} = Cx - v.$$

Now square both sides to get

$$1 + v^2 = C^2x^2 + v^2 - 2Cxv$$

which shows

$$1 = C^2x^2 - 2Cx\frac{y-p}{x} = C^2x^2 - 2C\left(y-p\right).$$

Solving this for y yields

$$y = \frac{C}{2}x^2 + \left(p - \frac{1}{2C}\right)$$

and for this to correspond to reflection as described above, it must be that $C > 0$. As described in an earlier section, this is just the equation of a parabola. Note it is possible to choose C as desired adjusting the shape of the mirror.

18.2.4 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 18.2.12 *Let $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$. Thus $\mathbf{x} = (x_1, \cdots, x_n)$ where each $x_k \in \mathbb{C}$ and a similar formula holding for $\mathbf{y}$. Then the dot product of these two vectors is defined to be*

$$\mathbf{x} \cdot \mathbf{y} \equiv \sum_j x_j \overline{y_j} \equiv x_1\overline{y_1} + \cdots + x_n\overline{y_n}.$$

Notice how you put the conjugate on the entries of the vector, $\mathbf{y}$. 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 $\mathbf{y}$ 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. $\mathbf{u} \cdot \mathbf{v} = \overline{\mathbf{v} \cdot \mathbf{u}}$.
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 $\mathbf{u} = \mathbf{0}$.

The norm is defined in the usual way.

Definition 18.2.13 *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 18.2.14 *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 $\theta = \dfrac{\overline{z}}{|z|}$. Recall that for $z = x+iy, \overline{z} = x-iy$ and $\overline{z}z = |z|^2$.

Theorem 18.2.15 *(Cauchy Schwarz)The following inequality holds for* x_i *and* $y_i \in \mathbb{C}$.

$$|(\mathbf{x} \cdot \mathbf{y})| = \left| \sum_{i=1}^{n} x_i \overline{y}_i \right| \leq \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} \left(\sum_{i=1}^{n} |y_i|^2 \right)^{1/2} = |\mathbf{x}|\,|\mathbf{y}| \tag{18.15}$$

Proof: Let $\theta \in \mathbb{C}$ such that $|\theta| = 1$ and

$$\theta \sum_{i=1}^{n} x_i \overline{y}_i = \left| \sum_{i=1}^{n} x_i \overline{y}_i \right|$$

Thus

$$\theta \sum_{i=1}^{n} x_i \overline{y}_i = \sum_{i=1}^{n} x_i \overline{(\overline{\theta} y_i)} = \left| \sum_{i=1}^{n} x_i \overline{y}_i \right|.$$

Consider $p(t) \equiv \sum_{i=1}^{n} \left(x_i + t\overline{\theta} y_i\right) \left(\overline{x_i + t\overline{\theta} y_i}\right)$ where $t \in \mathbb{R}$.

$$\begin{aligned} 0 &\leq p(t) = \sum_{i=1}^{n} |x_i|^2 + 2t \operatorname{Re}\left(\theta \sum_{i=1}^{n} x_i \overline{y}_i \right) + t^2 \sum_{i=1}^{n} |y_i|^2 \\ &= |\mathbf{x}|^2 + 2t \left| \sum_{i=1}^{n} x_i \overline{y}_i \right| + t^2 |\mathbf{y}|^2 \end{aligned}$$

If $|\mathbf{y}| = 0$ then (18.15) is obviously true because both sides equal zero. Therefore, assume $|\mathbf{y}| \neq 0$ and then $p(t)$ is a polynomial of degree two whose graph opens up. Therefore, it either has no zeroes, two zeros or one repeated zero. If it has two zeros, the above inequality must be violated because in this case the graph must dip below the x axis. Therefore, it either has no zeros or exactly one. From the quadratic formula this happens exactly when

$$4\left|\sum_{i=1}^{n} x_i \overline{y}_i\right|^2 - 4\left|\mathbf{x}\right|^2 \left|\mathbf{y}\right|^2 \leq 0$$

and so

$$\left|\sum_{i=1}^{n} x_i \overline{y}_i\right| \leq \left|\mathbf{x}\right| \left|\mathbf{y}\right|$$

as claimed. This proves the inequality.

By analogy to the case of $\mathbb{R}^n$, length or magnitude of vectors in $\mathbb{C}^n$ can be defined.

Definition 18.2.16 *Let $\mathbf{z} \in \mathbb{C}^n$. Then $|\mathbf{z}| \equiv (\mathbf{z} \cdot \mathbf{z})^{1/2}$.*

Theorem 18.2.17 *For length defined in Definition 18.2.16, the following hold.*

$$|\mathbf{z}| \geq 0 \text{ and } |\mathbf{z}| = 0 \text{ if and only if } \mathbf{z} = \mathbf{0} \tag{18.16}$$

$$\text{If } \alpha \text{ is a scalar, } |\alpha\mathbf{z}| = |\alpha| \left|\mathbf{z}\right| \tag{18.17}$$

$$|\mathbf{z} + \mathbf{w}| \leq |\mathbf{z}| + |\mathbf{w}|. \tag{18.18}$$

Proof: The first two claims are left as exercises. To establish the third, you use the same argument which was used in $\mathbb{R}^n$.

$$\begin{aligned} |\mathbf{z} + \mathbf{w}|^2 &= (\mathbf{z} + \mathbf{w}, \mathbf{z} + \mathbf{w}) \\ &= \mathbf{z} \cdot \mathbf{z} + \mathbf{w} \cdot \mathbf{w} + \mathbf{w} \cdot \mathbf{z} + \mathbf{z} \cdot \mathbf{w} \\ &= |\mathbf{z}|^2 + |\mathbf{w}|^2 + 2\,\mathrm{Re}\,\mathbf{w} \cdot \mathbf{z} \\ &\leq |\mathbf{z}|^2 + |\mathbf{w}|^2 + 2\left|\mathbf{w} \cdot \mathbf{z}\right| \\ &\leq |\mathbf{z}|^2 + |\mathbf{w}|^2 + 2\left|\mathbf{w}\right| \left|\mathbf{z}\right| = \left(|\mathbf{z}| + |\mathbf{w}|\right)^2. \end{aligned}$$

All other considerations such as open and closed sets and the like are identical in this more general context with the corresponding definition in $\mathbb{R}^n$. The main difference is that here the scalars are complex numbers.

Definition 18.2.18 *Suppose you have a vector space, V and for $\mathbf{z}, \mathbf{w} \in V$ and α a scalar a norm is a way of measuring distance or magnitude which satisfies the properties (18.16) - (18.18). Thus a norm is something which does the following.*

$$||\mathbf{z}|| \geq 0 \text{ and } ||\mathbf{z}|| = 0 \text{ if and only if } \mathbf{z} = \mathbf{0} \tag{18.19}$$

$$\text{If } \alpha \text{ is a scalar, } ||\alpha\mathbf{z}|| = |\alpha| \left|\left|\mathbf{z}\right|\right| \tag{18.20}$$

$$||\mathbf{z} + \mathbf{w}|| \leq ||\mathbf{z}|| + ||\mathbf{w}||. \tag{18.21}$$

Here is is understood that for all $\mathbf{z} \in V, ||\mathbf{z}|| \in [0, \infty)$.

18.3 Exercises

1. Use formula (18.12) 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, $\mathbf{u} * \mathbf{v} \equiv u_1 v_1 + 2u_2 v_3 + 3u_3 v_3$. Show the axioms for a dot product all hold for this funny product. Prove $|\mathbf{u} * \mathbf{v}| \leq (\mathbf{u} * \mathbf{v})^{1/2} (\mathbf{v} * \mathbf{v})^{1/2}$. **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 $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{v} = (1, 0, -2)$ and $\mathbf{u} = (1, 2, 3)$.
6. Find $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{v} = (1, 2, -2)$ and $\mathbf{u} = (1, 0, 3)$.
7. Find $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{v} = (1, 2, -2, 1)$ and $\mathbf{u} = (1, 2, 3, 0)$.
8. Does it make sense to speak of $\text{proj}_{\mathbf{0}}(\mathbf{v})$?
9. 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 θ 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 10 pounds. How much work does this force do?
11. A boy 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. 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?
13. 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.
14. 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.
15. 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.
16. If $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ are vectors. Show that $(\mathbf{b} + \mathbf{c})_\perp = \mathbf{b}_\perp + \mathbf{c}_\perp$ where $\mathbf{b}_\perp = \mathbf{b} - \text{proj}_{\mathbf{a}}(\mathbf{b})$.
17. In the discussion of the reflecting mirror which directs all rays to a particular point, $(0, p)$. Show that for any choice of positive C this point is the focus of the parabola and the directrix is $y = p - \frac{1}{C}$.

18. Suppose you wanted to make a solar powered oven to cook food. Are there reasons for using a mirror which is not parabolic? Also describe how you would design a good flash light with a beam which does not spread out too quickly.

19. Find $(1,2,3,4)\cdot(2,0,1,3)$.

20. Show that $(\mathbf{a}\cdot\mathbf{b}) = \frac{1}{4}\left[|\mathbf{a}+\mathbf{b}|^2 - |\mathbf{a}-\mathbf{b}|^2\right]$.

21. Prove from the axioms of the dot product the parallelogram identity, $|\mathbf{a}+\mathbf{b}|^2 + |\mathbf{a}-\mathbf{b}|^2 = 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2$.

22. Let A and be a real $m\times n$ matrix and let $\mathbf{x}\in\mathbb{R}^n$ and $\mathbf{y}\in\mathbb{R}^m$. Show $(A\mathbf{x},\mathbf{y})_{\mathbb{R}^m} = \left(\mathbf{x},A^T\mathbf{y}\right)_{\mathbb{R}^n}$ where $(\cdot,\cdot)_{\mathbb{R}^k}$ denotes the dot product in $\mathbb{R}^k$. In the notation above, $A\mathbf{x}\cdot\mathbf{y} = \mathbf{x}\cdot A^T\mathbf{y}$. Use the definition of matrix multiplication to do this.

23. Use the result of Problem 22 to verify directly that $(AB)^T = B^TA^T$ without making any reference to subscripts.

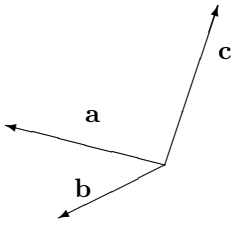
24. Suppose f,g are two continuous functions defined on $[0,1]$. Define $(f\cdot g) = \int_0^1 f(x)\,g(x)\,dx$. Show this dot product satisfies conditons (18.1) - (18.5). Explain why the Cauchy Schwarz inequality continues to hold in this context and state the Cauchy Schwarz inequality in terms of integrals.

18.4 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 18.4.1 *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 our 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 18.4.2 *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two vectors in* $\mathbb{R}^n$. *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* θ *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 spanned by $\mathbf{a}$ and $\mathbf{b}$.
The cross product satisfies the following properties.

$$\mathbf{a}\times\mathbf{b} = -(\mathbf{b}\times\mathbf{a})\ ,\ \mathbf{a}\times\mathbf{a} = \mathbf{0}, \tag{18.22}$$

For α a scalar,

$$(\alpha\mathbf{a})\times\mathbf{b} = \alpha(\mathbf{a}\times\mathbf{b}) = \mathbf{a}\times(\alpha\mathbf{b}), \tag{18.23}$$

For $\mathbf{a},\mathbf{b}$, and $\mathbf{c}$ vectors, one obtains the distributive laws,

$$\mathbf{a}\times(\mathbf{b}+\mathbf{c}) = \mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c}, \tag{18.24}$$

$$(\mathbf{b}+\mathbf{c})\times\mathbf{a} = \mathbf{b}\times\mathbf{a}+\mathbf{c}\times\mathbf{a}. \tag{18.25}$$

Formula (18.22) 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 (18.23) is also fairly clear. If α is a nonnegative scalar, the direction of $(\alpha\mathbf{a})\times\mathbf{b}$ is the same as the direction of $\mathbf{a}\times\mathbf{b}$,$\alpha(\mathbf{a}\times\mathbf{b})$ and $\mathbf{a}\times(\alpha\mathbf{b})$ while the magnitude is just α 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 (18.23). 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 (18.22),

$$\begin{aligned}(\mathbf{b}+\mathbf{c})\times\mathbf{a} &= -\mathbf{a}\times(\mathbf{b}+\mathbf{c})\\ &= -(\mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c})\\ &= \mathbf{b}\times\mathbf{a}+\mathbf{c}\times\mathbf{a}.\end{aligned}$$

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,

$$\begin{array}{ll}\mathbf{i}\times\mathbf{j}=\mathbf{k} & \mathbf{j}\times\mathbf{i}=-\mathbf{k}\\ \mathbf{k}\times\mathbf{i}=\mathbf{j} & \mathbf{i}\times\mathbf{k}=-\mathbf{j}\\ \mathbf{j}\times\mathbf{k}=\mathbf{i} & \mathbf{k}\times\mathbf{j}=-\mathbf{i}\end{array}$$

With this information, the following gives the coordinate description of the cross product.

Proposition 18.4.3 *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*

$$\begin{aligned}\mathbf{a}\times\mathbf{b} = &(a_2b_3-a_3b_2)\,\mathbf{i}+(a_3b_1-a_1b_3)\,\mathbf{j}+\\ &+(a_1b_2-a_2b_1)\,\mathbf{k}.\end{aligned} \tag{18.26}$$

Proof: From the above table and the properties of the cross product listed,

$$(a_1\mathbf{i}+a_2\mathbf{j}+a_3\mathbf{k})\times(b_1\mathbf{i}+b_2\mathbf{j}+b_3\mathbf{k})=$$

$$a_1b_2\mathbf{i}\times\mathbf{j}+a_1b_3\mathbf{i}\times\mathbf{k}+a_2b_1\mathbf{j}\times\mathbf{i}+a_2b_3\mathbf{j}\times\mathbf{k}+$$

$$+a_3b_1\mathbf{k}\times\mathbf{i}+a_3b_2\mathbf{k}\times\mathbf{j}$$

$$=a_1b_2\mathbf{k}-a_1b_3\mathbf{j}-a_2b_1\mathbf{k}+a_2b_3\mathbf{i}+a_3b_1\mathbf{j}-a_3b_2\mathbf{i}$$

$$=(a_2b_3-a_3b_2)\,\mathbf{i}+(a_3b_1-a_1b_3)\,\mathbf{j}+(a_1b_2-a_2b_1)\,\mathbf{k} \tag{18.27}$$

This proves the proposition.

It is probably impossible for most people to remember (18.26). Fortunately, there is a somewhat easier way to remember it.

$$\mathbf{a}\times\mathbf{b}=\begin{vmatrix}\mathbf{i}&\mathbf{j}&\mathbf{k}\\a_1&a_2&a_3\\b_1&b_2&b_3\end{vmatrix} \tag{18.28}$$

where you expand the determinant along the top row. This yields

$$(a_2b_3-a_3b_2)\,\mathbf{i}-(a_1b_3-a_3b_1)\,\mathbf{j}+(a_1b_2-a_2b_1)\,\mathbf{k} \tag{18.29}$$

which is the same as (18.27).

Example 18.4.4 *Find* $(\mathbf{i}-\mathbf{j}+2\mathbf{k})\times(3\mathbf{i}-2\mathbf{j}+\mathbf{k})$.

Use (18.28) to compute this.

$$\begin{vmatrix}\mathbf{i}&\mathbf{j}&\mathbf{k}\\1&-1&2\\3&-2&1\end{vmatrix}=\begin{vmatrix}-1&2\\-2&1\end{vmatrix}\mathbf{i}-\begin{vmatrix}1&2\\3&1\end{vmatrix}\mathbf{j}+\begin{vmatrix}1&-1\\3&-2\end{vmatrix}\mathbf{k}$$

$$=3\mathbf{i}+5\mathbf{j}+\mathbf{k}.$$

Example 18.4.5 *Find the area of the parallelogram determined by the vectors,* $(\mathbf{i}-\mathbf{j}+2\mathbf{k})$ *and* $(3\mathbf{i}-2\mathbf{j}+\mathbf{k})$. *These are the same two vectors in Example 18.4.4.*

From Example 18.4.4 and the geometric description of the cross product, the area is just the norm of the vector obtained in Example 18.4.4. Thus the area is $\sqrt{9+25+1}=\sqrt{35}$.

Example 18.4.6 *Find the area of the triangle determined by* $(1,2,3)$, $(0,2,5)$, *and* $(5,1,2)$.

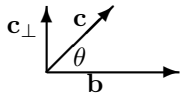
This triangle is obtained by connecting the three points with lines. Picking $(1,2,3)$ as a starting point, there are two displacement vectors, $(-1,0,2)$ and $(4,-1,-1)$ such that the given vector added to these displacement vectors gives the other two vectors. The area of the triangle is half the area 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 $\frac{1}{2}\sqrt{4+49+1}=\frac{3}{2}\sqrt{6}$.

18.4.1 The Distributive Law For The Cross Product

This section gives a proof for (18.24), 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 [7]. Another approach, based on volumes of parallelepipeds is found in [25] and is discussed a little later.

Lemma 18.4.7 *Let* $\mathbf{b}$ *and* $\mathbf{c}$ *be two vectors. Then* $\mathbf{b}\times\mathbf{c}=\mathbf{b}\times\mathbf{c}_\perp$ *where* $\mathbf{c}_{||}+\mathbf{c}_\perp=\mathbf{c}$ *and* $\mathbf{c}_\perp\cdot\mathbf{b}=0$.

Proof: Consider the following picture.



Now $\mathbf{c}_\perp=\mathbf{c}-\mathbf{c}\cdot\frac{\mathbf{b}}{|\mathbf{b}|}\frac{\mathbf{b}}{|\mathbf{b}|}$ and so $\mathbf{c}_\perp$ is in the plane determined by $\mathbf{c}$ and $\mathbf{b}$. Therefore, from the geometric definition of the cross product, $\mathbf{b}\times\mathbf{c}$ and $\mathbf{b}\times\mathbf{c}_\perp$ have the same direction. Now, referring to the picture,

$$\begin{aligned}|\mathbf{b}\times\mathbf{c}_\perp| &= |\mathbf{b}|\,|\mathbf{c}_\perp| \\ &= |\mathbf{b}|\,|\mathbf{c}|\sin\theta \\ &= |\mathbf{b}\times\mathbf{c}|\,.\end{aligned}$$

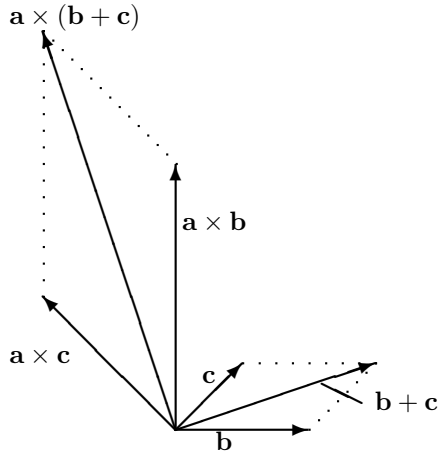
Therefore, $\mathbf{b}\times\mathbf{c}$ and $\mathbf{b}\times\mathbf{c}_\perp$ also have the same magnitude and so they are the same vector.

With this, the proof of the distributive law is in the following theorem.

Theorem 18.4.8 *Let* $\mathbf{a},\mathbf{b}$, *and* $\mathbf{c}$ *be vectors in* $\mathbb{R}^3$. *Then*

$$\mathbf{a}\times(\mathbf{b}+\mathbf{c})=\mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c} \tag{18.30}$$

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,

$$|\mathbf{a}\times\mathbf{b}| = |\mathbf{a}|\,|\mathbf{b}|\,, |\mathbf{a}\times(\mathbf{b}+\mathbf{c})| = |\mathbf{a}|\,|\mathbf{b}+\mathbf{c}|\,, \text{ and}$$

$$|\mathbf{a}\times\mathbf{c}| = |\mathbf{a}|\,|\mathbf{c}|\,.$$

Therefore,

$$\frac{|\mathbf{a}\times(\mathbf{b}+\mathbf{c})|}{|\mathbf{b}+\mathbf{c}|} = \frac{|\mathbf{a}\times\mathbf{c}|}{|\mathbf{c}|} = \frac{|\mathbf{a}\times\mathbf{b}|}{|\mathbf{b}|} = |\mathbf{a}|$$

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 (18.30).

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,

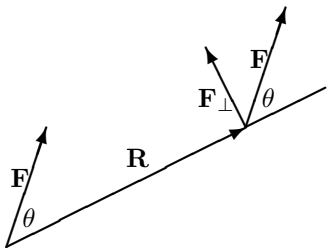
$$\begin{aligned}\mathbf{a}\times(\mathbf{b}+\mathbf{c}) &= \mathbf{a}\times(\mathbf{b}+\mathbf{c})_{\perp}\\ &= \mathbf{a}\times(\mathbf{b}_{\perp}+\mathbf{c}_{\perp})\\ &= \mathbf{a}\times\mathbf{b}_{\perp} + \mathbf{a}\times\mathbf{c}_{\perp}\\ &= \mathbf{a}\times\mathbf{b} + \mathbf{a}\times\mathbf{c}.\end{aligned}$$

This proves the theorem.

The result of Problem 16 of the exercises 18.3 is used to go from the first to the second line.

18.4.2 Torque

Imagine you are using a wrench to loosen a nut. The idea is to turn the nut by applying a force to the end of the wrench. If you push or pull the wrench directly toward or away from the nut, it should be obvious from experience that no progress will be made in turning the nut. The important thing is the component of force perpendicular to the wrench. It is this component of force which will cause the nut to turn. For example see the following picture.



In the picture a force, $\mathbf{F}$ is applied at the end of a wrench represented by the position vector, $\mathbf{R}$ and the angle between these two is θ. Then the tendency to turn will be $|\mathbf{R}|\,|\mathbf{F}_{\perp}| =$

$|\mathbf{R}|\,|\mathbf{F}|\sin\theta$, which you recognize as the magnitude of the cross product of $\mathbf{R}$ and $\mathbf{F}$. If there were just one force acting at one point whose position vector is $\mathbf{R}$, perhaps this would be sufficient, but what if there are numerous forces acting at many different points with neither the position vectors nor the force vectors in the same plane; what then? To keep track of this sort of thing, define for each $\mathbf{R}$ and $\mathbf{F}$, the torque vector,

$$\tau \equiv \mathbf{R} \times \mathbf{F}.$$

This is also called the moment of the force, $\mathbf{F}$. That way, if there are several forces acting at several points, the total torque can be obtained by simply adding up the torques associated with the different forces and positions.

Example 18.4.9 *Suppose* $\mathbf{R}_1 = 2\mathbf{i} - \mathbf{j} + 3\mathbf{k}, \mathbf{R}_2 = \mathbf{i} + 2\mathbf{j} - 6\mathbf{k}$ *meters and at the points determined by these vectors there are forces,* $\mathbf{F}_1 = \mathbf{i} - \mathbf{j} + 2\mathbf{k}$ *and* $\mathbf{F}_2 = \mathbf{i} - 5\mathbf{j} + \mathbf{k}$ *Newtons respectively. Find the total torque about the origin produced by these forces acting at the given points.*

It is necessary to take $\mathbf{R}_1 \times \mathbf{F}_1 + \mathbf{R}_2 \times \mathbf{F}_2$. Thus the total torque equals

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 2 & -1 & 3 \\ 1 & -1 & 2 \end{vmatrix} + \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 2 & -6 \\ 1 & -5 & 1 \end{vmatrix} = -27\mathbf{i} - 8\mathbf{j} - 8\mathbf{k} \text{ Newton meters}$$

Example 18.4.10 *Find if possible a single force vector,* $\mathbf{F}$ *which if applied at the point* $\mathbf{i} + \mathbf{j} + \mathbf{k}$ *will produce the same torque as the above two forces acting at the given points.*

This is fairly routine. The problem is to find $\mathbf{F} = F_1\mathbf{i} + F_2\mathbf{j} + F_3\mathbf{k}$ which produces the above torque vector. Therefore,

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 1 & 1 \\ F_1 & F_2 & F_3 \end{vmatrix} = -27\mathbf{i} - 8\mathbf{j} - 8\mathbf{k}$$

which reduces to $(F_3 - F_2)\,\mathbf{i} + (F_1 - F_3)\,\mathbf{j} + (F_2 - F_1)\,\mathbf{k} = -27\mathbf{i} - 8\mathbf{j} - 8\mathbf{k}$. This amounts to solving the system of three equations in three unknowns, F_1, F_2, and F_3,

$$\begin{aligned} F_3 - F_2 &= -27 \\ F_1 - F_3 &= -8 \\ F_2 - F_1 &= -8 \end{aligned}$$

However, there is no solution to these three equations. (Why?) Therefore no single force acting at the point $\mathbf{i} + \mathbf{j} + \mathbf{k}$ will produce the given torque.

18.4.3 Center Of Mass

The mass of an object is a measure of how much stuff there is in the object. An object has mass equal to one kilogram, a unit of mass in the metric system, if it would exactly balance a known one kilogram object when placed on a balance. The known object is one kilogram by definition. The mass of an object does not depend on where the balance is used. It would be one kilogram on the moon as well as on the earth. The weight of an object is something else. It is the force exerted on the object by gravity and has magnitude gm where g is a constant called the acceleration of gravity. Thus the weight of a one kilogram object would be different on the moon which has much less gravity, smaller g, than on the earth. An important idea is that of the center of mass. This is the point at which an object will balance no matter how it is turned.

Definition 18.4.11 *Let an object consist of p point masses, $m_1, \cdots, m_p$ with the position of the k^{th} of these at $\mathbf{R}_k$. The center of mass of this object, $\mathbf{R}_0$ is the point satisfying*

$$\sum_{k=1}^{p} (\mathbf{R}_k - \mathbf{R}_0) \times g m_k \mathbf{u} = \mathbf{0}$$

for all unit vectors, $\mathbf{u}$.

The above definition indicates that no matter how the object is suspended, the total torque on it due to gravity is such that no rotation occurs. Using the properties of the cross product,

$$\left(\sum_{k=1}^{p} \mathbf{R}_k g m_k - \mathbf{R}_0 \sum_{k=1}^{p} g m_k \right) \times \mathbf{u} = \mathbf{0} \tag{18.31}$$

for any choice of unit vector, $\mathbf{u}$. You should verify that if $\mathbf{a} \times \mathbf{u} = \mathbf{0}$ for all $\mathbf{u}$, then it must be the case that $\mathbf{a} = \mathbf{0}$. Then the above formula requires that

$$\sum_{k=1}^{p} \mathbf{R}_k g m_k - \mathbf{R}_0 \sum_{k=1}^{p} g m_k = \mathbf{0}.$$

dividing by g, and then by $\sum_{k=1}^{p} m_k$,

$$\mathbf{R}_0 = \frac{\sum_{k=1}^{p} \mathbf{R}_k m_k}{\sum_{k=1}^{p} m_k}. \tag{18.32}$$

This is the formula for the center of mass of a collection of point masses. To consider the center of mass of a solid consisting of continuously distributed masses, you need the methods of calculus.

Example 18.4.12 *Let $m_1 = 5, m_2 = 6$, and $m_3 = 3$ where the masses are in kilograms. Suppose m_1 is located at $2\mathbf{i}+3\mathbf{j}+\mathbf{k}$, m_2 is located at $\mathbf{i}-3\mathbf{j}+2\mathbf{k}$ and m_3 is located at $2\mathbf{i}-\mathbf{j}+3\mathbf{k}$. Find the center of mass of these three masses.*

Using (18.32)

$$\begin{aligned} \mathbf{R}_0 &= \frac{5(2\mathbf{i}+3\mathbf{j}+\mathbf{k}) + 6(\mathbf{i}-3\mathbf{j}+2\mathbf{k}) + 3(2\mathbf{i}-\mathbf{j}+3\mathbf{k})}{5+6+3} \\ &= \frac{11}{7}\mathbf{i} - \frac{3}{7}\mathbf{j} + \frac{13}{7}\mathbf{k} \end{aligned}$$

18.4.4 Angular Velocity

Definition 18.4.13 *In a rotating body, a vector, $\mathbf{\Omega}$ is called an **angular velocity vector** if the velocity of a point having position vector, $\mathbf{u}$ relative to the body is given by $\mathbf{\Omega} \times \mathbf{u}$.*

The existence of an angular velocity vector is the key to understanding motion in a moving system of coordinates. It is used to explain the motion on the surface of the rotating earth. For example, have you ever wondered why low pressure areas rotate counter clockwise in the upper hemisphere but clockwise in the lower hemisphere? To quantify these things, you will need the concept of an angular velocity vector. Details are presented later for interesting examples. Here we take a simple example. In the above example, think of a coordinate system fixed in the rotating body. Thus if you were riding on the rotating body, you would observe this coordinate system as fixed but it is not fixed.

Example 18.4.14 *A wheel rotates counter clockwise about the vector* $\mathbf{i}+\mathbf{j}+\mathbf{k}$ *at 60 revolutions per minute. This means that if the thumb of your right hand were to point in the direction of* $\mathbf{i}+\mathbf{j}+\mathbf{k}$ *your fingers of this hand would wrap in the direction of rotation. Find the angular velocity vector for this wheel. Assume the unit of distance is meters and the unit of time is minutes.*

Let $\omega = 60 \times 2\pi = 120\pi$. This is the number of radians per minute corresponding to 60 revolutions per minute. Then the angular velocity vector is $\frac{120\pi}{\sqrt{3}}(\mathbf{i}+\mathbf{j}+\mathbf{k})$. Note this gives what you would expect in the case the position vector to the point is perpendicular to $\mathbf{i}+\mathbf{j}+\mathbf{k}$ and at a distance of r. This is because of the geometric description of the cross product. The magnitude of the vector is $r120\pi$ meters per minute and corresponds to the speed and an excercise with the right hand shows the direction is correct also. However, if this body is rigid, this will work for every other point in it, even those for which the position vector is not perpendicular to the given vector. A complete analysis of this is given later.

Example 18.4.15 *A wheel rotates counter clockwise about the vector* $\mathbf{i}+\mathbf{j}+\mathbf{k}$ *at 60 revolutions per minute exactly as in Example 18.4.14. Let* $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\}$ *denote an orthogonal right handed system attached to the rotating wheel in which* $\mathbf{u}_3 = \frac{1}{\sqrt{3}}(\mathbf{i}+\mathbf{j}+\mathbf{k})$. *Thus* $\mathbf{u}_1$ *and* $\mathbf{u}_2$ *depend on time. Find the velocity of the point of the wheel located at the point* $2\mathbf{u}_1+3\mathbf{u}_2-\mathbf{u}_3$. *Note this point is not fixed in space. It is moving.*

Since $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\}$ is a right handed system like $\mathbf{i},\mathbf{j},\mathbf{k}$, everything applies to this system in the same way as with $\mathbf{i},\mathbf{j},\mathbf{k}$. Thus the cross product is given by

$$\begin{aligned}&(a\mathbf{u}_1+b\mathbf{u}_2+c\mathbf{u}_3)\times(d\mathbf{u}_1+e\mathbf{u}_2+f\mathbf{u}_3)\\ =\ &\begin{vmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3\\ a & b & c\\ d & e & f\end{vmatrix}\end{aligned}$$

Therefore, in terms of the given vectors $\mathbf{u}_i$, the angular velocity vector is

$$120\pi\mathbf{u}_3$$

the velocity of the given point is

$$\begin{aligned}&\begin{vmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3\\ 0 & 0 & 120\pi\\ 2 & 3 & -1\end{vmatrix}\\ =\ &-360\pi\mathbf{u}_1+240\pi\mathbf{u}_2\end{aligned}$$

in meters per minute. Note how this gives the answer in terms of these vectors which are fixed in the body, not in space. Since $\mathbf{u}_i$ depends on t, this shows the answer in this case does also. Of course this is right. Just think of what is going on with the wheel rotating. Those vectors which are fixed in the wheel are moving in space. The velocity of a point in the wheel should be constantly changing. However, its speed will not change. The speed will be the magnitude of the velocity and this is

$$\sqrt{(-360\pi\mathbf{u}_1+240\pi\mathbf{u}_2)\cdot(-360\pi\mathbf{u}_1+240\pi\mathbf{u}_2)}$$

which from the properties of the dot product equals

$$\sqrt{(-360\pi)^2+(240\pi)^2}=120\sqrt{13}\pi$$

because the $\mathbf{u}_i$ are given to be orthogonal.

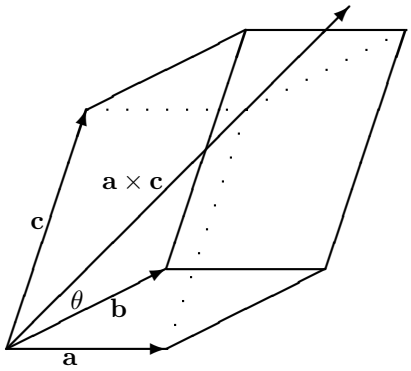
18.4.5 The Box Product

Definition 18.4.16 *A parallelepiped determined by the three vectors,* $\mathbf{a}, \mathbf{b}$, *and* $\mathbf{c}$ *consists of*

$$\{r\mathbf{a}+s\mathbf{b}+t\mathbf{b} : r, s, t \in [0,1]\}.$$

That is, if you pick three numbers, r, s, *and* t *each in* $[0,1]$ *and form* $r\mathbf{a}+s\mathbf{b}+t\mathbf{b}$, *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.



You notice the area of the base of the parallelepiped, the parallelogram determined by the vectors, $\mathbf{a}$ and $\mathbf{c}$ has area equal to $|\mathbf{a}\times\mathbf{c}|$ while the altitude of the parallelepiped is $|\mathbf{b}|\cos\theta$ where θ is the angle shown in the picture between $\mathbf{b}$ and $\mathbf{a}\times\mathbf{c}$. Therefore, the volume of this parallelepiped is the area of the base times the altitude which is just

$$|\mathbf{a}\times\mathbf{c}|\,|\mathbf{b}|\cos\theta = \mathbf{a}\times\mathbf{c}\cdot\mathbf{b}.$$

This expression is known as the box product and is sometimes written as $[\mathbf{a},\mathbf{c},\mathbf{b}]$. 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 18.4.17 *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.

$$\begin{aligned}(\mathbf{i}+2\mathbf{j}-5\mathbf{k})\times(\mathbf{i}+3\mathbf{j}-6\mathbf{k}) &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 2 & -5 \\ 1 & 3 & -6 \end{vmatrix} \\ &= 3\mathbf{i}+\mathbf{j}+\mathbf{k}\end{aligned}$$

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 18.4.18 *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 parallellepiped or they both give -1 times the volume.

An Alternate 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,

$$\begin{aligned} \mathbf{x} \cdot \mathbf{a} \times (\mathbf{b} + \mathbf{c}) &= (\mathbf{x} \times \mathbf{a}) \cdot (\mathbf{b} + \mathbf{c}) \\ &= (\mathbf{x} \times \mathbf{a}) \cdot \mathbf{b} + (\mathbf{x} \times \mathbf{a}) \cdot \mathbf{c} \\ &= \mathbf{x} \cdot \mathbf{a} \times \mathbf{b} + \mathbf{x} \cdot \mathbf{a} \times \mathbf{c} \\ &= \mathbf{x} \cdot (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}) . \end{aligned}$$

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.

Observation 18.4.19 *Suppose you have three vectors,* $\mathbf{u} = (a, b, c)$, $\mathbf{v} = (d, e, f)$, *and* $\mathbf{w} = (g, h, i)$. *Then* $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w}$ *is given by the following.*

$$\begin{aligned} \mathbf{u} \cdot \mathbf{v} \times \mathbf{w} &= (a, b, c) \cdot \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ d & e & f \\ g & h & i \end{vmatrix} \\ &= a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= \det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} . \end{aligned}$$

The message is that to take the box product, you can simply take the determinant of the matrix which results by letting the rows be the rectangular components of the given vectors in the order in which they occur in the box product.

18.5 Vector Identities And Notation

To begin with consider $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ and it is desired to simplify this quantity. It turns out this is an important quantity which comes up in many different contexts. Let $\mathbf{u} = (u_1, u_2, u_3)$ and let $\mathbf{v}$ and $\mathbf{w}$ be defined similarly.

$$\begin{aligned} \mathbf{v} \times \mathbf{w} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} \\ &= (v_2 w_3 - v_3 w_2) \mathbf{i} + (w_1 v_3 - v_1 w_3) \mathbf{j} + (v_1 w_2 - v_2 w_1) \mathbf{k} \end{aligned}$$

Next consider $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ which is given by

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ (v_2 w_3 - v_3 w_2) & (w_1 v_3 - v_1 w_3) & (v_1 w_2 - v_2 w_1) \end{vmatrix} .$$

When you multiply this out, you get

$$\begin{aligned}&\mathbf{i}\left(v_1u_2w_2+u_3v_1w_3-w_1u_2v_2-u_3w_1v_3\right)+\mathbf{j}\left(v_2u_1w_1+v_2w_3u_3-w_2u_1v_1-u_3w_2v_3\right)\\&+\mathbf{k}\left(u_1w_1v_3+v_3w_2u_2-u_1v_1w_3-v_2w_3u_2\right)\end{aligned}$$

and if you are clever, you see right away that

$$\left(\mathbf{i}v_1+\mathbf{j}v_2+\mathbf{k}v_3\right)\left(u_1w_1+u_2w_2+u_3w_3\right)-\left(\mathbf{i}w_1+\mathbf{j}w_2+\mathbf{k}w_3\right)\left(u_1v_1+u_2v_2+u_3v_3\right).$$

Thus

$$\mathbf{u}\times\left(\mathbf{v}\times\mathbf{w}\right)=\mathbf{v}\left(\mathbf{u}\cdot\mathbf{w}\right)-\mathbf{w}\left(\mathbf{u}\cdot\mathbf{v}\right). \tag{18.33}$$

A related formula is

$$\begin{aligned}\left(\mathbf{u}\times\mathbf{v}\right)\times\mathbf{w} &= -\left[\mathbf{w}\times\left(\mathbf{u}\times\mathbf{v}\right)\right]\\ &= -\left[\mathbf{u}\left(\mathbf{w}\cdot\mathbf{v}\right)-\mathbf{v}\left(\mathbf{w}\cdot\mathbf{u}\right)\right]\\ &= \mathbf{v}\left(\mathbf{w}\cdot\mathbf{u}\right)-\mathbf{u}\left(\mathbf{w}\cdot\mathbf{v}\right).\end{aligned} \tag{18.34}$$

This derivation is simply wretched and it does nothing for other identities which may arise in applications. Actually, the above two formulas, (18.33) and (18.34) are sufficient for most applications if you are creative in using them, but there is another way. This other way allows you to discover such vector identities as the above without any creativity or any cleverness. Therefore, it is far superior to the above nasty computation. It is a vector identity discovering machine and it is this which is the main topic in what follows.

There are two special symbols, δ_{ij} and ε_{ijk} which are very useful in dealing with vector identities. To begin with, here is the definition of these symbols.

Definition 18.5.1 *The symbol, δ_{ij}, called the Kroneker delta symbol is defined as follows.*

$$\delta_{ij}\equiv\begin{cases}1 \text{ if } i=j\\ 0 \text{ if } i\neq j\end{cases}.$$

With the Kroneker symbol, i and j can equal any integer in $\{1,2,\cdots,n\}$ for any $n\in\mathbb{N}$.

Definition 18.5.2 *For i,j, and k integers in the set, $\{1,2,3\}$, ε_{ijk} is defined as follows.*

$$\varepsilon_{ijk}\equiv\begin{cases}1 \text{ if } (i,j,k)=(1,2,3),(2,3,1), \text{ or } (3,1,2)\\ -1 \text{ if } (i,j,k)=(2,1,3),(1,3,2), \text{ or } (3,2,1)\\ 0 \text{ if there are any repeated integers}\end{cases}.$$

The subscripts ijk and ij in the above are called indices. A single one is called an index. This symbol, ε_{ijk} is also called the permutation symbol.

The way to think of ε_{ijk} is that $\varepsilon_{123}=1$ and if you switch any two of the numbers in the list i,j,k, it changes the sign. Thus $\varepsilon_{ijk}=-\varepsilon_{jik}$ and $\varepsilon_{ijk}=-\varepsilon_{kji}$ etc. You should check that this rule reduces to the above definition. For example, it immediately implies that if there is a repeated index, the answer is zero. This follows because $\varepsilon_{iij}=-\varepsilon_{iij}$ and so $\varepsilon_{iij}=0$.

It is useful to use the Einstein summation convention when dealing with these symbols. Simply stated, the convention is that you sum over the repeated index. Thus a_ib_i means $\sum_i a_ib_i$. Also, $\delta_{ij}x_j$ means $\sum_j\delta_{ij}x_j=x_i$. When you use this convention, there is one very important thing to never forget. It is this: Never have an index be repeated more than once. Thus a_ib_i is all right but $a_{ii}b_i$ is not. The reason for this is that you end up getting confused about what is meant. If you want to write $\sum_i a_ib_ic_i$ it is best to simply use the summation notation. There is a very important reduction identity connecting these two symbols.

Lemma 18.5.3 *The following holds.*

$$\varepsilon_{ijk}\varepsilon_{irs} = (\delta_{jr}\delta_{ks} - \delta_{kr}\delta_{js}).$$

Proof: If $\{j,k\} \neq \{r,s\}$ then every term in the sum on the left must have either ε_{ijk} or ε_{irs} contains a repeated index. Therefore, the left side equals zero. The right side also equals zero in this case. To see this, note that if the two sets are not equal, then there is one of the indices in one of the sets which is not in the other set. For example, it could be that j is not equal to either r or s. Then the right side equals zero.

Therefore, it can be assumed $\{j,k\} = \{r,s\}$. If $i = r$ and $j = s$ for $s \neq r$, then there is exactly one term in the sum on the left and it equals 1. The right also reduces to 1 in this case. If $i = s$ and $j = r$, there is exactly one term in the sum on the left which is nonzero and it must equal -1. The right side also reduces to -1 in this case. If there is a repeated index in $\{j,k\}$, then every term in the sum on the left equals zero. The right also reduces to zero in this case because then $j = k = r = s$ and so the right side becomes $(1)(1) - (-1)(-1) = 0$.

Proposition 18.5.4 *Let* $\mathbf{u}, \mathbf{v}$ *be vectors in* $\mathbb{R}^n$ *where the Cartesian coordinates of* $\mathbf{u}$ *are* $(u_1, \cdots, u_n)$ *and the Cartesian coordinates of* $\mathbf{v}$ *are* $(v_1, \cdots, v_n)$. *Then* $\mathbf{u} \cdot \mathbf{v} = u_i v_i$. *If* $\mathbf{u}, \mathbf{v}$ *are vectors in* $\mathbb{R}^3$, *then*

$$(\mathbf{u} \times \mathbf{v})_i = \varepsilon_{ijk} u_j v_k.$$

Also, $\delta_{ik} a_k = a_i$.

Proof: The first claim is obvious from the definition of the dot product. The second is verified by simply checking it works. For example,

$$\mathbf{u} \times \mathbf{v} \equiv \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}$$

and so

$$(\mathbf{u} \times \mathbf{v})_1 = (u_2 v_3 - u_3 v_2).$$

From the above formula in the proposition,

$$\varepsilon_{1jk} u_j v_k \equiv u_2 v_3 - u_3 v_2,$$

the same thing. The cases for $(\mathbf{u} \times \mathbf{v})_2$ and $(\mathbf{u} \times \mathbf{v})_3$ are verified similarly. The last claim follows directly from the definition.

With this notation, you can easily discover vector identities and simplify expressions which involve the cross product.

Example 18.5.5 *Discover a formula which simplifies* $(\mathbf{u} \times \mathbf{v}) \times \mathbf{w}$.

From the above reduction formula,

$$\begin{aligned} ((\mathbf{u} \times \mathbf{v}) \times \mathbf{w})_i &= \varepsilon_{ijk} (\mathbf{u} \times \mathbf{v})_j w_k \\ &= \varepsilon_{ijk}\varepsilon_{jrs} u_r v_s w_k \\ &= -\varepsilon_{jik}\varepsilon_{jrs} u_r v_s w_k \\ &= -(\delta_{ir}\delta_{ks} - \delta_{is}\delta_{kr}) u_r v_s w_k \\ &= -(u_i v_k w_k - u_k v_i w_k) \\ &= \mathbf{u} \cdot \mathbf{w} v_i - \mathbf{v} \cdot \mathbf{w} u_i \\ &= ((\mathbf{u} \cdot \mathbf{w})\mathbf{v} - (\mathbf{v} \cdot \mathbf{w})\mathbf{u})_i. \end{aligned}$$

Since this holds for all i, it follows that

$$(\mathbf{u}\times\mathbf{v})\times\mathbf{w}=(\mathbf{u}\cdot\mathbf{w})\,\mathbf{v}-(\mathbf{v}\cdot\mathbf{w})\,\mathbf{u}.$$

This is good notation and it will be used in the rest of the book whenever convenient.

18.6 Exercises

1. Show that if $\mathbf{a}\times\mathbf{u}=\mathbf{0}$ for all unit vectors, $\mathbf{u}$, then $\mathbf{a}=\mathbf{0}$.
2. If you only assume (18.31) holds for $\mathbf{u}=\mathbf{i},\mathbf{j},\mathbf{k}$, show that this implies (18.31) holds for all unit vectors, $\mathbf{u}$.
3. Let $m_1=5, m_2=1$, and $m_3=4$ where the masses are in kilograms and the distance is in meters. Suppose m_1 is located at $2\mathbf{i}-3\mathbf{j}+\mathbf{k}$, m_2 is located at $\mathbf{i}-3\mathbf{j}+6\mathbf{k}$ and m_3 is located at $2\mathbf{i}+\mathbf{j}+3\mathbf{k}$. Find the center of mass of these three masses.
4. Let $m_1=2, m_2=3$, and $m_3=1$ where the masses are in kilograms and the distance is in meters. Suppose m_1 is located at $2\mathbf{i}-\mathbf{j}+\mathbf{k}$, m_2 is located at $\mathbf{i}-2\mathbf{j}+\mathbf{k}$ and m_3 is located at $4\mathbf{i}+\mathbf{j}+3\mathbf{k}$. Find the center of mass of these three masses.
5. Find the angular velocity vector of a rigid body which rotates counter clockwise about the vector $\mathbf{i}-2\mathbf{j}+\mathbf{k}$ at 40 revolutions per minute. Assume distance is measured in meters.
6. Let $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\}$ be a right handed system with $\mathbf{u}_3$ pointing in the direction of $\mathbf{i}-2\mathbf{j}+\mathbf{k}$ and $\mathbf{u}_1$ and $\mathbf{u}_2$ being fixed with the body which is rotating at 40 revolutions per minute. Assuming all distances are in meters, find the constant speed of the point of the body located at $3\mathbf{u}_1+\mathbf{u}_2-\mathbf{u}_3$ in meters per minute.
7. Find the area of the triangle determined by the three points, $(1,2,3),(4,2,0)$ and $(-3,2,1)$.
8. Find the area of the triangle determined by the three points, $(1,0,3),(4,1,0)$ and $(-3,1,1)$.
9. 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?
10. Find the area of the parallelogram determined by the vectors, $(1,2,3)$ and $(3,-2,1)$.
11. Find the area of the parallelogram determined by the vectors, $(1,0,3)$ and $(4,-2,1)$.
12. Find the area of the parallelogram determined by the vectors, $(1,-2,2)$ and $(3,1,1)$.
13. Find the volume of the parallelepiped determined by the vectors, $\mathbf{i}-7\mathbf{j}-5\mathbf{k},\mathbf{i}-2\mathbf{j}-6\mathbf{k}$,$3\mathbf{i}+2\mathbf{j}+3\mathbf{k}$.
14. Find the volume of the parallelepiped determined by the vectors, $\mathbf{i}+\mathbf{j}-5\mathbf{k},\mathbf{i}+5\mathbf{j}-6\mathbf{k}$,$3\mathbf{i}+\mathbf{j}+3\mathbf{k}$.
15. Find the volume of the parallelepiped determined by the vectors, $\mathbf{i}+6\mathbf{j}+5\mathbf{k},\mathbf{i}+5\mathbf{j}-6\mathbf{k}$,$3\mathbf{i}+\mathbf{j}+\mathbf{k}$.
16. 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?

17. What does it mean geometrically if the box product of three vectors gives zero?

18. It is desired to find an equation of a plane containing the two vectors, $\mathbf{a}$ and $\mathbf{b}$ and the point $\mathbf{0}$. Using Problem 17, show an equation for this plane is

$$\begin{vmatrix} x & y & z \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = 0$$

That is, the set of all (x, y, z) such that the above expression equals zero.

19. 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.

20. 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 $(\mathbf{i} \times \mathbf{j}) \times \mathbf{j}$.

21. 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

$$\begin{aligned} |\mathbf{a} \times \mathbf{b}|^2 &= |\mathbf{a}|^2 |\mathbf{b}|^2 - (\mathbf{a} \cdot \mathbf{b})^2 \\ &= |\mathbf{a}|^2 |\mathbf{b}|^2 \left(1 - \cos^2(\theta)\right) \end{aligned}$$

where θ 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.

22. Discover a vector identity for $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$.

23. Discover a vector identity for $(\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{z} \times \mathbf{w})$.

24. Discover a vector identity for $(\mathbf{u} \times \mathbf{v}) \times (\mathbf{z} \times \mathbf{w})$ in terms of box products.

25. Simplify $(\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{v} \times \mathbf{w}) \times (\mathbf{w} \times \mathbf{z})$.

26. Simplify $|\mathbf{u} \times \mathbf{v}|^2 + (\mathbf{u} \times \mathbf{v})^2 - |\mathbf{u}|^2 |\mathbf{v}|^2$.

27. Prove that $\varepsilon_{ijk}\varepsilon_{ijr} = 2\delta_{kr}$.

28. If A is a 3×3 matrix such that $A = \begin{pmatrix} \mathbf{u} & \mathbf{v} & \mathbf{w} \end{pmatrix}$ where these are the columns of the matrix, A. Show that $\det(A) = \varepsilon_{ijk} u_i v_j w_k$.

29. If A is a 3×3 matrix, show $\varepsilon_{rps} \det(A) = \varepsilon_{ijk} A_{ri} A_{pj} A_{sk}$.

30. Suppose A is a 3×3 matrix and $\det(A) \neq 0$. Show using 29 and 27 that

$$\left(A^{-1}\right)_{ks} = \frac{1}{2\det(A)} \varepsilon_{rps} \varepsilon_{ijk} A_{pj} A_{ri}.$$

31. When you have a rotating rigid body with angular velocity vector, $\boldsymbol{\Omega}$ then the velocity, $\mathbf{u}'$ is given by $\mathbf{u}' = \boldsymbol{\Omega} \times \mathbf{u}$. It turns out that all the usual calculus rules such as the product rule hold. Also, $\mathbf{u}''$ is the acceleration. Show using the product rule that for $\boldsymbol{\Omega}$ a constant vector,

$$\mathbf{u}'' = \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{u}).$$

It turns out this is the centripetal acceleration. Note how it involves cross products. Things get really interesting when you move about on the rotating body. Weird forces are felt. This is in the section on moving coordinate systems.

Bases For $\mathbb{R}^n$

19.0.1 Outcomes

1. Recall and use the definition of an orthonormal basis.
2. Show that any orthonormal set of vectors is linearly independent.
3. Find an orthonormal basis using the Gram Schmidt process.
4. Define the dual basis for a given basis.
5. Find and define the metric tensor.
6. Find the dual basis for a given basis.
7. Explain how to use the metric tensor to write the dot product in terms of components with respect to a given basis which might not be orthonormal.

19.1 Orthonormal Bases

Not all bases for $\mathbb{F}^n$ are created equal. Recall $\mathbb{F}$ equals either $\mathbb{C}$ or $\mathbb{R}$ and the dot product is given by

$$\mathbf{x} \cdot \mathbf{y} = \sum_j x_j \overline{y_j}$$

The best ones are orthonormal. Much of what follows will be for $\mathbb{F}^n$ in the interest of generality but you can substitute $\mathbb{R}$ for $\mathbb{F}$ if you like. Then later if you need it you can read it in full generality.

Definition 19.1.1 *Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is a set of vectors in $\mathbb{F}^n$. It is an orthonormal set if $\mathbf{v}_i \cdot \mathbf{v}_j = \delta_{ij}$.*

Every orthonormal set of vectors is automatically linearly independent.

Proposition 19.1.2 *Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is an orthonormal set of vectors. Then it is linearly independent.*

Proof: Suppose $c_i \mathbf{v}_i = \mathbf{0}$ where summation takes place here and below over the repeated index. Then taking dot products with $\mathbf{v}_j, 0 = \mathbf{0} \cdot \mathbf{v}_j = c_i \mathbf{v}_i \cdot \mathbf{v}_j = c_i \delta_{ij} = c_j$. Since j is arbitrary, this shows the set is linearly independent as claimed.

It turns out that if X is any subspace of $\mathbb{F}^m$, then there exists an orthonormal basis for X.

Lemma 19.1.3 *Let X be a subspace of $\mathbb{F}^m$ of dimension n whose basis is $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$. Then there exists an orthonormal basis for X, $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ which has the property that for each $k \leq n$, span$(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$.*

Proof: Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a basis for X. Let $\mathbf{u}_1 \equiv \mathbf{x}_1 / |\mathbf{x}_1|$. Thus for $k = 1$, $\text{span}(\mathbf{u}_1) = \text{span}(\mathbf{x}_1)$ and $\{\mathbf{u}_1\}$ is an orthonormal set. Now suppose for some $k < n$, $\mathbf{u}_1, \cdots, \mathbf{u}_k$ have been chosen such that $(\mathbf{u}_j, \mathbf{u}_l) = \delta_{jl}$ and $\text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$. Then define

$$\mathbf{u}_{k+1} \equiv \frac{\mathbf{x}_{k+1} - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)\, \mathbf{u}_j}{\left| \mathbf{x}_{k+1} - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)\, \mathbf{u}_j \right|}, \tag{19.1}$$

where the denominator is not equal to zero because the $\mathbf{x}_j$ form a basis and so

$$\mathbf{x}_{k+1} \notin \text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$$

Thus by induction,

$$\mathbf{u}_{k+1} \in \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{x}_{k+1}) = \text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{x}_{k+1}).$$

Also, $\mathbf{x}_{k+1} \in \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1})$ which is seen easily by solving (19.1) for $\mathbf{x}_{k+1}$ and it follows

$$\text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{x}_{k+1}) = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1}).$$

If $l \leq k$,

$$\begin{aligned}
(\mathbf{u}_{k+1} \cdot \mathbf{u}_l) &= C\left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)\, (\mathbf{u}_j \cdot \mathbf{u}_l) \right) \\
&= C\left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)\, \delta_{lj} \right) \\
&= C\left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - (\mathbf{x}_{k+1} \cdot \mathbf{u}_l) \right) = 0.
\end{aligned}$$

The vectors, $\{\mathbf{u}_j\}_{j=1}^{n}$, generated in this way are therefore an orthonormal basis because each vector has unit length.

The process by which these vectors were generated is called the Gram Schmidt process. The Gram Schmidt process of the above lemma has major significance.

Lemma 19.1.4 *Let A be an $m \times n$ matrix and let $A(\mathbb{F}^n)$ denote the set of vectors in $\mathbb{F}^m$ which are of the form $A\mathbf{x}$ for some $\mathbf{x} \in \mathbb{F}^n$. Then $A(\mathbb{F}^n)$ is a subspace of $\mathbb{F}^m$.*

Proof: Let $A\mathbf{x}$ and $A\mathbf{y}$ be two elements of $A(\mathbb{F}^n)$. It suffices to verify that if a, b are scalars, then $aA\mathbf{x} + bA\mathbf{y}$ is also in $A(\mathbb{F}^n)$. But $aA\mathbf{x} + bA\mathbf{y} = A(a\mathbf{x} + b\mathbf{y})$ because A is linear. This proves the lemma.

Theorem 19.1.5 *Let $\mathbf{y} \in \mathbb{F}^m$ and let A be an $m \times n$ matrix. Then there exists $\mathbf{x} \in \mathbb{F}^n$ minimizing the function, $|\mathbf{y} - A\mathbf{x}|^2$. Furthermore, $\mathbf{x}$ minimizes this function if and only if*

$$(\mathbf{y} - A\mathbf{x}) \cdot A\mathbf{w} = 0$$

for all $\mathbf{w} \in \mathbb{F}^n$.

Proof: Let $\{\mathbf{f}_1, \cdots, \mathbf{f}_r\}$ be an orthonormal basis for $A(\mathbb{F}^n)$. Since $A(\mathbb{F}^n) = \text{span}(\mathbf{f}_1, \cdots, \mathbf{f}_r)$, it follows that if you can find $y_1, \cdots, y_r$ in such a way as to minimize

$$\left| \mathbf{y} - \sum_{k=1}^{r} y_k \mathbf{f}_k \right|^2,$$

then letting $A\mathbf{x} = \sum_{k=1}^{r} y_k \mathbf{f}_k$, it will follow that this $\mathbf{x}$ is the desired solution. Let $y_1, \cdots, y_r$ be a list of scalars. Then from the definition of $|\cdot|$ and the properties of the dot product,

$$\begin{aligned}
\left| \mathbf{y} - \sum_{k=1}^{r} y_k \mathbf{f}_k \right|^2 &= \left(\mathbf{y} - \sum_{k=1}^{r} y_k \mathbf{f}_k \right) \cdot \left(\mathbf{y} - \sum_{k=1}^{r} y_k \mathbf{f}_k \right) \\
&= |\mathbf{y}|^2 - 2 \operatorname{Re} \sum_{k=1}^{r} y_k (\mathbf{y} \cdot \mathbf{f}_k) + \sum_k \sum_l y_k y_l \overbrace{\mathbf{f}_k \cdot \mathbf{f}_l}^{\delta_{kl}} \\
&= |\mathbf{y}|^2 - 2 \operatorname{Re} \sum_{k=1}^{r} y_k (\mathbf{y} \cdot \mathbf{f}_k) + \sum_{k=1}^{r} |y_k|^2 \\
&= |\mathbf{y}|^2 + \sum_{k=1}^{r} |y_k|^2 - 2 \operatorname{Re} y_k (\mathbf{y} \cdot \mathbf{f}_k)
\end{aligned}$$

Now complete the square to obtain

$$\begin{aligned}
&= |\mathbf{y}|^2 + \sum_{k=1}^{r} \left(|y_k|^2 - 2 \operatorname{Re} y_k (\mathbf{y} \cdot \mathbf{f}_k) + |\mathbf{y} \cdot \mathbf{f}_k|^2 \right) - \sum_{k=1}^{r} |\mathbf{y} \cdot \mathbf{f}_k|^2 \\
&= |\mathbf{y}|^2 + \sum_{k=1}^{r} |y_k - (\mathbf{y} \cdot \mathbf{f}_k)|^2 - \sum_{k=1}^{r} (\mathbf{y} \cdot \mathbf{f}_k)^2 .
\end{aligned}$$

This shows that the minimum is obtained when $y_k = (\mathbf{y} \cdot \mathbf{f}_k)$. This proves the existence part of the Theorem.

To verify the other part, let $t \in \mathbb{R}$ and consider

$$\begin{aligned}
|\mathbf{y} - A(\mathbf{x} + t\mathbf{w})|^2 &= (\mathbf{y} - A\mathbf{x} - tA\mathbf{w}) \cdot (\mathbf{y} - A\mathbf{x} - tA\mathbf{w}) \\
&= |\mathbf{y} - A\mathbf{x}|^2 - 2t \operatorname{Re} (\mathbf{y} - A\mathbf{x}) \cdot A\mathbf{w} + t^2 |A\mathbf{w}|^2 .
\end{aligned}$$

Then from the above equation, $|\mathbf{y} - A\mathbf{x}|^2 \leq |\mathbf{y} - A\mathbf{z}|^2$ for all $\mathbf{z} \in \mathbb{F}^n$ if and only if for all $\mathbf{w} \in \mathbb{F}^n$ and $t \in \mathbb{R}$

$$|\mathbf{y} - A\mathbf{x}|^2 - 2t \operatorname{Re} (\mathbf{y} - A\mathbf{x}) \cdot A\mathbf{w} + t^2 |\mathbf{w}|^2 \geq |\mathbf{y} - A\mathbf{x}|^2$$

and this happens if and only if for all $t \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{F}^n$,

$$-2t \operatorname{Re} (\mathbf{y} - A\mathbf{x}) \cdot A\mathbf{w} + t^2 |A\mathbf{w}|^2 \geq 0,$$

which occurs if and only if $\operatorname{Re} (\mathbf{y} - A\mathbf{x}) \cdot A\mathbf{w} = 0$ for all $\mathbf{w} \in \mathbb{R}^n$. (Why?)

This implies that $(\mathbf{y} - A\mathbf{x}) \cdot A\mathbf{w} = 0$ for every $\mathbf{w} \in \mathbb{F}^n$ because there exists a complex number, θ of magnitude 1 such that

$$\begin{aligned}
|(\mathbf{y} - A\mathbf{x}) \cdot A\mathbf{w}| &= \theta (\mathbf{y} - A\mathbf{x}) \cdot A\mathbf{w} = (\mathbf{y} - A\mathbf{x}) \cdot A\overline{\theta}\mathbf{w} \\
&= \operatorname{Re} (\mathbf{y} - A\mathbf{x}) \cdot A\overline{\theta}\mathbf{w} = 0.
\end{aligned}$$

Definition 19.1.6 *Let A be an $m \times n$ matrix. Then*

$$A^* \equiv \overline{(A^T)}.$$

This means you take the transpose of A and then replace each entry by its conjugate. This matrix is called the ***adjoint****. Thus in the case of real matrices having only real entries, the adjoint is just the transpose.*

Lemma 19.1.7 *Let A be an $m \times n$ matrix. Then*

$$A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A^* \mathbf{y}$$

Proof: This follows from the definition. Using the repeated index summation convention,

$$\begin{aligned} A\mathbf{x} \cdot \mathbf{y} &= A_{ij} x_j \overline{y_i} \\ &= x_j \overline{A^*_{ji}} \overline{y_i} \\ &= \mathbf{x} \cdot A^* \mathbf{y}. \end{aligned}$$

This proves the lemma.

Corollary 19.1.8 *A value of $\mathbf{x}$ which solves the problem of Theorem 19.1.5 is obtained by solving the equation*

$$A^* A\mathbf{x} = A^* \mathbf{y}$$

and furthermore, there exists a solution to this system of equations.

Proof: For $\mathbf{x}$ the unique minimizer of Theorem 19.1.5, $(\mathbf{y} - A\mathbf{x}) \cdot A\mathbf{w} = 0$ for all $\mathbf{w} \in \mathbb{F}^n$ and from Lemma 19.1.7, this is the same as saying

$$A^* (\mathbf{y} - A\mathbf{x}) \cdot \mathbf{w} = 0$$

for all $\mathbf{w} \in \mathbb{F}^n$. Therefore, there is a unique solution to the equation of this corollary and it solves the minimization problem of Theorem 19.1.5.

19.1.1 The Least Squares Regression Line

For the situation of the least squares regression line discussed here I will specialize to the case of $\mathbb{R}^n$ rather than $\mathbb{F}^n$ because it seems this case is by far the most interesting and the extra details are not justified by an increase in utility. Thus, everywhere you see A^* it suffices to place A^T.

An important application of the above theorem is the problem of finding the least squares regression line in statistics. Suppose you are given points in the plane, $\{(x_i, y_i)\}_{i=1}^n$ and you would like to find constants m and b such that the line $y = mx + b$ goes through all these points. Of course this will be impossible in general. Therefore, try to find m, b to get as close as possible. The desired system is

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix}$$

which is of the form $\mathbf{y} = A\mathbf{x}$ and it is desired to choose m and b to make

$$\left| A \begin{pmatrix} m \\ b \end{pmatrix} - \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right|^2$$

as small as possible. According to Theorem 19.1.5 and Corollary 19.1.8, the best values for m and b occur as the solution to

$$A^T A \begin{pmatrix} m \\ b \end{pmatrix} = A^T \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

where

$$A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}.$$

Thus, computing $A^T A$,

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix}$$

Solving this system of equations for m and b,

$$m = \frac{-\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) + \left(\sum_{i=1}^n x_i y_i\right) n}{\left(\sum_{i=1}^n x_i^2\right) n - \left(\sum_{i=1}^n x_i\right)^2}$$

and

$$b = \frac{-\left(\sum_{i=1}^n x_i\right)\sum_{i=1}^n x_i y_i + \left(\sum_{i=1}^n y_i\right)\sum_{i=1}^n x_i^2}{\left(\sum_{i=1}^n x_i^2\right) n - \left(\sum_{i=1}^n x_i\right)^2}.$$

One could clearly do a least squares fit for curves of the form $y = ax^2 + bx + c$ in the same way. In this case you want to solve as well as possible for a, b, and c the system

$$\begin{pmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

and one would use the same technique as above. Many other similar problems are important, including many in higher dimensions and they are all solved the same way.

19.1.2 The Fredholm Alternative

The next major result is called the Fredholm alternative. It comes from Theorem 19.1.5 and Lemma 19.1.7.

Theorem 19.1.9 *Let A be an $m \times n$ matrix. Then there exists $\mathbf{x} \in \mathbb{F}^n$ such that $A\mathbf{x} = \mathbf{y}$ if and only if whenever $A^*\mathbf{z} = \mathbf{0}$ it follows that $\mathbf{z} \cdot \mathbf{y} = 0$.*

Proof: First suppose that for some $\mathbf{x} \in \mathbb{F}^n$, $A\mathbf{x} = \mathbf{y}$. Then letting $A^*\mathbf{z} = \mathbf{0}$ and using Lemma 19.1.7

$$\mathbf{y} \cdot \mathbf{z} = A\mathbf{x} \cdot \mathbf{z} = \mathbf{x} \cdot A^*\mathbf{z} = \mathbf{x} \cdot \mathbf{0} = 0.$$

This proves half the theorem.

To do the other half, suppose that whenever, $A^*\mathbf{z} = \mathbf{0}$ it follows that $\mathbf{z} \cdot \mathbf{y} = 0$. It is necessary to show there exists $\mathbf{x} \in \mathbb{F}^n$ such that $\mathbf{y} = A\mathbf{x}$. From Theorem 19.1.5 there exists $\mathbf{x}$ minimizing $|\mathbf{y} - A\mathbf{x}|^2$ which therefore satisfies

$$(\mathbf{y} - A\mathbf{x}) \cdot A\mathbf{w} = 0 \tag{19.2}$$

for all $\mathbf{w} \in \mathbb{F}^n$. Therefore, for all $\mathbf{w} \in \mathbb{F}^n$,

$$A^* (\mathbf{y} - A\mathbf{x}) \cdot \mathbf{w} = 0$$

which shows that $A^* (\mathbf{y} - A\mathbf{x}) = \mathbf{0}$. (Why?) Therefore, by assumption,

$$(\mathbf{y} - A\mathbf{x}) \cdot \mathbf{y} = 0.$$

Now by (19.2) with $\mathbf{w} = \mathbf{x}$,

$$(\mathbf{y} - A\mathbf{x}) \cdot (\mathbf{y} - A\mathbf{x}) = (\mathbf{y} - A\mathbf{x}) \cdot \mathbf{y} - (\mathbf{y} - A\mathbf{x}) \cdot A\mathbf{x} = 0$$

showing that $\mathbf{y} = A\mathbf{x}$. This proves the theorem.

The following corollary is also called the Fredholm alternative.

Corollary 19.1.10 *Let A be an $m \times n$ matrix. Then A is onto if and only if A^* is one to one.*

Proof: Suppose first A is onto. Then by Theorem 19.1.9, it follows that for all $\mathbf{y} \in \mathbb{F}^m$, $\mathbf{y} \cdot \mathbf{z} = 0$ whenever $A^*\mathbf{z} = \mathbf{0}$. Therefore, let $\mathbf{y} = \mathbf{z}$ where $A^*\mathbf{z} = \mathbf{0}$ and conclude that $\mathbf{z} \cdot \mathbf{z} = 0$ whenever $A^*\mathbf{z} = 0$. If $A^*\mathbf{x} = A^*\mathbf{y}$, then $A^* (\mathbf{x} - \mathbf{y}) = \mathbf{0}$ and so $\mathbf{x} - \mathbf{y} = \mathbf{0}$. Thus A^* is one to one.

Now let $\mathbf{y} \in \mathbb{F}^m$ be given. $\mathbf{y} \cdot \mathbf{z} = 0$ whenever $A^*\mathbf{z} = \mathbf{0}$ because, since A^* is assumed to be one to one, and $\mathbf{0}$ is a solution to this equation, it must be the only solution. Therefore, by Theorem 19.1.9 there exists $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{y}$ therefore, A is onto.

19.2 The Dual Basis

That which follows on the dual basis can be extended to $\mathbb{C}^n$ but this will not be done here.

Given a basis, there is something called a dual basis which is very important in applications. A given basis might not be orthonormal so you can't say $\mathbf{v}_i \cdot \mathbf{v}_j = \delta_{ij}$ but you really want to say this. A useful way of getting many of the same advantages is to define something called a dual basis. A dual basis for $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a set of vectors, $\{\mathbf{v}^1, \cdots, \mathbf{v}^n\}$ which has the property that $\mathbf{v}^k \cdot \mathbf{v}_j = \delta^k_j$. Here

$$\delta^j_i \equiv \begin{cases} 1 \text{ if } j = i \\ 0 \text{ if } j \neq i \end{cases}.$$

Similarly,

$$\delta^{ji} \equiv \begin{cases} 1 \text{ if } j = i \\ 0 \text{ if } j \neq i \end{cases}.$$

In this subject it is convenient as well as traditional to keep track of the level on which the index occurs. Thus $\mathbf{v}_i \neq \mathbf{v}^i$! Of course, sometimes these are the same but not generally.

Definition 19.2.1 *Let $\{\mathbf{e}_i\}_{i=1}^n$ form a basis for $\mathbb{R}^n$. Then $\{\mathbf{e}^i\}_{i=1}^n$ is called the dual basis if*

$$\mathbf{e}^i \cdot \mathbf{e}_j = \delta^i_j \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}. \tag{19.3}$$

Theorem 19.2.2 *If $\{\mathbf{e}_i\}_{i=1}^n$ is a basis then $\{\mathbf{e}^i\}_{i=1}^n$ is also a basis provided (19.3) holds. Furthermore, for each vector, $\mathbf{v}$,*

$$\mathbf{v} = (\mathbf{v} \cdot \mathbf{e}_j)\, \mathbf{e}^j.$$

Also,

$$\mathbf{v} = \left(\mathbf{v} \cdot \mathbf{e}^j\right) \mathbf{e}_j$$

Proof: First we verify that $\{\mathbf{e}^i\}_{i=1}^n$ is linearly independent. Suppose

$$\mathbf{0} = v_i \mathbf{e}^i. \tag{19.4}$$

Then taking the dot product of both sides of (19.4) with $\mathbf{e}_j$,yields

$$0 = v_i \mathbf{e}^i \cdot \mathbf{e}_j = v_i \delta^i_j = v_j.$$

Since j was arbitrary, this shows each $v_j = 0$ and so the set is linearly independent as claimed.

It remains to verify $\{\mathbf{e}^i\}_{i=1}^n$ spans $\mathbb{R}^n$. Let $\mathbf{v} \in \mathbb{R}^n$ be arbitrary and consider the element in the span of these vectors, $(\mathbf{v} \cdot \mathbf{e}_j)\, \mathbf{e}^j$. Then

$$\left(\mathbf{v} - (\mathbf{v} \cdot \mathbf{e}_j)\, \mathbf{e}^j\right) \cdot \mathbf{e}_k = \mathbf{v} \cdot \mathbf{e}_k - (\mathbf{v} \cdot \mathbf{e}_j)\, \mathbf{e}^j \cdot \mathbf{e}_k = \mathbf{v} \cdot \mathbf{e}_k - (\mathbf{v} \cdot \mathbf{e}_j)\, \delta^j_k = 0$$

and so, since $\{\mathbf{e}_i\}_{i=1}^n$ is a basis,

$$\left(\mathbf{v} - \mathbf{v} \cdot \mathbf{e}_j \mathbf{e}^j\right) \cdot \mathbf{w} = 0$$

for all vectors, $\mathbf{w}$. In particular, this would hold for $\mathbf{w} = \left(\mathbf{v} - \mathbf{v} \cdot \mathbf{e}_j \mathbf{e}^j\right)$. It follows $\mathbf{v} - \mathbf{v} \cdot \mathbf{e}_j \mathbf{e}^j = \mathbf{0}$ and this shows $\{\mathbf{e}^i\}_{i=1}^n$ is a basis.

In the above argument we obtained formulas for the components of a vector $\mathbf{v}$, v_i, with respect to the dual basis, found to be $v_j = \mathbf{v} \cdot \mathbf{e}_j$. In the same way, we may find the components of a vector with respect to the basis $\{\mathbf{e}_i\}_{i=1}^n$. Let $\mathbf{v}$ be any vector and let

$$\mathbf{v} = v^j \mathbf{e}_j. \tag{19.5}$$

Then taking the dot product of both sides of (19.5) with $\mathbf{e}^i$, $v^i = \mathbf{e}^i \cdot \mathbf{v}$.

Does there exist a dual basis and is it uniquely determined?

Lemma 19.2.3 *Let $\{\mathbf{e}_i\}_{i=1}^n$ be a basis for $\mathbb{R}^n$. The matrix, $G \equiv (g_{ij}) = (\mathbf{e}_i \cdot \mathbf{e}_j)$ is an invertible matrix. Furthermore* $\det G > 0$.

Proof: Each of these vectors is in $\mathbb{R}^n$ and so can be written as a column matrix of numbers with respect to the usual basis for $\mathbb{R}^n$. Now note $\mathbf{e}_i \cdot \mathbf{e}_j = \mathbf{e}_i^T \mathbf{e}_j$. Therefore, the above matrix is nothing more than

$$G = (g_{ij}) = \begin{pmatrix} \mathbf{e}_1^T \\ \vdots \\ \mathbf{e}_n^T \end{pmatrix} \begin{pmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_n \end{pmatrix} = U^T U$$

where U is the matrix which has the $\mathbf{e}_i$ as columns. Therefore,

$$\det(G) = \det\left(U^T\right) \det(U) = \det(U)^2 \geq 0.$$

Since $\{\mathbf{e}_i\}_{i=1}^n$ is a basis, it follows the matrix, U above is one to one. Therefore, the matrix, U has an inverse and so $\det(U) \neq 0$. It follows $\det(G) > 0$.

Definition 19.2.4 *The matrix, G above is called the metric tensor. Its inverse, G^{-1} is denoted by $\left(g^{ij}\right)$. That is the ij^{th} entry of G^{-1} is denoted as g^{ij}. Thus from the definition of matrix multiplication,*

$$g^{ik} g_{kj} = \delta^i_j.$$

Theorem 19.2.5 *If $\{\mathbf{e}_i\}_{i=1}^n$ is a basis for $\mathbb{R}^n$, then there exists a unique dual basis, $\{\mathbf{e}^j\}_{j=1}^n$ satisfying*

$$\mathbf{e}^j \cdot \mathbf{e}_i = \delta_i^j.$$

Furthermore, $\mathbf{e}^i = g^{ij}\mathbf{e}_j$.

Proof:

$$g^{ij}\mathbf{e}_j \cdot \mathbf{e}_k = g^{ij}g_{jk} = \delta_k^i.$$

This proves the existence of the dual basis. Uniqueness was established earlier.

If $\mathbf{v}$ is any vector, there exist unique scalars, v_k such that $v_k\mathbf{e}^k = \mathbf{v}$. Also, scalars v^k such that $v^k\mathbf{e}_k = \mathbf{v}$. We saw these scalars are given by $v_k = \mathbf{v} \cdot \mathbf{e}_k$ and $v^k = \mathbf{v} \cdot \mathbf{e}^k$. Do you begin to get the idea on the notation? You sum over indices on different levels.

Definition 19.2.6 *If $\mathbf{v}$ is any vector,*

$$\mathbf{v} = v_j\mathbf{e}^j,\ \mathbf{v} = v^j\mathbf{e}_j. \tag{19.6}$$

The components of $\mathbf{v}$ which have the index on the top are called the contravariant components of the vector while the components which have the index on the bottom are called the covariant components. In general $v_i \neq v^j$!

Theorem 19.2.7 *The following hold.*

$$g^{ij}\mathbf{e}_j = \mathbf{e}^i,\ g_{ij}\mathbf{e}^j = \mathbf{e}_i, \tag{19.7}$$

$$g^{ij}v_j = v^i,\ g_{ij}v^j = v_i, \tag{19.8}$$

$$\det(g_{ij}) > 0,\ \det\left(g^{ij}\right) > 0. \tag{19.9}$$

$$g^{ij} = \mathbf{e}^i \cdot \mathbf{e}^j,\ g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j \tag{19.10}$$

Proof: It was shown that $g^{ij}\mathbf{e}_j = \mathbf{e}^i$ in Theorem 19.2.5. The second claim of (19.7) follows from Theorem 19.2.2.

$$\mathbf{e}_i = (\mathbf{e}_i \cdot \mathbf{e}_j)\,\mathbf{e}^j = g_{ij}\mathbf{e}^j.$$

This verifies (19.7). To verify (19.8), use Theorem 19.2.2.

$$v^i = \mathbf{e}^i \cdot \mathbf{v} = g^{ij}\mathbf{e}_j \cdot \mathbf{v} = g^{ij}v_j.$$

To establish the other formula in (19.8), use Theorem 19.2.2 again.

$$v_i = \mathbf{e}_i \cdot \mathbf{v} = g_{ij}\mathbf{e}^j \cdot \mathbf{v} = g_{ij}v^j.$$

It was shown that $\det(g_{ij}) > 0$ already. But $\det\left(g^{ij}\right)\det(g_{ij}) = 1$ so this proves the second claim in (19.9). The second of the claims in (19.10) is the way g_{ij} was defined. It only remains to verify the first equation. Using Theorem 19.2.2,

$$\begin{aligned}\left(\mathbf{e}^i \cdot \mathbf{e}^k\right)(\mathbf{e}_k \cdot \mathbf{e}_j) &= \left(\left(\mathbf{e}^i \cdot \mathbf{e}^k\right)\mathbf{e}_k \cdot \mathbf{e}_j\right) \\ &= \mathbf{e}^i \cdot \mathbf{e}_j = \delta_j^i.\end{aligned}$$

Since $\left(\mathbf{e}^i \cdot \mathbf{e}^k\right)$ acts like the inverse of (g_{ij}) it follows it is the inverse. This proves the theorem.

The process of writing $g^{ij}v_j = v^i$ is sometimes called raising the index while the process $g_{ij}v^j = v_i$ is called lowering the index.

Example 19.2.8 *Let* $\mathbf{e}_1 = (1,2,1)^T, \mathbf{e}_2 = (0,1,1)^T$, *and* $\mathbf{e}_3 = (3,-1,1)^T$. *Find the dual basis.*

As explained above, the metric tensor is

$$G = \begin{pmatrix} 6 & 3 & 2 \\ 3 & 2 & 0 \\ 2 & 0 & 11 \end{pmatrix}$$

Taking the inverse,

$$G^{-1} = \begin{pmatrix} \frac{22}{25} & -\frac{33}{25} & -\frac{4}{25} \\ -\frac{33}{25} & \frac{62}{25} & \frac{6}{25} \\ -\frac{4}{25} & \frac{6}{25} & \frac{3}{25} \end{pmatrix}$$

Therefore,

$$\mathbf{e}^1 = g^{1j}\mathbf{e}_j = \frac{22}{25}\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + \left(-\frac{33}{25}\right)\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + \left(-\frac{4}{25}\right)\begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{2}{5} \\ \frac{3}{5} \\ -\frac{3}{5} \end{pmatrix}$$

$$\mathbf{e}^2 = g^{2j}\mathbf{e}_j = \left(-\frac{33}{25}\right)\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + \left(\frac{62}{25}\right)\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + \left(\frac{6}{25}\right)\begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -\frac{3}{5} \\ -\frac{2}{5} \\ \frac{7}{5} \end{pmatrix}$$

$$\mathbf{e}^3 = g^{3j}\mathbf{e}_j = \left(-\frac{4}{25}\right)\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + \left(\frac{6}{25}\right)\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + \left(\frac{3}{25}\right)\begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{5} \\ -\frac{1}{5} \\ \frac{1}{5} \end{pmatrix}$$

Another way to find the dual basis is as follows. First make the matrix, M which has as columns the given basis. Then multiply on the right by G^{-1}. The resulting matrix will have as columns the dual basis. For example, this procedure yields$\mathbf{e}_1 = (1,2,1)^T, \mathbf{e}_2 = (0,1,1)^T$, and $\mathbf{e}_3 = (3,-1,1)^T$.

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} \frac{22}{25} & -\frac{33}{25} & -\frac{4}{25} \\ -\frac{33}{25} & \frac{62}{25} & \frac{6}{25} \\ -\frac{4}{25} & \frac{6}{25} & \frac{3}{25} \end{pmatrix} = \begin{pmatrix} \frac{2}{5} & -\frac{3}{5} & \frac{1}{5} \\ \frac{3}{5} & -\frac{2}{5} & -\frac{1}{5} \\ -\frac{3}{5} & \frac{7}{5} & \frac{1}{5} \end{pmatrix}.$$

This follows from the symmetry of G^{-1} and the definition of matrix multiplication.

Example 19.2.9 *Let* $\mathbf{e}_1 = (1,2,0,1)^T, \mathbf{e}_2 = (2,1,0,0)^T, \mathbf{e}_3 = (0,1,1,2)^T$, *and* $\mathbf{e}_4 = (0,0,3,1)^T$. *Find the dual basis and metric tensor.*

First find the metric tensor. As explained above, this can be done by multiplying the two matrices,

$$\begin{pmatrix} 1 & 2 & 0 & 1 \\ 2 & 1 & 0 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 3 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 1 & 1 & 0 \\ 0 & 0 & 1 & 3 \\ 1 & 0 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 4 & 4 & 1 \\ 4 & 5 & 1 & 0 \\ 4 & 1 & 6 & 5 \\ 1 & 0 & 5 & 10 \end{pmatrix}.$$

Now you invert this matrix to get the inverse of the metric tensor.

$$\begin{pmatrix} \frac{55}{27} & -\frac{35}{27} & -\frac{5}{3} & \frac{17}{27} \\ -\frac{35}{27} & \frac{28}{27} & 1 & -\frac{10}{27} \\ -\frac{5}{3} & 1 & \frac{5}{3} & -\frac{2}{3} \\ \frac{17}{27} & -\frac{10}{27} & -\frac{2}{3} & \frac{10}{27} \end{pmatrix}.$$

Then the dual basis consists of the columns of the matrix,

$$\begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 1 & 1 & 0 \\ 0 & 0 & 1 & 3 \\ 1 & 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} \frac{55}{27} & -\frac{35}{27} & -\frac{5}{3} & \frac{17}{27} \\ -\frac{35}{27} & \frac{28}{27} & 1 & -\frac{10}{27} \\ -\frac{5}{3} & 1 & \frac{5}{3} & -\frac{2}{3} \\ \frac{17}{27} & -\frac{10}{27} & -\frac{2}{3} & \frac{10}{27} \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{5}{9} & \frac{7}{9} & \frac{1}{3} & -\frac{1}{9} \\ \frac{10}{9} & -\frac{5}{9} & -\frac{2}{3} & \frac{2}{9} \\ \frac{2}{9} & -\frac{1}{9} & -\frac{1}{3} & \frac{4}{9} \\ -\frac{2}{3} & \frac{1}{3} & 1 & -\frac{1}{3} \end{pmatrix}. \tag{19.11}$$

Lets check this by multiplying by the matrix which has rows equal to the basis.

$$\begin{pmatrix} 1 & 2 & 0 & 1 \\ 2 & 1 & 0 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 3 & 1 \end{pmatrix} \begin{pmatrix} -\frac{5}{9} & \frac{7}{9} & \frac{1}{3} & -\frac{1}{9} \\ \frac{10}{9} & -\frac{5}{9} & -\frac{2}{3} & \frac{2}{9} \\ \frac{2}{9} & -\frac{1}{9} & -\frac{1}{3} & \frac{4}{9} \\ -\frac{2}{3} & \frac{1}{3} & 1 & -\frac{1}{3} \end{pmatrix}$$

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

Therefore, the columns of (19.11) are the dual basis as hoped.

Example 19.2.10 *In the above example, find the covariant and contravariant components of the vector,* $(2,1,1) = 2\mathbf{i} + \mathbf{j} + \mathbf{k}$.

$$v_1 = \mathbf{e}_1 \cdot \mathbf{v} = (1,2,1) \cdot (2,1,1) = 5$$
$$v_2 = \mathbf{e}_2 \cdot \mathbf{v} = (0,1,1) \cdot (2,1,1) = 2$$
$$v_3 = \mathbf{e}_3 \cdot \mathbf{v} = (3,-1,1) \cdot (2,1,1) = 6.$$

These are the covariant components. To check whether these work, form

$$5\overbrace{\begin{pmatrix} \frac{2}{5} \\ \frac{3}{5} \\ -\frac{3}{5} \end{pmatrix}}^{\mathbf{e}^1} + 2\overbrace{\begin{pmatrix} -\frac{3}{5} \\ \frac{3}{5} \\ \frac{7}{5} \end{pmatrix}}^{\mathbf{e}^2} + 6\overbrace{\begin{pmatrix} \frac{1}{5} \\ -\frac{1}{5} \\ \frac{1}{5} \end{pmatrix}}^{\mathbf{e}^3} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}.$$

Success! Now find the contravariant components. To do this, you can simply raise the index.

$$\begin{pmatrix} v^1 \\ v^2 \\ v^3 \end{pmatrix} = \begin{pmatrix} \frac{22}{25} & -\frac{33}{25} & -\frac{4}{25} \\ -\frac{33}{25} & \frac{62}{25} & \frac{6}{25} \\ -\frac{4}{25} & \frac{6}{25} & \frac{3}{25} \end{pmatrix} \begin{pmatrix} 5 \\ 2 \\ 6 \end{pmatrix} = \begin{pmatrix} \frac{4}{5} \\ -\frac{1}{5} \\ \frac{2}{5} \end{pmatrix}.$$

Did it work?

$$\frac{4}{5}\overbrace{\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}}^{\mathbf{e}_1} + \left(-\frac{1}{5}\right)\overbrace{\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}}^{\mathbf{e}_2} + \frac{2}{5}\overbrace{\begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}}^{\mathbf{e}_3} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}.$$

Again, success has occured.

The reason G is called the metric tensor is contained in the following proposition.

Proposition 19.2.11 *Let* $\mathbf{v}, \mathbf{w}$ *be two vectors in* $\mathbb{R}^n$ *and let* $\{\mathbf{e}_i\}_{i=1}^n$ *be a basis for* $\mathbb{R}^n$. *Then*

$$\mathbf{v} \cdot \mathbf{w} = g^{ij} v_i w_j = g_{ij} v^i w^j.$$

Proof:

$$\mathbf{v} \cdot \mathbf{w} = v_i \mathbf{e}^i \cdot w_k \mathbf{e}^k = \mathbf{e}^i \cdot \mathbf{e}^k v_i w_k = g^{ik} v_i w_k.$$

This proves the first equation. The second is similar.

If $\{\mathbf{e}_i\}_{i=1}^n$ is the usual orthonormal basis for $\mathbb{R}^n$, then the metric tensor is just the identity matrix and so you get the usual version of the dot product. The metric tensor allows you to consider the dot product in terms of components of arbitrary bases.

19.3 Exercises

1. The proof of Theorem 19.1.5 concluded with the following observation. If $-ta+t^2 b \geq 0$ for all $t \in \mathbb{R}$ and $b \geq 0$, then $a = 0$. Why is this so?

2. In the proof of Theorem 19.1.9 the following argument was used. If $\mathbf{x} \cdot \mathbf{w} = 0$ for all $\mathbf{w} \in \mathbb{R}^n$, then $\mathbf{x} = \mathbf{0}$. Why is this so?

3. Suppose $L : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation. Show the following are equivalent.

   (a) $L\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$.

   (b) L is one to one.

4. Using Corollary 19.1.10 and Problem 3, show that an $m \times n$ matrix is onto if and only if its transpose is one to one.

5. Suppose A is a 3×2 matrix. Is it possible that A^T is one to one? What does this say about A being onto? Prove your answer.

6. Explain why there always exists a solution to the equation, $A^T \mathbf{y} = A^T A \mathbf{x}$ and also explain why this is called a least squares solution to the equation, $\mathbf{y} = A\mathbf{x}$.

7. Referring to Problem 6, find the least squares solution to the following system.

$$\begin{aligned} x + 2y &= 1 \\ 2x + 3y &= 2 \\ 3x + 5y &= 4 \end{aligned}$$

8. You are doing experiments and have obtained the ordered pairs, $(0,1), (1,2), (2,3.5)$, and $(3,4)$. Find m and b such that $y = mx + b$ approximates these four points as well as possible. Now do the same thing for $y = ax^2 + bx + c$, finding a, b, and c to give the best approximation.

9. Suppose you have several ordered triples, (x_i, y_i, z_i). Describe how to find a polynomial,

$$z = a + bx + cy + dxy + ex^2 + fy^2$$

   for example giving the best fit to the given ordered triples. Is there any reason you have to use a polynomial? Would similar approaches work for other combinations of functions just as well?

10. Using the Gram Schmidt process, find an orthonormal basis for the span of the vectors, $(1,2,1), (2,-1,3)$, and $(1,0,0)$.

11. Using the Gram Schmidt process, find an orthonormal basis for the span of the vectors, $(1,2,1,0), (2,-1,3,1)$, and $(1,0,0,1)$.

12. The set, $V \equiv \{(x,y,z) : 2x+3y-z=0\}$ is a subspace of $\mathbb{R}^3$. Find an orthonormal basis for this subspace.

13. The two level surfaces, $2x+3y-z+w=0$ and $3x-y+z+2w=0$ intersect in a subspace of $\mathbb{R}^4$, find a basis for this subspace. Next find an orthonormal basis for this subspace.

14. Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ be a linearly independent set of vectors. Let $\mathbf{u}_1 = \mathbf{v}_1$ and if $\mathbf{u}_1, \cdots, \mathbf{u}_k$ have been chosen for $k < m$, define

$$\mathbf{u}_{k+1} \equiv \mathbf{v}_{k+1} - \sum_{j=1}^{k} \frac{\mathbf{v}_{k+1} \cdot \mathbf{u}_j}{|\mathbf{u}_j|^2} \mathbf{u}_j.$$

Show that each $\mathbf{u}_k$ is non zero, $\mathbf{u}_k \cdot \mathbf{u}_l = 0$ if $k \neq l$, and for each $k \leq m$

$$\text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_k) = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k).$$

15. Let $\mathbf{e}_1 = \mathbf{i}+\mathbf{j}, \mathbf{e}_2 = \mathbf{i}-\mathbf{j}, \mathbf{e}_3 = \mathbf{j}+\mathbf{k}$. Find $\mathbf{e}^1, \mathbf{e}^2, \mathbf{e}^3$, (g_{ij}), $\left(g^{ij}\right)$. If $\mathbf{v} = \mathbf{i}+2\mathbf{j}+\mathbf{k}$, find v^i and v_j, the contravariant and covariant components of the vector.

16. Let $\mathbf{e}^1 = 2\mathbf{i}+\mathbf{j}, \mathbf{e}^2 = \mathbf{i}-2\mathbf{j}, \mathbf{e}^3 = \mathbf{k}$. Find $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, (g_{ij}), $\left(g^{ij}\right)$. If $\mathbf{v} = 2\mathbf{i}-2\mathbf{j}+\mathbf{k}$, find v^i and v_j, the contravariant and covariant components of the vector.

17. Suppose $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ have the property that $\mathbf{e}_i \cdot \mathbf{e}_j = 0$ whenever $i \neq j$. Show that then the metric tensor is a diagonal matrix.

18. Suppose $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ have the property that $\mathbf{e}_i \cdot \mathbf{e}_j = 0$ whenever $i \neq j$. Show the same is true of the dual basis and that in fact, $\mathbf{e}^i$ is a multiple of $\mathbf{e}_i$.

19. Let $\mathbf{v} = v_i \mathbf{e}^i$ and let $\mathbf{w} = w^j \mathbf{e}_j$. Show that $\mathbf{v} \cdot \mathbf{w} = v_i w^j$.

20. Show if $\{\mathbf{e}_i\}_{i=1}^3$ is a basis in $\mathbb{R}^3$

$$\mathbf{e}^1 = \frac{\mathbf{e}_2 \times \mathbf{e}_3}{\mathbf{e}_2 \times \mathbf{e}_3 \cdot \mathbf{e}_1}, \ \mathbf{e}^2 = \frac{\mathbf{e}_1 \times \mathbf{e}_2}{\mathbf{e}_1 \times \mathbf{e}_3 \cdot \mathbf{e}_2}, \ \mathbf{e}^3 = \frac{\mathbf{e}_1 \times \mathbf{e}_2}{\mathbf{e}_1 \times \mathbf{e}_2 \cdot \mathbf{e}_3}.$$

21. Let $\{\mathbf{e}_i\}_{i=1}^n$ be a basis and define

$$\mathbf{e}_i^* \equiv \frac{\mathbf{e}_i}{|\mathbf{e}_i|}, \ \mathbf{e}^{*i} \equiv \mathbf{e}^i |\mathbf{e}_i|.$$

Show $\mathbf{e}^{*i} \cdot \mathbf{e}_j^* = \delta_j^i$.

22. If $\mathbf{v}$ is a vector, v_i^* and v^{*i}, are defined by

$$\mathbf{v} \equiv v_i^* \mathbf{e}^{*i} \equiv v^{*i} \mathbf{e}_i^*.$$

These are called the physical components of $\mathbf{v}$. Show

$$v_i^* = \frac{v_i}{|\mathbf{e}_i|}, \ v^{*i} = v^i |\mathbf{e}_i| \ \text{ (No summation on } i \text{).}$$

Linear Transformations

20.0.1 Outcomes

1. Define linear transformation. Interpret a matrix as a linear transformation.
2. Find a matrix that represents a linear transformation given by a geometric description.
3. Write the solution space of a homogeneous system as the span of a set of basis vectors. Determine the dimension of the solution space.
4. Relate the solutions of a non-homogeneous system to the solutions of a homogeneous system.

20.1 Linear Transformations

An $m \times n$ matrix can be used to transform vectors in $\mathbb{F}^n$ to vectors in $\mathbb{F}^m$ through the use of matrix multiplication.

Example 20.1.1 *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 in to 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.*

$$\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 5 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 10 \\ 5 \\ -3 \end{pmatrix} = \begin{pmatrix} 20 \\ 25 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 7 \\ 3 \end{pmatrix} = \begin{pmatrix} 14 \\ 7 \end{pmatrix},$$

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 20.1.2 *Let X and Y be vector spaces and let $T : X \to Y$ be a function. Thus for each $\mathbf{x} \in X, T\mathbf{x} \in Y$. Then T is a* ***linear transformation*** *if whenever α, β are scalars and $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors in X,*

$$T(\alpha\mathbf{x}_1 + \beta\mathbf{x}_2) = \alpha_1 T\mathbf{x}_1 + \beta T\mathbf{x}_2.$$

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 (14.14) on Page 375. 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 20.1.3 *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 $\mathbf{x} \neq \mathbf{y}$*

$$T\mathbf{x} \neq T\mathbf{y}.$$

Equivalently, if $T(\mathbf{x}) = T(\mathbf{y})$, then $\mathbf{x} = \mathbf{y}$.

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 20.1.4 *A linear transformation mapping X to Y is called* ***onto*** *if whenever $\mathbf{y} \in Y$ there exists $\mathbf{x} \in X$ such that $T(\mathbf{x}) = \mathbf{y}$.*

Thus T is onto if everything in Y 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 TX, $T(X)$,or $\mathrm{Im}(T)$ as the set of vectors of Y which are of the form $T\mathbf{x}$ for some $\mathbf{x} \in X$. In the case that T is obtained from multiplication by an $m \times n$ matrix, A, it is standard to simply write $A(\mathbb{F}^n)$ $A\mathbb{F}^n$, or $\mathrm{Im}(A)$ to denote those vectors in $\mathbb{F}^m$ which are obtained in the form $A\mathbf{x}$ for some $\mathbf{x} \in \mathbb{F}^n$.

20.2 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

$$A\mathbf{x} = T\mathbf{x} \tag{20.1}$$

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 (20.1). 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,

$$\begin{aligned} T\mathbf{x} &= \sum_{i=1}^{n} x_i T(\mathbf{e}_i) \\ &= \begin{pmatrix} | & & | \\ T(\mathbf{e}_1) & \cdots & T(\mathbf{e}_n) \\ | & & | \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\ &\equiv A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \end{aligned}$$

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 20.2.1 *Let T be a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$. Then the matrix, A satisfying (20.1) is given by*

$$\begin{pmatrix} | & & | \\ T(\mathbf{e}_1) & \cdots & T(\mathbf{e}_n) \\ | & & | \end{pmatrix}$$

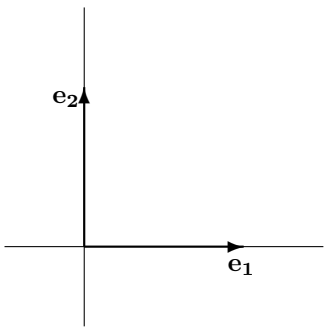
where $T\mathbf{e}_i$ is the i^{th} column of A.

20.2.1 Rotations Of $\mathbb{R}^2$

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.

Example 20.2.2 *Determine the matrix which represents the linear transformation defined by rotating every vector through an angle of θ.*

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 drawing a picture and doing a little geometry, you see that

$$T\mathbf{e}_1 = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, T\mathbf{e}_2 = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}.$$

Therefore, from Theorem 20.2.1,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Example 20.2.3 *Find the matrix of the linear transformation which is obtained by first rotating all vectors through an angle of ϕ and then through an angle θ. Thus you want the linear transformation which rotates all angles through an angle of $\theta + \phi$.*

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_ϕ and then do T_θ where T_ϕ is the linear transformation which rotates through an angle of ϕ and T_θ is the linear transformation which rotates through an angle of θ. Denoting the corresponding matrices by $A_{\theta+\phi}$, A_ϕ, and A_θ, you must have for every $\mathbf{x}$

$$A_{\theta+\phi}\mathbf{x} = T_{\theta+\phi}\mathbf{x} = T_\theta T_\phi \mathbf{x} = A_\theta A_\phi \mathbf{x}.$$

Consequently, you must have

$$\begin{aligned} A_{\theta+\phi} &= \begin{pmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{pmatrix} = A_\theta A_\phi \\ &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}. \end{aligned}$$

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.

20.2.2 Projections

In Physics it is important to consider the work done by a force field on an object. This involves the concept of projection onto a vector. Suppose you want to find the projection of a vector, $\mathbf{v}$ onto the given vector, $\mathbf{u}$, denoted by $\mathrm{proj}_{\mathbf{u}}(\mathbf{v})$ This is done using the dot product as follows.

$$\mathrm{proj}_{\mathbf{u}}(\mathbf{v}) = \left(\frac{\mathbf{v}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\right)\mathbf{u}$$

Because of properties of the dot product, the map $\mathbf{v} \to \mathrm{proj}_{\mathbf{u}}(\mathbf{v})$ is linear,

$$\begin{aligned}\mathrm{proj}_{\mathbf{u}}(\alpha\mathbf{v}+\beta\mathbf{w}) &= \left(\frac{\alpha\mathbf{v}+\beta\mathbf{w}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\right)\mathbf{u} = \alpha\left(\frac{\mathbf{v}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\right)\mathbf{u} + \beta\left(\frac{\mathbf{w}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\right)\mathbf{u} \\ &= \alpha\,\mathrm{proj}_{\mathbf{u}}(\mathbf{v}) + \beta\,\mathrm{proj}_{\mathbf{u}}(\mathbf{w}).\end{aligned}$$

Example 20.2.4 *Let the projection map be defined above and let* $\mathbf{u} = (1,2,3)^T$. *Does this linear transformation come from multiplication by a matrix? If so, what is the matrix?*

You can find this matrix in the same way as in the previous example. Let $\mathbf{e}_i$ denote the vector in $\mathbb{R}^n$ which has a 1 in the i^{th} position and a zero everywhere else. Thus a typical vector, $\mathbf{x} = (x_1, \cdots, x_n)^T$ can be written in a unique way as

$$\mathbf{x} = \sum_{j=1}^{n} x_j \mathbf{e}_j.$$

From the way you multiply a matrix by a vector, it follows that $\mathrm{proj}_{\mathbf{u}}(\mathbf{e}_i)$ gives the i^{th} column of the desired matrix. Therefore, it is only necessary to find

$$\mathrm{proj}_{\mathbf{u}}(\mathbf{e}_i) \equiv \left(\frac{\mathbf{e}_i\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\right)\mathbf{u}$$

For the given vector in the example, this implies the columns of the desired matrix are

$$\frac{1}{14}\begin{pmatrix}1\\2\\3\end{pmatrix}, \frac{2}{14}\begin{pmatrix}1\\2\\3\end{pmatrix}, \frac{3}{14}\begin{pmatrix}1\\2\\3\end{pmatrix}.$$

Hence the matrix is

$$\frac{1}{14}\begin{pmatrix}1&2&3\\2&4&6\\3&6&9\end{pmatrix}.$$

20.2.3 Matrices Which Are One To One Or Onto

Lemma 20.2.5 *Let* A *be an* $m\times n$ *matrix. Then* $A(\mathbb{F}^n) = \mathrm{span}(\mathbf{a}_1, \cdots, \mathbf{a}_n)$ *where* $\mathbf{a}_1, \cdots, \mathbf{a}_n$ *denote the columns of* A. *In fact, for* $\mathbf{x} = (x_1, \cdots, x_n)^T$,

$$A\mathbf{x} = \sum_{k=1}^{n} x_k \mathbf{a}_k.$$

Proof: This follows from the definition of matrix multiplication in Definition 14.1.9 on Page 370.

The following is a theorem of major significance. First here is an interesting observation.

Observation 20.2.6 *Let* A *be an* $m\times n$ *matrix. Then* A *is one to one if and only if* $A\mathbf{x} = \mathbf{0}$ *implies* $\mathbf{x} = \mathbf{0}$.

Here is why: $A\mathbf{0} = A(\mathbf{0}+\mathbf{0}) = A\mathbf{0} + A\mathbf{0}$ and so $A\mathbf{0} = \mathbf{0}$.

Now suppose A is one to one and $A\mathbf{x} = \mathbf{0}$. Then since $A\mathbf{0} = \mathbf{0}$, it follows $\mathbf{x} = \mathbf{0}$. Thus if A is one to one and $A\mathbf{x} = \mathbf{0}$, then $\mathbf{x} = \mathbf{0}$.

Next suppose the condition that $A\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$ is valid. Then if $A\mathbf{x} = A\mathbf{y}$, then $A(\mathbf{x}-\mathbf{y}) = \mathbf{0}$ and so from the condition, $\mathbf{x}-\mathbf{y} = \mathbf{0}$ so that $\mathbf{x} = \mathbf{y}$. Thus A is one to one.

Theorem 20.2.7 *Suppose A is an $n \times n$ matrix. Then A is one to one if and only if A is onto. Also, if B is an $n \times n$ matrix and $AB = I$, then it follows $BA = I$.*

Proof: First suppose A is one to one. Consider the vectors, $\{A\mathbf{e}_1, \cdots, A\mathbf{e}_n\}$ where $\mathbf{e}_k$ is the column vector which is all zeros except for a 1 in the k^{th} position. This set of vectors is linearly independent because if

$$\sum_{k=1}^{n} c_k A\mathbf{e}_k = \mathbf{0},$$

then since A is linear,

$$A\left(\sum_{k=1}^{n} c_k \mathbf{e}_k\right) = \mathbf{0}$$

and since A is one to one, it follows

$$\sum_{k=1}^{n} c_k \mathbf{e}_k = \mathbf{0}$$

which implies each $c_k = 0$. Therefore, $\{A\mathbf{e}_1, \cdots, A\mathbf{e}_n\}$ must be a basis for $\mathbb{F}^n$ because if not there would exist a vector, $\mathbf{y} \notin \text{span}(A\mathbf{e}_1, \cdots, A\mathbf{e}_n)$ and then by Lemma 16.1.41, $\{A\mathbf{e}_1, \cdots, A\mathbf{e}_n, \mathbf{y}\}$ would be an independent set of vectors having $n+1$ vectors in it, contrary to the exchange theorem. It follows that for $\mathbf{y} \in \mathbb{F}^n$ there exist constants, c_i such that

$$\mathbf{y} = \sum_{k=1}^{n} c_k A\mathbf{e}_k = A\left(\sum_{k=1}^{n} c_k \mathbf{e}_k\right)$$

showing that, since $\mathbf{y}$ was arbitrary, A is onto.

Next suppose A is onto. By Lemma 20.2.5, this means the span of the columns of A equals $\mathbb{F}^n$. If these columns are not linearly independent, then by Lemma 16.1.28 on Page 426, one of the columns is a linear combination of the others and so the span of the columns of A equals the span of the $n-1$ other columns. This violates the exchange theorem because $\{\mathbf{e}_1, \cdots, \mathbf{e}_n\}$ would be a linearly independent set of vectors contained in the span of only $n-1$ vectors. Therefore, the columns of A must be independent and by Lemma 20.2.5 this is equivalent to saying that $A\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{0}$. This implies A is one to one because if $A\mathbf{x} = A\mathbf{y}$, then $A(\mathbf{x} - \mathbf{y}) = \mathbf{0}$ and so $\mathbf{x} - \mathbf{y} = \mathbf{0}$.

Now suppose $AB = I$. Why is $BA = I$? Since $AB = I$ it follows B is one to one since otherwise, there would exist, $\mathbf{x} \neq \mathbf{0}$ such that $B\mathbf{x} = \mathbf{0}$ and then $AB\mathbf{x} = A\mathbf{0} = \mathbf{0} \neq I\mathbf{x}$. Therefore, from what was just shown, B is also onto. In addition to this, A must be one to one because if $A\mathbf{y} = \mathbf{0}$, then $\mathbf{y} = B\mathbf{x}$ for some $\mathbf{x}$ and then $\mathbf{x} = AB\mathbf{x} = A\mathbf{y} = \mathbf{0}$ showing $\mathbf{y} = \mathbf{0}$. Now from what is given to be so, it follows $(AB)A = A$ and so using the associative law for matrix multiplication,

$$A(BA) - A = A(BA - I) = 0.$$

But this means $(BA - I)\mathbf{x} = \mathbf{0}$ for all $\mathbf{x}$ since otherwise, A would not be one to one. Hence $BA = I$ as claimed. This proves the theorem.

This theorem shows that if an $n \times n$ matrix, B acts like an inverse when multiplied on one side of A it follows that $B = A^{-1}$and it will act like an inverse on both sides of A.

The conclusion of this theorem pertains to square matrices only. For example, let

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \; B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & -1 \end{pmatrix} \tag{20.2}$$

Then

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

but

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

There is also an important characterization in terms of determinants.

Theorem 20.2.8 *Let A be an $n \times n$ matrix and let T_A denote the linear transformation determined by A. Then the following are equivalent.*

1. *T_A is one to one.*
2. *T_A is onto.*
3. *$\det(A) \neq 0$.*

20.2.4 The General Solution Of A Linear System

Recall the following definition which was discussed above.

Definition 20.2.9 *T is a **linear transformation** if whenever $\mathbf{x}, \mathbf{y}$ are vectors and a, b scalars,*

$$T(a\mathbf{x} + b\mathbf{y}) = aT\mathbf{x} + bT\mathbf{y}.$$

Thus linear transformations distribute across addition and pass scalars to the outside. A linear system is one which is of the form

$$T\mathbf{x} = \mathbf{b}.$$

*If $T\mathbf{x}_p = \mathbf{b}$, then $\mathbf{x}_p$ is called a **particular solution** to the linear system.*

For example, if A is an $m \times n$ matrix and T_A is determined by

$$T_A(\mathbf{x}) = A\mathbf{x},$$

then from the properties of matrix multiplication, T_A is a linear transformation. In this setting, we will usually write A for the linear transformation as well as the matrix. There are many other examples of linear transformations other than this. In differential equations, you will encounter linear transformations which act on functions to give new functions. In this case, the functions are considered as vectors.

Definition 20.2.10 *Let T be a linear transformation. Define*

$$\ker(T) \equiv \{\mathbf{x} : T\mathbf{x} = \mathbf{0}\}.$$

Thus $\ker(T)$ consists of the set of all vectors which T sends to $\mathbf{0}$.

The above definition states that $\ker(T)$ is the set of solutions to the equation,

$$T\mathbf{x} = \mathbf{0}.$$

In the case where T is really a matrix, you have been solving such equations for quite some time. However, sometimes linear transformations act on vectors which are not in $\mathbb{F}^n$.

Example 20.2.11 *Let* $\frac{d}{dx}$ *denote the linear transformation defined on* X*, the functions which are defined on* $\mathbb{R}$ *and have a continuous derivative. Find* $\ker\left(\frac{d}{dx}\right)$.

The example asks for functions, f which the property that $\frac{df}{dx} = 0$. As you know from calculus, these functions are the constant functions. Thus $\ker\left(\frac{d}{dx}\right)$ = constant functions.

When T is a linear transformation, systems of the form $T\mathbf{x} = \mathbf{0}$ are called **homogeneous systems**. Thus the solution to the homogeneous system is known as $\ker(T)$.

Systems of the form $T\mathbf{x} = \mathbf{b}$ where $\mathbf{b} \neq \mathbf{0}$ are called **nonhomogeneous systems**. It turns out there is a very interesting and important relation between the solutions to the homogeneous systems and the solutions to the nonhomogeneous systems.

Theorem 20.2.12 *Suppose* $\mathbf{x}_p$ *is a solution to the linear system,*

$$T\mathbf{x} = \mathbf{b}$$

Then if $\mathbf{y}$ *is any other solution to the linear system, there exists* $\mathbf{x} \in \ker(T)$ *such that*

$$\mathbf{y} = \mathbf{x}_p + \mathbf{x}.$$

Proof: Consider $\mathbf{y} - \mathbf{x}_p \equiv \mathbf{y} + (-1)\mathbf{x}_p$. Then $T\left(\mathbf{y} - \mathbf{x}_p\right) = T\mathbf{y} - T\mathbf{x}_p = \mathbf{b} - \mathbf{b} = \mathbf{0}$. Let $\mathbf{x} \equiv \mathbf{y} - \mathbf{x}_p$. This proves the theorem.

Sometimes people remember the above theorem in the following form. The solutions to the nonhomogeneous system, $T\mathbf{x} = \mathbf{b}$ are given by $\mathbf{x}_p + \ker(T)$ where $\mathbf{x}_p$ is a particular solution to $T\mathbf{x} = \mathbf{b}$.

We have been vague about what T is and what $\mathbf{x}$ is on purpose. This theorem is completely algebraic in nature and will work whenever you have linear transformations. In particular, it will be important in differential equations. For now, here is a familiar example.

Example 20.2.13 *Let*

$$A = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 2 & 1 & 1 & 2 \\ 4 & 5 & 7 & 2 \end{pmatrix}$$

Find $\ker(A)$.

This asks you to find $\{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$. In other words you are asked to solve the system, $A\mathbf{x} = \mathbf{0}$. Let $\mathbf{x} = (x, y, z, w)^T$. Then this amounts to solving

$$\begin{pmatrix} 1 & 2 & 3 & 0 \\ 2 & 1 & 1 & 2 \\ 4 & 5 & 7 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This is the linear system

$$\begin{aligned} x + 2y + 3z &= 0 \\ 2x + y + z + 2w &= 0 \\ 4x + 5y + 7z + 2w &= 0 \end{aligned}$$

and you know how to solve this using row operations, (Gauss Elimination). Set up the augmented matrix,

$$\left(\begin{array}{cccc|c} 1 & 2 & 3 & 0 & 0 \\ 2 & 1 & 1 & 2 & 0 \\ 4 & 5 & 7 & 2 & 0 \end{array}\right)$$

Then row reduce to obtain a reduced echelon form,

$$\left(\begin{array}{cccc|c} 1 & 0 & -\frac{1}{3} & \frac{4}{3} & 0 \\ 0 & 1 & \frac{5}{3} & -\frac{2}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

This yields $x = \frac{1}{3}z - \frac{4}{3}w$ and $y = \frac{2}{3}w - \frac{5}{3}z$. Thus $\ker(A)$ consists of vectors of the form,

$$\begin{pmatrix} \frac{1}{3}z - \frac{4}{3}w \\ \frac{2}{3}w - \frac{5}{3}z \\ z \\ w \end{pmatrix} = z\begin{pmatrix} \frac{1}{3} \\ -\frac{5}{3} \\ 1 \\ 0 \end{pmatrix} + w\begin{pmatrix} -\frac{4}{3} \\ \frac{2}{3} \\ 0 \\ 1 \end{pmatrix}.$$

Example 20.2.14 *The **general solution** of a linear system of equations is just the set of all solutions. Find the general solution to the linear system,*

$$\begin{pmatrix} 1 & 2 & 3 & 0 \\ 2 & 1 & 1 & 2 \\ 4 & 5 & 7 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 9 \\ 7 \\ 25 \end{pmatrix}$$

given that $\begin{pmatrix} 1 & 1 & 2 & 1 \end{pmatrix}^T = \begin{pmatrix} x & y & z & w \end{pmatrix}^T$ *is one solution.*

Note the matrix on the left is the same as the matrix in Example 20.2.13. Therefore, from Theorem 20.2.12, you will obtain all solutions to the above linear system in the form

$$z\begin{pmatrix} \frac{1}{3} \\ -\frac{5}{3} \\ 1 \\ 0 \end{pmatrix} + w\begin{pmatrix} -\frac{4}{3} \\ \frac{2}{3} \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 2 \\ 1 \end{pmatrix}.$$

20.3 Exercises

1. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/3$.
2. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/4$.
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 $2\pi/3$.
5. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/12$. **Hint:** Note that $\pi/12 = \pi/3 - \pi/4$.
6. 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.

7. 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.

8. 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.

9. 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.

10. 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 $\pi/4$.

11. 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 $\pi/4$.

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 $\pi/6$.

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 $\pi/6$.

14. 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 $5\pi/12 = 2\pi/3 - \pi/4$.

15. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1, -2, 3)^T$.

16. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1, 5, 3)^T$.

17. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1, 0, 3)^T$.

18. Show that the function $T_{\mathbf{u}}$ defined by $T_{\mathbf{u}}(\mathbf{v}) \equiv \mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v})$ is also a linear transformation.

19. If $\mathbf{u} = (1, 2, 3)^T$, as in Example 20.2.4 and $T_{\mathbf{u}}$ is given in the above problem, find the matrix, $A_{\mathbf{u}}$ which satisfies $A_{\mathbf{u}}\mathbf{x} = T(\mathbf{x})$.

20. If A, B, and C are each $n \times n$ matrices and ABC is invertible, why are each of A, B, and C invertible.

21. Show that $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ by doing the computation $ABC\left(C^{-1}B^{-1}A^{-1}\right)$.

22. If A is invertible, show $\left(A^T\right)^{-1} = \left(A^{-1}\right)^T$.

23. If A is invertible, show $\left(A^2\right)^{-1} = \left(A^{-1}\right)^2$.

24. If A is invertible, show $\left(A^{-1}\right)^{-1} = A$.

25. Give an example of a 3×2 matrix with the property that the linear transformation determined by this matrix is one to one but not onto.

26. Explain why $A\mathbf{x} = \mathbf{0}$ always has a solution.

27. Suppose $\det(A - \lambda I) = 0$. Show using Theorem 20.2.8 there exists $\mathbf{x} \neq \mathbf{0}$ such that $(A - \lambda I)\mathbf{x} = \mathbf{0}$.

28. Let A be an $n \times n$ matrix and let $\mathbf{x}$ be a nonzero vector such that $A\mathbf{x} = \lambda\mathbf{x}$ for some scalar, λ. When this occurs, the vector, $\mathbf{x}$ is called an **eigenvector** and the scalar, λ is called an **eigenvalue**. It turns out that not every number is an eigenvalue. Only certain ones are. Why? **Hint:** Show that if $A\mathbf{x} = \lambda\mathbf{x}$, then $(A - \lambda I)\mathbf{x} = \mathbf{0}$. Explain why this shows that $(A - \lambda I)$ is not one to one and not onto. Now use Theorem 20.2.8 to argue $\det(A - \lambda I) = 0$. What sort of equation is this? How many solutions does it have?

29. 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 $A_1\mathbf{x}$ is the vector,

$$A_1\mathbf{x} = \begin{pmatrix} A\mathbf{x} \\ \mathbf{0} \end{pmatrix}$$

which equals zero if and only if $A\mathbf{x} = \mathbf{0}$.

30. Find $\ker(A)$ for

$$A = \begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix}.$$

Recall $\ker(A)$ is just the set of solutions to $A\mathbf{x} = \mathbf{0}$.

31. Suppose $A\mathbf{x} = \mathbf{b}$ has a solution. Explain why the solution is unique precisely when $A\mathbf{x} = \mathbf{0}$ has only the trivial (zero) solution.

32. Using Problem 30, find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 11 \\ 7 \\ 18 \\ 7 \end{pmatrix}$$

33. Using Problem 30, find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 6 \\ 7 \\ 13 \\ 7 \end{pmatrix}$$

34. Show that if A is an $m \times n$ matrix, then $\ker(A)$ is a subspace.

35. Sometimes it is required to consider rotations in three dimensions. An example is the Euler angles in the mechanics of a rotating body. Describe how you could use the concepts of matrix multiplication to systematically study such rotations.

36. Verify the linear transformation determined by the matrix of (20.2) maps $\mathbb{R}^3$ onto $\mathbb{R}^2$ but the linear transformation determined by this matrix is not one to one.

Spectral Theory

21.0.1 Outcomes

1. Describe the eigenvalue problem geometrically and algebraically.
2. Evaluate the spectrum and eigenvectors for a square matrix.
3. Use the determinant of the Grammian matrix to compute volumes of k dimensional parallelepipeds.
4. Recall and use block multiplication.
5. Read and understand the proof of Schur's theorem.

21.1 Eigenvalues And Eigenvectors Of A Matrix

Spectral Theory refers to the study of eigenvalues and eigenvectors of a matrix. It is of fundamental importance in many areas. Row operations will no longer be such a useful tool in this subject.

21.1.1 Definition Of Eigenvectors And Eigenvalues

In this section, $\mathbb{F} = \mathbb{C}$.

To illustrate the idea behind what will be discussed, consider the following example.

Example 21.1.1 *Here is a matrix.*

$$\begin{pmatrix} 0 & 5 & -10 \\ 0 & 22 & 16 \\ 0 & -9 & -2 \end{pmatrix}.$$

Multiply this matrix by the vector

$$\begin{pmatrix} -5 \\ -4 \\ 3 \end{pmatrix}$$

and see what happens. Then multiply it by

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

and see what happens. Does this matrix act this way for some other vector?

First

$$\begin{pmatrix} 0 & 5 & -10 \\ 0 & 22 & 16 \\ 0 & -9 & -2 \end{pmatrix} \begin{pmatrix} -5 \\ -4 \\ 3 \end{pmatrix} = \begin{pmatrix} -50 \\ -40 \\ 30 \end{pmatrix} = 10 \begin{pmatrix} -5 \\ -4 \\ 3 \end{pmatrix}.$$

Next

$$\begin{pmatrix} 0 & 5 & -10 \\ 0 & 22 & 16 \\ 0 & -9 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = 0 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

When you multiply the first vector by the given matrix, it stretched the vector, multiplying it by 10. When you multiplied the matrix by the second vector it sent it to the zero vector. Now consider

$$\begin{pmatrix} 0 & 5 & -10 \\ 0 & 22 & 16 \\ 0 & -9 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -5 \\ 38 \\ -11 \end{pmatrix}.$$

In this case, multiplication by the matrix did not result in merely multiplying the vector by a number.

In the above example, the first two vectors were called eigenvectors and the numbers, 10 and 0 are called eigenvalues. Not every number is an eigenvalue and not every vector is an eigenvector.

Definition 21.1.2 *Let M be an $n \times n$ matrix and let $\mathbf{x} \in \mathbb{C}^n$ be a* ***nonzero vector*** *for which*

$$M\mathbf{x} = \lambda\mathbf{x} \tag{21.1}$$

for some scalar, λ. Then $\mathbf{x}$ is called an ***eigenvector*** *and λ is called an* ***eigenvalue*** *(**characteristic value**) of the matrix, M.*

Note: Eigenvectors are never equal to zero!

The set of all eigenvalues of an $n \times n$ matrix, M, is denoted by $\sigma(M)$ and is referred to as the ***spectrum*** *of M.*

The eigenvectors of a matrix M are those vectors, $\mathbf{x}$ for which multiplication by M results in a vector in the same direction or opposite direction to $\mathbf{x}$. Since the zero vector, $\mathbf{0}$ has no direction this would make no sense for the zero vector. As noted above, $\mathbf{0}$ is never allowed to be an eigenvector. How can eigenvectors be identified? Suppose $\mathbf{x}$ satisfies (21.1). Then

$$(M - \lambda I)\mathbf{x} = \mathbf{0}$$

for some $\mathbf{x} \neq \mathbf{0}$. (Equivalently, you could write $(\lambda I - M)\mathbf{x} = \mathbf{0}$.) Sometimes we will use

$$(\lambda I - M)\mathbf{x} = \mathbf{0}$$

and sometimes

$$(M - \lambda I)\mathbf{x} = \mathbf{0}.$$

It makes absolutely no difference and you should use whichever you like better. Therefore, the matrix $M - \lambda I$ cannot have an inverse because if it did, the equation could be solved,

$$\mathbf{x} = \left((M - \lambda I)^{-1}(M - \lambda I)\right)\mathbf{x} = (M - \lambda I)^{-1}((M - \lambda I)\mathbf{x}) = (M - \lambda I)^{-1}\mathbf{0} = \mathbf{0},$$

and this would require $\mathbf{x} = \mathbf{0}$, contrary to the requirement that $\mathbf{x} \neq \mathbf{0}$. By Theorem 15.2.1 on Page 393,

$$\det(M - \lambda I) = 0. \tag{21.2}$$

(Equivalently you could write $\det(\lambda I - M) = 0$.) The expression, $\det(\lambda I - M)$ or equivalently, $\det(M - \lambda I)$ is a polynomial called the **characteristic polynomial** and the above equation is called the characteristic equation. For M an $n \times n$ matrix, it follows from the theorem on expanding a matrix by its cofactor that $\det(M - \lambda I)$ is a polynomial of degree n. As such, the equation, (21.2) has a solution, $\lambda \in \mathbb{C}$ by the fundamental theorem of algebra. Is it actually an eigenvalue? The answer is yes and this follows from Observation 20.2.6 on Page 501 along with Theorem 15.2.1 on Page 393. Since $\det(M - \lambda I) = 0$ the matrix, $\det(M - \lambda I)$ cannot be one to one and so there exists a nonzero vector, $\mathbf{x}$ such that $(M - \lambda I)\mathbf{x} = \mathbf{0}$. This proves the following corollary.

Corollary 21.1.3 *Let M be an $n \times n$ matrix and $\det(M - \lambda I) = 0$. Then there exists a nonzero vector, $\mathbf{x} \in \mathbb{C}^n$ such that $(M - \lambda I)\mathbf{x} = \mathbf{0}$.*

21.1.2 Finding Eigenvectors And Eigenvalues

As an example, consider the following.

Example 21.1.4 *Find the eigenvalues and eigenvectors for the matrix,*

$$A = \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix}.$$

You first need to identify the eigenvalues. Recall this requires the solution of the equation

$$\det(A - \lambda I) = 0.$$

In this case this equation is

$$\det\left(\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right) = 0$$

When you expand this determinant and simplify, you find the equation you need to solve is

$$(\lambda - 5)\left(\lambda^2 - 20\lambda + 100\right) = 0$$

and so the eigenvalues are

$$5, 10, 10.$$

We have listed 10 twice because it is a zero of multiplicity two due to

$$\lambda^2 - 20\lambda + 100 = (\lambda - 10)^2.$$

Having found the eigenvalues, it only remains to find the eigenvectors. First find the eigenvectors for $\lambda = 5$. As explained above, this requires you to solve the equation,

$$\left(\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} - 5\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

That is you need to find the solution to

$$\begin{pmatrix} 0 & -10 & -5 \\ 2 & 9 & 2 \\ -4 & -8 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

By now this is an old problem. You set up the augmented matrix and row reduce to get the solution. Thus the matrix you must row reduce is

$$\left(\begin{array}{ccc|c} 0 & -10 & -5 & 0 \\ 2 & 9 & 2 & 0 \\ -4 & -8 & 1 & 0 \end{array}\right). \tag{21.3}$$

A reduced echelon form is

$$\left(\begin{array}{ccc|c} 1 & 0 & -\frac{5}{4} & 0 \\ 0 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

and so the solution is any vector of the form

$$\begin{pmatrix} \frac{5}{4}t \\ \frac{-1}{2}t \\ t \end{pmatrix} = t \begin{pmatrix} \frac{5}{4} \\ \frac{-1}{2} \\ 1 \end{pmatrix}$$

where $t \in \mathbb{F}$. You would obtain the same collection of vectors if you replaced t with $4t$. Thus a simpler description for the solutions to this system of equations whose augmented matrix is in (21.3) is

$$t \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix} \tag{21.4}$$

where $t \in \mathbb{F}$. Now you need to remember that you can't take $t = 0$ because this would result in the zero vector and

Eigenvectors are never equal to zero!

Other than this value, every other choice of z in (21.4) results in an eigenvector. It is a good idea to check your work! To do so, we will take the original matrix and multiply by this vector and see if we get 5 times this vector.

$$\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix} = \begin{pmatrix} 25 \\ -10 \\ 20 \end{pmatrix} = 5 \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix}$$

so it appears this is correct. Always check your work on these problems if you care about getting the answer right.

The parameter, t is sometimes called a **free variable.** The set of vectors in (21.4) is called the **eigenspace** and it equals $\ker(A - \lambda I)$. You should observe that in this case the eigenspace has dimension 1 because the eigenspace is the span of a single vector. In general, you obtain the solution from the row echelon form and the number of different free variables gives you the dimension of the eigenspace. Just remember that not every vector in the eigenspace is an eigenvector. The vector, $\mathbf{0}$ is not an eigenvector although it is in the eigenspace because

Eigenvectors are never equal to zero!

Next consider the eigenvectors for $\lambda = 10$. These vectors are solutions to the equation,

$$\left(\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} - 10\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

That is you must find the solutions to

$$\begin{pmatrix} -5 & -10 & -5 \\ 2 & 4 & 2 \\ -4 & -8 & -4 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

which reduces to consideration of the augmented matrix,

$$\left(\begin{array}{ccc|c} -5 & -10 & -5 & 0 \\ 2 & 4 & 2 & 0 \\ -4 & -8 & -4 & 0 \end{array}\right)$$

A reduced echelon form for this matrix is

$$\begin{pmatrix} 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and so the eigenvectors are of the form

$$\begin{pmatrix} -2s - t \\ s \\ t \end{pmatrix} = s\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + t\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

You can't pick t and s both equal to zero because this would result in the zero vector and

Eigenvectors are never equal to zero!

However, every other choice of t and s does result in an eigenvector for the eigenvalue $\lambda = 10$. As in the case for $\lambda = 5$ you should check your work if you care about getting it right.

$$\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix}\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -10 \\ 0 \\ 10 \end{pmatrix} = 10\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

so it worked. The other vector will also work. Check it.

21.1.3 A Warning

The above example shows how to find eigenvectors and eigenvalues algebraically. You may have noticed it is a bit long. Sometimes students try to first row reduce the matrix before looking for eigenvalues. This is a **terrible idea** because row operations destroy the eigenvalues. The eigenvalue problem is really not about row operations.

The general eigenvalue problem is the hardest problem in algebra and people still do research on ways to find eigenvalues and their eigenvectors. If you are doing anything which would yield a way to find eigenvalues and eigenvectors for general matrices without too much trouble, the thing you are doing will certainly be wrong. The problems you will see in these notes are not too hard because they are cooked up by us to be easy. Later we

will describe general methods to compute eigenvalues and eigenvectors numerically. These methods work even when the problem is not cooked up to be easy.

If you are so fortunate as to find the eigenvalues as in the above example, then finding the eigenvectors does reduce to row operations and this part of the problem is easy. However, finding the eigenvalues along with the eigenvectors is anything but easy because for an $n \times n$ matrix, it involves solving a polynomial equation of degree n. If you only find a good approximation to the eigenvalue, it won't work. It either is or is not an eigenvalue and if it is not, the only solution to the equation, $(M - \lambda I)\,\mathbf{x} = \mathbf{0}$ will be the zero solution as explained above and

Eigenvectors are never equal to zero!

Here is another example.

Example 21.1.5 *Let*

$$A = \begin{pmatrix} 2 & 2 & -2 \\ 1 & 3 & -1 \\ -1 & 1 & 1 \end{pmatrix}$$

First find the eigenvalues.

$$\det\left(\begin{pmatrix} 2 & 2 & -2 \\ 1 & 3 & -1 \\ -1 & 1 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right) = 0$$

This reduces to $\lambda^3 - 6\lambda^2 + 8\lambda = 0$ and the solutions are 0, 2, and 4.

0 **Can be an Eigenvalue!**

Now find the eigenvectors. For $\lambda = 0$ the augmented matrix for finding the solutions is

$$\left(\begin{array}{ccc|c} 2 & 2 & -2 & 0 \\ 1 & 3 & -1 & 0 \\ -1 & 1 & 1 & 0 \end{array}\right)$$

and the a reduced echelon form is

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

Therefore, the eigenvectors are of the form

$$t \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

where $t \neq 0$.

Next find the eigenvectors for $\lambda = 2$. The augmented matrix for the system of equations needed to find these eigenvectors is

$$\left(\begin{array}{ccc|c} 0 & 2 & -2 & 0 \\ 1 & 1 & -1 & 0 \\ -1 & 1 & -1 & 0 \end{array}\right)$$

and the a reduced echelon form is

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

and so the eigenvectors are of the form

$$t \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

where $t \neq 0$.

Finally find the eigenvectors for $\lambda = 4$. The augmented matrix for the system of equations needed to find these eigenvectors is

$$\left(\begin{array}{ccc|c} -2 & 2 & -2 & 0 \\ 1 & -1 & -1 & 0 \\ -1 & 1 & -3 & 0 \end{array}\right)$$

and a reduced echelon form is

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

Therefore, the eigenvectors are of the form

$$t \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

where $t \neq 0$.

21.1.4 Complex Eigenvalues

Sometimes you have to consider eigenvalues which are complex numbers. This occurs in differential equations for example. You do these problems exactly the same way as you do the ones in which the eigenvalues are real. Here is an example.

Example 21.1.6 *Find the eigenvalues and eigenvectors of the matrix*

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix}.$$

You need to find the eigenvalues. Solve

$$\det\left(\left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{array}\right) - \lambda \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right)\right) = 0.$$

This reduces to $(\lambda - 1)\left(\lambda^2 - 4\lambda + 5\right) = 0$. The solutions are $\lambda = 1, \lambda = 2 + i, \lambda = 2 - i$.

There is nothing new about finding the eigenvectors for $\lambda = 1$ so consider the eigenvalue $\lambda = 2 + i$. You need to solve

$$\left((2+i)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

In other words, you must consider the augmented matrix,

$$\left(\begin{array}{ccc|c} 1+i & 0 & 0 & 0 \\ 0 & i & 1 & 0 \\ 0 & -1 & i & 0 \end{array}\right)$$

for the solution. Divide the top row by $(1+i)$ and then take $-i$ times the second row and add to the bottom. This yields

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & i & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Now multiply the second row by $-i$ to obtain

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & -i & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Therefore, the eigenvectors are of the form

$$t\begin{pmatrix} 0 \\ i \\ 1 \end{pmatrix}.$$

You should find the eigenvectors for $\lambda = 2-i$. These are

$$t\begin{pmatrix} 0 \\ -i \\ 1 \end{pmatrix}.$$

As usual, if you want to get it right you had better check it.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 0 \\ -i \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ -1-2i \\ 2-i \end{pmatrix} = (2-i)\begin{pmatrix} 0 \\ -i \\ 1 \end{pmatrix}$$

so it worked.

21.2 Volumes

The determinant and the concept of eigenvalues and eigenvectors provide a way to give a unified treatment of the concept of volumes in various dimensions. First here is a useful theorem which is of considerable interest for its own sake.

Theorem 21.2.1 *Let A be an $n \times n$ matrix. Then*

$$\det(A) = \prod_{i=1}^{n} \lambda_i$$

where λ_i are the eigenvalues of A. In words, the determinant of a matrix equals the product of its eigenvalues.

Proof: The characteristic polynomial is $\det(\lambda I - A) = \prod_{j=1}^{n}(\lambda - \lambda_j)$ where λ_i are the eigenvalues. This follows from the fundamental theorem of algebra which says every polynomial can be factored. Then, letting $\lambda = 0$ it follows $\det(-A) = (-1)^n \det(A) = \prod_{j=1}^{n}(0 - \lambda_j) = (-1)^n \prod_{j=1}^{n} \lambda_j$ and this proves the theorem.

Recall the geometric definition of the cross product of two vectors found on Page 471. As explained there, the magnitude of the cross product of two vectors was the area of the parallelogram determined by the two vectors. There was also a coordinate description of the cross product. In terms of the notation of Proposition 18.5.4 on Page 481 the i^{th} coordinate of the cross product is given by

$$\varepsilon_{ijk} u_j v_k$$

where the two vectors are (u_1, u_2, u_3) and (v_1, v_2, v_3). Therefore, using the reduction identity of Lemma 18.5.3 on Page 481

$$\begin{aligned} |\mathbf{u} \times \mathbf{v}|^2 &= \varepsilon_{ijk} u_j v_k \varepsilon_{irs} u_r v_s \\ &= (\delta_{jr}\delta_{ks} - \delta_{kr}\delta_{js}) u_j v_k u_r v_s \\ &= u_j v_k u_j v_k - u_j v_k u_k v_j \\ &= (\mathbf{u} \cdot \mathbf{u})(\mathbf{v} \cdot \mathbf{v}) - (\mathbf{u} \cdot \mathbf{v})^2 \end{aligned}$$

which equals

$$\det \begin{pmatrix} \mathbf{u} \cdot \mathbf{u} & \mathbf{u} \cdot \mathbf{v} \\ \mathbf{u} \cdot \mathbf{v} & \mathbf{v} \cdot \mathbf{v} \end{pmatrix}.$$

Now recall the box product and how the box product was $\pm$ the volume of the parallelepiped spanned by the three vectors. From the definition of the box product

$$\begin{aligned} \mathbf{u} \times \mathbf{v} \cdot \mathbf{w} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \cdot (w_1\mathbf{i} + w_2\mathbf{j} + w_3\mathbf{k}) \\ &= \det \begin{pmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{pmatrix}. \end{aligned}$$

Therefore,

$$|\mathbf{u} \times \mathbf{v} \cdot \mathbf{w}|^2 = \det \begin{pmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{pmatrix}^2$$

which from the theory of determinants equals

$$\det \begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix} \det \begin{pmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{pmatrix} =$$

$$\det \left(\begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix} \begin{pmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{pmatrix} \right) =$$

$$\det \begin{pmatrix} u_1^2 + u_2^2 + u_3^2 & u_1v_1 + u_2v_2 + u_3v_3 & u_1w_1 + u_2w_2 + u_3w_3 \\ u_1v_1 + u_2v_2 + u_3v_3 & v_1^2 + v_2^2 + v_3^2 & v_1w_1 + v_2w_2 + v_3w_3 \\ u_1w_1 + u_2w_2 + u_3w_3 & v_1w_1 + v_2w_2 + v_3w_3 & w_1^2 + w_2^2 + w_3^2 \end{pmatrix}$$

$$= \det \begin{pmatrix} \mathbf{u} \cdot \mathbf{u} & \mathbf{u} \cdot \mathbf{v} & \mathbf{u} \cdot \mathbf{w} \\ \mathbf{u} \cdot \mathbf{v} & \mathbf{v} \cdot \mathbf{v} & \mathbf{v} \cdot \mathbf{w} \\ \mathbf{u} \cdot \mathbf{w} & \mathbf{v} \cdot \mathbf{w} & \mathbf{w} \cdot \mathbf{w} \end{pmatrix}$$

You see there is a definite pattern emerging here. These earlier cases were for a parallelepiped determined by either two or three vectors in $\mathbb{R}^3$. It makes sense to speak of a parallelepiped in any number of dimensions.

Definition 21.2.2 *Let $\mathbf{u}_1, \cdots, \mathbf{u}_p$ be vectors in $\mathbb{R}^k$. The parallelepiped determined by these vectors will be denoted by $P(\mathbf{u}_1, \cdots, \mathbf{u}_p)$ and it is defined as*

$$P(\mathbf{u}_1, \cdots, \mathbf{u}_p) \equiv \left\{ \sum_{j=1}^{p} s_j \mathbf{u}_j : s_j \in [0,1] \right\}.$$

The volume of this parallelepiped is defined as

$$\text{volume of } P(\mathbf{u}_1, \cdots, \mathbf{u}_p) \equiv \left(\det(\mathbf{u}_i \cdot \mathbf{u}_j)\right)^{1/2}.$$

In this definition, $\mathbf{u}_i \cdot \mathbf{u}_j$ is the ij^{th} entry of a $p \times p$ matrix. Note this definition agrees with all earlier notions of area and volume for parallelepipeds and it makes sense in any number of dimensions. However, it is important to verify the above determinant is nonnegative. After all, the above definition requires a square root of this determinant.

Lemma 21.2.3 *Let $\mathbf{u}_1, \cdots, \mathbf{u}_p$ be vectors in $\mathbb{R}^k$ for some k. Then $\det(\mathbf{u}_i \cdot \mathbf{u}_j) \geq 0$.*

Proof: Recall $\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \mathbf{w}$. Therefore, in terms of matrix multiplication, the matrix $(\mathbf{u}_i \cdot \mathbf{u}_j)$ is just the following

$$\overbrace{\begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_p^T \end{pmatrix}}^{p \times k} \overbrace{\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix}}^{k \times p}$$

which is of the form

$$U^T U.$$

Now the eigenvalues of the matrix $U^T U$ are all nonnegative. Here is why. Suppose $U^T U \mathbf{x} = \lambda \mathbf{x}$. Then

$$0 \leq \overline{\mathbf{x}}^T U^T U \mathbf{x} = \lambda \overline{\mathbf{x}}^T \mathbf{x} = \lambda \sum_k |x_k|^2.$$

Therefore, from Theorem 21.2.1, $\det\left(U^T U\right) \geq 0$ because it is the product of nonnegative numbers. This proves the lemma and shows the definition of volume is well defined.

Note it gives the right answer in the case where all the vectors are perpendicular. Here is why. Suppose $\{\mathbf{u}_1, \cdots, \mathbf{u}_p\}$ are vectors which have the property that $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ if $i \neq j$. Thus $P(\mathbf{u}_1, \cdots, \mathbf{u}_p)$ is a box which has all p sides perpendicular. What should its p dimensional volume be? Shouldn't it equal the product of the lengths of the sides? What does $\det(\mathbf{u}_i \cdot \mathbf{u}_j)$ give? The matrix $(\mathbf{u}_i \cdot \mathbf{u}_j)$ is a diagonal matrix having the squares of the magnitudes of the sides down the diagonal. Therefore, $\det(\mathbf{u}_i \cdot \mathbf{u}_j)^{1/2}$ equals the product of the lengths of the sides as it should.

The matrix, $(\mathbf{u}_i \cdot \mathbf{u}_j)$ whose determinant gives the square of the volume of the parallelepiped spanned by the vectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_p\}$ is called the Grammian matrix and sometimes the metric tensor.

These considerations are of great significance because they allow the computation in a systematic manner of k dimensional volumes of parallelepipeds which happen to be in $\mathbb{R}^n$ for $n \neq k$. Think for example of a plane in $\mathbb{R}^3$ and the problem of finding the area of something on this plane.

Example 21.2.4 *Find the equation of the plane containing the three points,* $(1,2,3)$, $(0,2,1)$, *and* $(3,1,0)$.

These three points determine two vectors, the one from $(0,2,1)$ to $(1,2,3)$, $\mathbf{i}+0\mathbf{j}+2\mathbf{k}$, and the one from $(0,2,1)$ to $(3,1,0)$, $3\mathbf{i}+(-1)\mathbf{j}+(-1)\mathbf{k}$. If (x,y,z) denotes a point in the plane, then the volume of the parallelepiped spanned by the vector from $(0,2,1)$ to (x,y,z) and these other two vectors must be zero. Thus

$$\det\begin{pmatrix} x & y-2 & z-1 \\ 3 & -1 & -1 \\ 1 & 0 & 2 \end{pmatrix} = 0$$

Therefore, $-2x-7y+13+z=0$ is the equation of the plane. You should check it contains all three points.

Example 21.2.5 *In the above example find the two dimensional volume of the parallelogram determined by the two vectors,* $\mathbf{i}+0\mathbf{j}+2\mathbf{k}$ *and* $3\mathbf{i}+(-1)\mathbf{j}+(-1)\mathbf{k}$.

This is easy. Just take the square root of the determinant of the Grammian matrix. Thus the area is

$$\sqrt{\det\begin{pmatrix} 5 & 1 \\ 1 & 11 \end{pmatrix}} = 3\sqrt{6}.$$

21.3 Block Multiplication Of Matrices

Suppose A is a matrix of the form

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{r1} & \cdots & A_{rm} \end{pmatrix} \tag{21.5}$$

where A_{ij} is a $s_i \times p_j$ matrix where s_i does not depend on j and p_j does not depend on i. Such a matrix is called a **block matrix**. Let $n=\sum_j p_j$ and $k=\sum_i s_i$ so A is an $k\times n$ matrix. What is $A\mathbf{x}$ where $\mathbf{x}\in\mathbb{F}^n$? From the process of multiplying a matrix times a vector, the following lemma follows.

Lemma 21.3.1 *Let* A *be an* $m\times n$ *block matrix as in (21.5) and let* $\mathbf{x}\in\mathbb{F}^n$. *Then* $A\mathbf{x}$ *is of the form*

$$A\mathbf{x} = \begin{pmatrix} \sum_j A_{1j}\mathbf{x}_j \\ \vdots \\ \sum_j A_{rj}\mathbf{x}_j \end{pmatrix}$$

where $\mathbf{x}=(\mathbf{x}_1,\cdots,\mathbf{x}_m)^T$ *and* $\mathbf{x}_i\in\mathbb{F}^{p_i}$.

Suppose also that B is a $l\times k$ block matrix of the form

$$\begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{m1} & \cdots & B_{mp} \end{pmatrix} \tag{21.6}$$

and that for all i,j, it makes sense to multiply $B_{is}A_{sj}$ for all $s\in\{1,\cdots,m\}$. (That is the two matrices are conformable.) and that for each s, $B_{is}A_{sj}$ is the same size so that it makes sense to write $\sum_s B_{is}A_{sj}$.

Theorem 21.3.2 *Let B be an $l \times k$ block matrix as in (21.6) and let A be a $k \times n$ block matrix as in (21.5) such that B_{is} is conformable with A_{sj} and each product, $B_{is}A_{sj}$ is of the same size so they can be added. Then BA is a $l \times n$ block matrix having rp blocks such that the ij^{th} block is of the form*

$$\sum_s B_{is}A_{sj}. \tag{21.7}$$

Proof: Let B_{is} be a $q_i \times p_s$ matrix and A_{sj} be a $p_s \times r_j$ matrix. Also let $\mathbf{x} \in \mathbb{F}^n$ and let $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_m)^T$ and $\mathbf{x}_i \in \mathbb{F}^{r_i}$ so it makes sense to multiply $A_{sj}\mathbf{x}_j$. Then from the associative law of matrix multiplication and Lemma 21.3.1 applied twice,

$$\begin{aligned}
(BA)\,\mathbf{x} &= B\,(A\mathbf{x}) \\
&= \begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{m1} & \cdots & B_{mp} \end{pmatrix} \begin{pmatrix} \sum_j A_{1j}\mathbf{x}_j \\ \vdots \\ \sum_j A_{rj}\mathbf{x}_j \end{pmatrix} \\
&= \begin{pmatrix} \sum_s \sum_j B_{1s}A_{sj}\mathbf{x}_j \\ \vdots \\ \sum_s \sum_j B_{ms}A_{sj}\mathbf{x}_j \end{pmatrix} = \begin{pmatrix} \sum_j \left(\sum_s B_{1s}A_{sj}\right)\mathbf{x}_j \\ \vdots \\ \sum_j \left(\sum_s B_{ms}A_{sj}\right)\mathbf{x}_j \end{pmatrix}.
\end{aligned}$$

By Lemma 21.3.1, this shows that $(BA)\,\mathbf{x}$ equals the block matrix whose ij^{th} entry is given by (21.7) times $\mathbf{x}$. Since $\mathbf{x}$ is an arbitrary vector in $\mathbb{F}^n$, this proves the theorem.

The message of this theorem is that you can formally multiply block matrices as though the blocks were numbers. You just have to pay attention to the preservation of order.

This simple idea of block multiplication turns out to be very useful later. For now here is an interesting and significant application. In this theorem, $p_M(t)$ denotes the polynomial, $\det(tI - M)$. Thus the zeros of this polynomial are the eigenvalues of the matrix, M.

Theorem 21.3.3 *Let A be an $m \times n$ matrix and let B be an $n \times m$ matrix for $m \leq n$. Then*

$$p_{BA}(t) = t^{n-m}p_{AB}(t),$$

so the eigenvalues of BA and AB are the same including multiplicities except that BA has $n - m$ extra zero eigenvalues.

Proof: Use block multiplication to write

$$\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}$$

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}.$$

Therefore,

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$$

By Problem 12 of Page 417, it follows that $\begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$ and $\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix}$ have the same characteristic polynomials. Therefore, noting that BA is an $n \times n$ matrix and AB is an $m \times m$ matrix,

$$t^m \det(tI - BA) = t^n \det(tI - AB)$$

and so $\det(tI - BA) = p_{BA}(t) = t^{n-m}\det(tI - AB) = t^{n-m}p_{AB}(t)$. This proves the theorem.

21.4 Shur's Theorem*

Every matrix is related to an upper triangular matrix in a particularly significant way. This is Shur's theorem and it is the most important theorem in the spectral theory of matrices. Recall the Gram Schmidt procedure of Lemma 19.1.3 on Page 486 which is stated here for convenience.

Lemma 21.4.1 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a basis for $\mathbb{F}^n$. Then there exists an orthonormal basis for $\mathbb{F}^n$, $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ which has the property that for each $k \leq n$, span$(\mathbf{x}_1, \cdots, \mathbf{x}_k) =$ span $(\mathbf{u}_1, \cdots, \mathbf{u}_k)$.*

Definition 21.4.2 *An $n \times n$ matrix, U, is unitary if $UU^* = I = U^*U$ where U^* is defined to be the transpose of the conjugate of U. Thus $\overline{U_{ij}} = U^*_{ji}$.*

Theorem 21.4.3 *Let A be an $n \times n$ matrix. Then there exists a unitary matrix, U such that*

$$U^*AU = T, \tag{21.8}$$

where T is an upper triangular matrix having the eigenvalues of A on the main diagonal listed according to multiplicity as roots of the characteristic equation.

Proof: Let $\mathbf{v}_1$ be a unit eigenvector for A . Then there exists λ_1 such that

$$A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1, \ |\mathbf{v}_1| = 1.$$

Extend $\{\mathbf{v}_1\}$ to a basis and then use the Gram Schmidt procedure to obtain $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$, an orthonormal basis in $\mathbb{F}^n$. Let U_0 be a matrix whose i^{th} column is $\mathbf{v}_i$. Then from the above, it follows U_0 is unitary. Then $U_0^*AU_0$ 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. Repeat the process for the matrix, A_1 above. There exists a unitary matrix $\widetilde{U}_1$ such that $\widetilde{U}_1^* A_1 \ \widetilde{U}_1$ is of the form

$$\begin{pmatrix} \lambda_2 & * & \cdots & * \\ 0 & & & \\ \vdots & & A_2 & \\ 0 & & & \end{pmatrix}.$$

Now let U_1 be the $n \times n$ matrix of the form

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix}.$$

This is also a unitary matrix because by block multiplication,

$$\begin{aligned} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix}^* \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix} &= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1^* \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1^* \widetilde{U}_1 \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix} \end{aligned}$$

Then using block multiplication, $U_1^* U_0^* A U_0 U_1$ is of the form

$$\begin{pmatrix} \lambda_1 & * & * & \cdots & * \\ 0 & \lambda_2 & * & \cdots & * \\ 0 & 0 & & & \\ \vdots & \vdots & & A_2 & \\ 0 & 0 & & & \end{pmatrix}$$

where A_2 is an $n-2 \times n-2$ matrix. Continuing in this way, there exists a unitary matrix, U given as the product of the U_i in the above construction such that

$$U^* A U = T$$

where T is some upper triangular matrix. Since the matrix is upper triangular, the characteristic equation is $\prod_{i=1}^{n} (\lambda - \lambda_i)$ where the λ_i are the diagonal entries of T. Therefore, the λ_i are the eigenvalues.

What if A is a real matrix and you only want to consider real unitary matrices?

Theorem 21.4.4 *Let A be a real $n \times n$ matrix. Then there exists a real unitary matrix, Q and a matrix T of the form*

$$T = \begin{pmatrix} P_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & P_r \end{pmatrix} \tag{21.9}$$

where P_i equals either a real 1×1 matrix or P_i equals a real 2×2 matrix having two complex eigenvalues of A such that $Q^T A Q = T$. The matrix, T is called a real Schur form of the matrix A.

Proof: Suppose

$$A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1, \ |\mathbf{v}_1| = 1$$

where λ_1 is real. Then let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be an orthonormal basis of vectors in $\mathbb{R}^n$. Let Q_0 be a matrix whose i^{th} column is $\mathbf{v}_i$. Then $Q_0^* A Q_0$ is of the form

$$\begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$$

where A_1 is a real $n-1 \times n-1$ matrix. This is just like the proof of Theorem 21.4.3 up to this point.

Now in case $\lambda_1 = \alpha + i\beta$, it follows since A is real that $\mathbf{v}_1 = \mathbf{z}_1 + i\mathbf{w}_1$ and that $\overline{\mathbf{v}}_1 = \mathbf{z}_1 - i\mathbf{w}_1$ is an eigenvector for the eigenvalue, $\alpha - i\beta$. Here $\mathbf{z}_1$ and $\mathbf{w}_1$ are real vectors. It is clear that $\{\mathbf{z}_1, \mathbf{w}_1\}$ is an independent set of vectors in $\mathbb{R}^n$. Indeed,$\{\mathbf{v}_1, \overline{\mathbf{v}}_1\}$ is an independent set and it follows $\text{span}(\mathbf{v}_1, \overline{\mathbf{v}}_1) = \text{span}(\mathbf{z}_1, \mathbf{w}_1)$. Now using the Gram Schmidt theorem in $\mathbb{R}^n$, there exists $\{\mathbf{u}_1, \mathbf{u}_2\}$, an orthonormal set of real vectors such that $\text{span}(\mathbf{u}_1, \mathbf{u}_2) = \text{span}(\mathbf{v}_1, \overline{\mathbf{v}}_1)$. Now let $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$ be an orthonormal basis in $\mathbb{R}^n$ and let Q_0 be a unitary matrix whose i^{th} column is $\mathbf{u}_i$. Then $A\mathbf{u}_j$ are both in $\text{span}(\mathbf{u}_1, \mathbf{u}_2)$ for $j = 1, 2$ and so $\mathbf{u}_k^T A \mathbf{u}_j = 0$ whenever $k \geq 3$. It follows that $Q_0^* A Q_0$ is of the form

$$\begin{pmatrix} * & * & \cdots & * \\ * & * & & \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$$

where A_1 is now an $n-2\times n-2$ matrix. In this case, find $\widetilde{Q}_1$ an $n-2\times n-2$ matrix to put A_1 in an appropriate form as above and come up with A_2 either an $n-4\times n-4$ matrix or an $n-3\times n-3$ matrix. Then the only other difference is to let

$$Q_1=\begin{pmatrix}1&0&0&\cdots&0\\0&1&0&\cdots&0\\0&0&&&\\\vdots&\vdots&&\widetilde{Q}_1&\\0&0&&&\end{pmatrix}$$

thus putting a 2×2 identity matrix in the upper left corner rather than a one. Repeating this process with the above modification for the case of a complex eigenvalue leads eventually to (21.9) where Q is the product of real unitary matrices Q_i above. Finally,

$$\lambda I-T=\begin{pmatrix}\lambda I_1-P_1&\cdots&*\\&\ddots&\vdots\\0&&\lambda I_r-P_r\end{pmatrix}$$

where I_k is the 2×2 identity matrix in the case that P_k is 2×2 and is the number 1 in the case where P_k is a 1×1 matrix. Now, it follows that $\det(\lambda I-T)=\prod_{k=1}^{r}\det(\lambda I_k-P_k)$. Therefore, λ is an eigenvalue of T if and only if it is an eigenvalue of some P_k. This proves the theorem since the eigenvalues of T are the same as those of A because they have the same characteristic polynomial due to the similarity of A and T.

Definition 21.4.5 *When a linear transformation, A, mapping a linear space, V to V has a basis of eigenvectors, the linear transformation is called non defective. Otherwise it is called defective. An $n\times n$ matrix, A, is called normal if $AA^*=A^*A$. An important class of normal matrices is that of the Hermitian or self adjoint matrices. An $n\times n$ matrix, A is self adjoint or Hermitian if $A=A^*$.*

The next lemma is the basis for concluding that every normal matrix is unitarily similar to a diagonal matrix.

Lemma 21.4.6 *If T is upper triangular and normal, then T is a diagonal matrix.*

Proof: Since T is normal, $T^*T=TT^*$. Writing this in terms of components and using the description of the adjoint as the transpose of the conjugate, yields the following for the ik^{th} entry of $T^*T=TT^*$.

$$\sum_j t_{ij}t^*_{jk}=\sum_j t_{ij}\overline{t_{kj}}=\sum_j t^*_{ij}t_{jk}=\sum_j \overline{t_{ji}}t_{jk}.$$

Now use the fact that T is upper triangular and let $i=k=1$ to obtain the following from the above.

$$\sum_j |t_{1j}|^2=\sum_j |t_{j1}|^2=|t_{11}|^2$$

You see, $t_{j1}=0$ unless $j=1$ due to the assumption that T is upper triangular. This shows T is of the form

$$\begin{pmatrix}*&0&\cdots&0\\0&*&\cdots&*\\\vdots&\ddots&\ddots&\vdots\\0&\cdots&0&*\end{pmatrix}.$$

Now do the same thing only this time take $i = k = 2$ and use the result just established. Thus, from the above,

$$\sum_j |t_{2j}|^2 = \sum_j |t_{j2}|^2 = |t_{22}|^2,$$

showing that $t_{2j} = 0$ if $j > 2$ which means T has the form

$$\begin{pmatrix} * & 0 & 0 & \cdots & 0 \\ 0 & * & 0 & \cdots & 0 \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & * \end{pmatrix}.$$

Next let $i = k = 3$ and obtain that T looks like a diagonal matrix in so far as the first 3 rows and columns are concerned. Continuing in this way it follows T is a diagonal matrix.

Theorem 21.4.7 *Let A be a normal matrix. Then there exists a unitary matrix, U such that U^*AU is a diagonal matrix.*

Proof: From Theorem 21.4.3 there exists a unitary matrix, U such that U^*AU equals an upper triangular matrix. The theorem is now proved if it is shown that the property of being normal is preserved under unitary similarity transformations. That is, verify that if A is normal and if $B = U^*AU$, then B is also normal. But this is easy.

$$\begin{aligned} B^*B &= U^*A^*UU^*AU = U^*A^*AU \\ &= U^*AA^*U = U^*AUU^*A^*U = BB^*. \end{aligned}$$

Therefore, U^*AU is a normal and upper triangular matrix and by Lemma 21.4.6 it must be a diagonal matrix. This proves the theorem.

Corollary 21.4.8 *If A is Hermitian, then all the eigenvalues of A are real and there exists an orthonormal basis of eigenvectors.*

Proof: Since A is normal, there exists unitary, U such that $U^*AU = D$, a diagonal matrix whose diagonal entries are the eigenvalues of A. Therefore, $D^* = U^*A^*U = U^*AU = D$ showing D is real.

Finally, let

$$U = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{pmatrix}$$

where the $\mathbf{u}_i$ denote the columns of U and

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

The equation, $U^*AU = D$ implies

$$\begin{aligned} AU &= \begin{pmatrix} A\mathbf{u}_1 & A\mathbf{u}_2 & \cdots & A\mathbf{u}_n \end{pmatrix} \\ &= UD = \begin{pmatrix} \lambda_1\mathbf{u}_1 & \lambda_2\mathbf{u}_2 & \cdots & \lambda_n\mathbf{u}_n \end{pmatrix} \end{aligned}$$

where the entries denote the columns of AU and UD respectively. Therefore, $A\mathbf{u}_i = \lambda_i\mathbf{u}_i$ and since the matrix is unitary, the ij^{th} entry of U^*U equals δ_{ij} and so

$$\delta_{ij} = \overline{\mathbf{u}}_i^T\mathbf{u}_j = \overline{\mathbf{u}_i^T\overline{\mathbf{u}}_j} = \overline{\mathbf{u}_i \cdot \mathbf{u}_j}.$$

This proves the corollary because it shows the vectors $\{\mathbf{u}_i\}$ form an orthonormal basis.

Corollary 21.4.9 *If A is a real symmetric matrix, then A is Hermitian and there exists a real unitary matrix, U such that $U^TAU = D$ where D is a diagonal matrix.*

Proof: This follows from Theorem 21.4.4 and Corollary 21.4.8. Alternatively, you could use Corollary 21.4.8 to assert the eigenvalues are all real. Then if $A\mathbf{x} = \lambda\mathbf{x}$ the same is true of $\overline{\mathbf{x}}$ and so in the construction for Shur's theorem, you can always deal exclusively with real eigenvectors as long as your matrices are real and symmetric. When you construct the matrix which reduces the problem to a smaller one having A_1 in the lower right corner, use the Gram Schmidt process on $\mathbb{R}^n$ using the real dot product to construct vectors, $\mathbf{v}_2, \cdots, \mathbf{v}_n$ in $\mathbb{R}^n$ such that $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is an orthonormal basis for $\mathbb{R}^n$. The matrix A_1 is symmetric also. This is because for $j, k \geq 2$

$$A_{1kj} = \mathbf{v}_k^T A\mathbf{v}_j = \left(\mathbf{v}_k^T A\mathbf{v}_j\right)^T = \mathbf{v}_j^T A\mathbf{v}_k = A_{1jk}.$$

Therefore, continuing this way, the process of the proof delivers only real vectors and real matrices.

21.5 Exercises

1. State the eigenvalue problem from an algebraic perspective.

2. State the eigenvalue problem from a geometric perspective.

3. Suppose T is a linear transformation and it satisfies $T^2 = T$ and $T\mathbf{x} = \mathbf{x}$ for all $\mathbf{x}$ in a certain subspace, V. Show that 1 is an eigenvalue for T and show that all eigenvalues have absolute values no larger than 1.

4. Is it possible for a nonzero matrix to have only 0 as an eigenvalue?

5. Show that if $A\mathbf{x} = \lambda\mathbf{x}$ and $A\mathbf{y} = \lambda\mathbf{y}$, then whenever a, b are scalars,

$$A\left(a\mathbf{x} + b\mathbf{y}\right) = \lambda\left(a\mathbf{x} + b\mathbf{y}\right).$$

   Does this imply that $a\mathbf{x} + b\mathbf{y}$ is an eigenvector? Explain.

6. Let M be an $n \times n$ matrix and suppose $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are n eigenvectors which form a linearly independent set. Form the matrix S by making the columns these vectors. Show that S^{-1} exists and that $S^{-1}MS$ is a diagonal matrix (one having zeros everywhere except on the main diagonal) having the eigenvalues of M on the main diagonal. When this can be done the matrix is **diagonalizable**.

7. Show that a matrix, M is diagonalizable if and only if it has a basis of eigenvectors. **Hint:**To show that if the matrix can be diagonalized by some matrix, S giving $D = S^{-1}MS$ for D a diagonal matrix, then it has a basis of eigenvectors, try using the columns of the matrix S.

8. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} -19 & -14 & -1 \\ 8 & 4 & 8 \\ 15 & 30 & -3 \end{pmatrix}.$$

9. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} -3 & -30 & 15 \\ 0 & 12 & 0 \\ 15 & 30 & -3 \end{pmatrix}.$$

10. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 8 & 4 & 5 \\ 0 & 12 & 9 \\ -2 & 2 & 10 \end{pmatrix}.$$

11. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 7 & -2 & 0 \\ 8 & -1 & 0 \\ -2 & 4 & 6 \end{pmatrix}.$$
Can you find three independent eigenvectors?

12. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 3 & -2 & -1 \\ 0 & 5 & 1 \\ 0 & 2 & 4 \end{pmatrix}.$$
Can you find three independent eigenvectors in this case?

13. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 12 & -12 & 6 \\ 0 & 18 & 0 \\ 6 & 12 & 12 \end{pmatrix}$$

14. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} -5 & -1 & 10 \\ -15 & 9 & -6 \\ 8 & -8 & 2 \end{pmatrix}.$$

15. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} -10 & -8 & 8 \\ -4 & -14 & -4 \\ 0 & 0 & -18 \end{pmatrix}.$$

16. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 1 & 26 & -17 \\ 4 & -4 & 4 \\ -9 & -18 & 9 \end{pmatrix}.$$

17. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 8 & 4 & 5 \\ 0 & 12 & 9 \\ -2 & 2 & 10 \end{pmatrix}.$$

18. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 9 & 6 & -3 \\ 0 & 6 & 0 \\ -3 & -6 & 9 \end{pmatrix}.$$

19. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} -10 & -2 & 11 \\ -18 & 6 & -9 \\ 10 & -10 & -2 \end{pmatrix}.$$

20. Find the complex eigenvalues and eigenvectors of the matrix $\begin{pmatrix} 4 & -2 & -2 \\ 0 & 2 & -2 \\ 2 & 0 & 2 \end{pmatrix}$.

21. Let T be the linear transformation which reflects vectors about the x axis. Find a matrix for T and then find its eigenvalues and eigenvectors.

22. Let T be the linear transformation which rotates all vectors in $\mathbb{R}^2$ counterclockwise through an angle of $\pi/2$. Find a matrix of T and then find eigenvalues and eigenvectors.

23. Let T be the linear transformation which reflects all vectors in $\mathbb{R}^3$ through the xy plane. Find a matrix for T and then obtain its eigenvalues and eigenvectors.

24. Here are three vectors in $\mathbb{R}^4 : (1,2,0,3)^T, (2,1,-3,2)^T, (0,0,1,2)^T$. Find the volume of the parallelepiped determined by these three vectors.

25. Here are two vectors in $\mathbb{R}^4 : (1,2,0,3)^T, (2,1,-3,2)^T$. Find the volume of the parallelepiped determined by these two vectors.

26. Here are three vectors in $\mathbb{R}^2 : (1,2)^T, (2,1)^T, (0,1)^T$. Find the volume of the parallelepiped determined by these three vectors. Why should this volume equal zero?

27. If there are $n+1$ or more vectors in $\mathbb{R}^n$, Lemma 21.2.3 implies the parallelepiped determined by these $n+1$ vectors must have zero volume. What is the geometric significance of this assertion?

28. Find the equation of the plane through the three points $(1,2,3), (2,-3,1), (1,1,7)$.

29. Let
$$A = \begin{pmatrix} \boxed{\begin{matrix} 1 & 2 \\ 3 & 4 \end{matrix}} & \boxed{\begin{matrix} 2 \\ 0 \end{matrix}} \\ \boxed{\begin{matrix} 0 & 1 \end{matrix}} & \boxed{3} \end{pmatrix}$$
and let
$$B = \begin{pmatrix} \boxed{\begin{matrix} 0 & 1 \\ 1 & 1 \end{matrix}} \\ \boxed{\begin{matrix} 2 & 1 \end{matrix}} \end{pmatrix}$$
Multiply AB verifying the block multiplication formula. Here $A_{11} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $A_{12} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$, $A_{21} = \begin{pmatrix} 0 & 1 \end{pmatrix}$ and $A_{22} = (3)$.

30. Let A be an $r \times r$ matrix and let B be an $m \times m$ matrix such that $r+m=n$. Consider the following $n \times n$ block matrix

$$C = \begin{pmatrix} A & 0 \\ D & B \end{pmatrix}.$$

where the D is an $m \times r$ matrix, and the 0 is a $r \times m$ matrix. Letting I_k denote the $k \times k$ identity matrix, tell why

$$C = \begin{pmatrix} A & 0 \\ D & I_m \end{pmatrix} \begin{pmatrix} I_r & 0 \\ 0 & B \end{pmatrix}.$$

Now explain why $\det(C) = \det(A)\det(B)$. **Hint:** Part of this will require an explanation of why

$$\det \begin{pmatrix} A & 0 \\ D & I_m \end{pmatrix} = \det(A).$$

See Theorems 15.1.23 - 15.1.25.

31. If A is a real $n \times n$ matrix which has all real eigenvalues, show there exists a real unitary matrix, U such that $U^T A U = T$ where T is a real upper triangular matrix. If A is normal, explain why T is a diagonal matrix.

32. If A is an $n \times n$ Hermitian matrix, show there exists an orthonormal basis, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ such that

$$A = \sum_{j=1}^{n} \lambda_j \mathbf{v}_j^T \mathbf{v}_j.$$

If A is real and Hermitian, show that the vectors, $\mathbf{v}_j$ may all be taken to be real vectors.

Planes And Surfaces In $\mathbb{R}^n$

22.0.1 Outcomes

1. Find the angle between two lines.

2. Determine a point of intersection between a line and a surface.

3. Find the equation of a plane in 3 space given a point and a normal vector, three points, a sketch of a plane or a geometric description of the plane.

4. Determine the normal vector and the intercepts of a given plane.

5. Sketch the graph of a plane given its equation.

6. Determine the cosine of the angle between two planes.

7. Find the equation of a plane determined by lines.

8. Identify standard quadric surfaces given their functions or graphs.

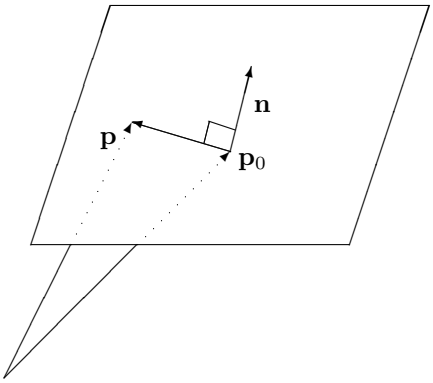
9. Sketch the graph of a quadric surface by identifying the intercepts, traces, sections, symmetry and boundedness or unboundedness of the surface.

22.1 Planes

You have an idea of what a plane is already. It is the span of some vectors. However, it can also be considered geometrically in terms of a dot product. To find the equation of a plane, you need two things, a point contained in the plane and a vector normal to the plane. Let $\mathbf{p}_0 = (x_0, y_0, z_0)$ denote the position vector of a point in the plane, let $\mathbf{p} = (x, y, z)$ be the position vector of an arbitrary point in the plane, and let $\mathbf{n}$ denote a vector normal to the plane. This means that

$$\mathbf{n} \cdot (\mathbf{p} - \mathbf{p}_0) = 0$$

whenever $\mathbf{p}$ is the position vector of a point in the plane. The following picture illustrates the geometry of this idea.



Expressed equivalently, the plane is just the set of all points $\mathbf{p}$ such that the vector, $\mathbf{p}-\mathbf{p}_0$ is perpendicular to the given normal vector, $\mathbf{n}$.

Example 22.1.1 *Find the equation of the plane with normal vector,* $\mathbf{n}=(1,2,3)$ *containing the point* $(2,-1,5)$.

From the above, the equation of this plane is just

$$(1,2,3)\cdot(x-2,y+1,z-3)=x-9+2y+3z=0$$

Example 22.1.2 $2x+4y-5z=11$ *is the equation of a plane. Find the normal vector and a point on this plane.*

You can write this in the form $2\left(x-\frac{11}{2}\right)+4(y-0)+(-5)(z-0)=0$. Therefore, a normal vector to the plane is $2\mathbf{i}+4\mathbf{j}-5\mathbf{k}$ and a point in this plane is $\left(\frac{11}{2},0,0\right)$. Of course there are many other points in the plane.

Definition 22.1.3 *Suppose two planes intersect. The angle between the planes is defined to be the angle between their normal vectors.*

Example 22.1.4 *Find the equation of the plane which contains the three points,* $(1,2,1),(3,-1,2),$ *and* $(4,2,1)$.

You just need to get a normal vector to this plane. This can be done by taking the cross products of the two vectors,

$$(3,-1,2)-(1,2,1) \text{ and } (4,2,1)-(1,2,1)$$

Thus a normal vector is $(2,-3,1)\times(3,0,0)=(0,3,9)$. Therefore, the equation of the plane is

$$0(x-1)+3(y-2)+9(z-1)=0$$

or $3y+9z=15$ which is the same as $y+3z=5$.

Example 22.1.5 *Find the equation of the plane which contains the three points,* $(1,2,1),(3,-1,2),$ *and* $(4,2,1)$ *another way.*

Letting (x,y,z) be a point on the plane, the volume of the parallelepiped spanned by $(x,y,z)-(1,2,1)$ and the two vectors, $(2,-3,1)$ and $(3,0,0)$ must be equal to zero. Thus the equation of the plane is

$$\det\begin{pmatrix}3&0&0\\2&-3&1\\x-1&y-2&z-1\end{pmatrix}=0.$$

Hence $-9z+15-3y=0$ and dividing by 3 yields the same answer as the above.

Proposition 22.1.6 *If* $(a,b,c) \neq (0,0,0)$, *then* $ax+by+cz=d$ *is the equation of a plane with normal vector* $a\mathbf{i}+b\mathbf{j}+c\mathbf{k}$. *Conversely, any plane can be written in this form.*

Proof: One of a,b,c is nonzero. Suppose for example that $c \neq 0$. Then the equation can be written as

$$a(x-0)+b(y-0)+c\left(z-\frac{d}{c}\right)=0$$

Therefore, $\left(0,0,\frac{d}{c}\right)$ is a point on the plane and a normal vector is $a\mathbf{i}+b\mathbf{j}+c\mathbf{k}$. The converse follows from the above discussion involving the point and a normal vector. This proves the proposition.

Example 22.1.7 *Find the equation of the plane which contains the three points,* $(1,2,1)$, $(3,-1,2)$, *and* $(4,2,1)$ *another way.*

You need to find numbers, a,b,c,d not all zero such that each of the given three points satisfies the equation,

$$ax+by+cz=d.$$

Then you must have for (x,y,z) a point on this plane,

$$\begin{gathered} a+2b+c-d=0, \\ 3a-b+2c-d=0, \\ 4a+2b+c-d=0, \\ xa+yb+zc-d=0. \end{gathered}$$

You need a nonzero solution to the above system of four equations for the unknowns, a,b,c, and d. Therefore,

$$\det\begin{pmatrix} 1 & 2 & 1 & -1 \\ 3 & -1 & 2 & -1 \\ 4 & 2 & 1 & -1 \\ x & y & z & -1 \end{pmatrix}=0$$

because the matrix sends a nonzero vector, $(a,b,c,-d)$ to zero and is therefore, not one to one. Consequently from Theorem 15.2.1 on Page 393, its determinant equals zero. Hence upon evaluating the determinant,

$$-15+9z+3y=0$$

which reduces to $3z+y=5$.

Example 22.1.8 *Find the equation of the plane containing the points* $(1,2,3)$ *and the line* $(0,1,1)+t(2,1,2)=(x,y,z)$.

There are several ways to do this. One is to find three points and use any of the above procedures. Let $t=0$ and then let $t=1$ to get two points on the line. This yields $(1,2,3)$, $(0,1,1)$, and $(2,2,3)$. Then the equation of the plane is

$$\det\begin{pmatrix} x & y & z & -1 \\ 1 & 2 & 3 & -1 \\ 0 & 1 & 1 & -1 \\ 2 & 2 & 3 & -1 \end{pmatrix}=2y-z-1=0.$$

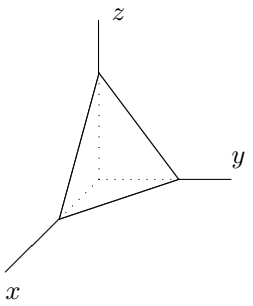
Example 22.1.9 *Find the equation of the plane which contains the two lines, given by the following parametric expressions in which* $t \in \mathbb{R}$.

$$(2t,1+t,1+2t)=(x,y,z),\ (2t+2,1,3+2t)=(x,y,z)$$

Note first that you don't know there even is such a plane. However, if there is, you could find it by obtaining three points, two on one line and one on another and then using any of the above procedures for finding the plane. From the first line, two points are $(0,1,1)$ and $(2,2,3)$ while a third point can be obtained from second line, $(2,1,3)$. You need a normal vector and then use any of these points. To get a normal vector, form $(2,0,2)\times(2,1,2) = (-2,0,2)$. Therefore, the plane is $-2x+0\,(y-1)+2\,(z-1)=0$. This reduces to $z-x=1$. If there is a plane, this is it. Now you can simply verify that both of the lines are really in this plane. From the first, $(1+2t)-2t=1$ and the second, $(3+2t)-(2t+2)=1$ so both lines lie in the plane.

One way to understand how a plane looks is to connect the points where it intercepts the x, y, and z axes. This allows you to visualize the plane somewhat and is a good way to sketch the plane. Not surprisingly these points are called intercepts.

Example 22.1.10 *Sketch the plane which has intercepts* $(2,0,0)\,,(0,3,0)\,,$ *and* $(0,0,4)\,.$



You see how connecting the intercepts gives a fairly good geometric description of the plane. These lines which connect the intercepts are also called the traces of the plane. Thus the line which joins $(0,3,0)$ to $(0,0,4)$ is the intersection of the plane with the yz plane. It is the trace on the yz plane.

Example 22.1.11 *Identify the intercepts of the plane,* $3x-4y+5z=11$.

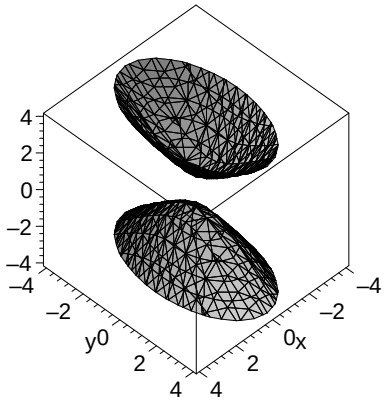
The easy way to do this is to divide both sides by 11.

$$\frac{x}{(11/3)}+\frac{y}{(-11/4)}+\frac{z}{(11/5)}=1$$

The intercepts are $(11/3,0,0)\,,(0,-11/4,0)$ and $(0,0,11/5)\,.$ You can see this by letting both y and z equal to zero to find the point on the x axis which is intersected by the plane. The other axes are handled similarly.

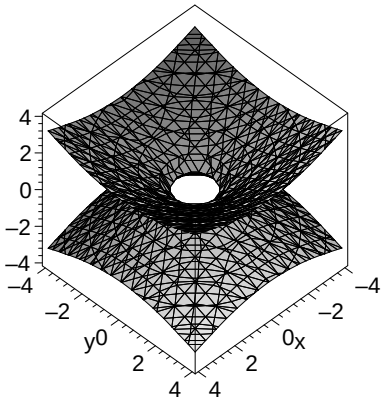
22.2 Quadric Surfaces

In the above it was shown that the equation of an arbitrary plane is an equation of the form $ax+by+cz=d$. Such equations are called level surfaces. There are some standard level surfaces which involve certain variables being raised to a power of 2 which are sufficiently important that they are given names, usually involving the portentous semi-word "oid". These are graphed below using Maple, a computer algebra system.



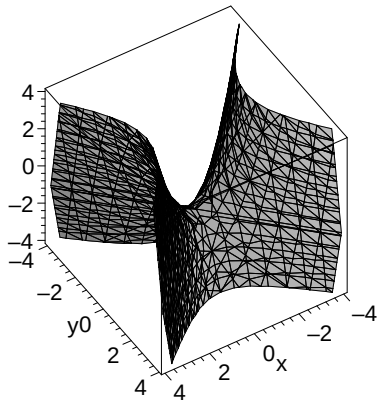
$z^2/a^2 - x^2/b^2 - y^2/c^2 = 1$

hyperboloid of two sheets



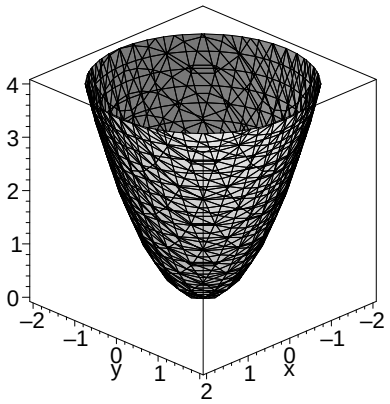
$x^2/b^2 + y^2/c^2 - z^2/a^2 = 1$

hyperboloid of one sheet



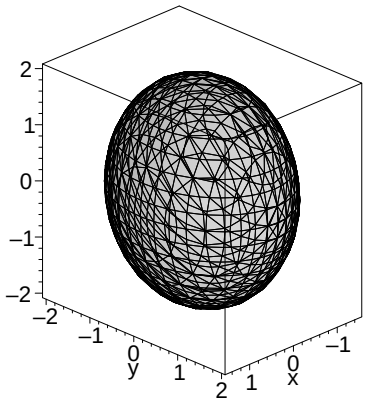
$z = x^2/a^2 - y^2/b^2$

hyperbolic paraboloid



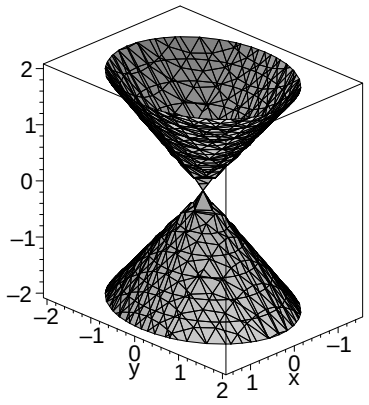
$x^2/b^2 + y^2/c^2 = z$

elliptic paraboloid



$x^2/a^2 + y^2/b^2 + z^2/c^2 = 1$

ellipsoid



$x^2/b^2 + y^2/c^2 = z^2/a^2$

elliptic cone

Why do the graphs of these level surfaces look the way they do? Consider first the

hyperboloid of two sheets. The equation defining this surface can be written in the form

$$\frac{z^2}{a^2} - 1 = \frac{x^2}{b^2} + \frac{y^2}{c^2}.$$

Suppose you fix a value for z. What ordered pairs, (x, y) will satisfy the equation? If $\frac{z^2}{a^2} < 1$, there is no such ordered pair because the above equation would require a negative number to equal a nonnegative one. This is why there is a gap and there are two sheets. If $\frac{z^2}{a^2} > 1$, then the above equation is the equation for an ellipse. That is why if you slice the graph by letting $z = z_0$ the result is an ellipse in the plane $z = z_0$.

Consider the hyperboloid of one sheet.

$$\frac{x^2}{b^2} + \frac{y^2}{c^2} = 1 + \frac{z^2}{a^2}.$$

This time, it doesn't matter what value z takes. The resulting equation for (x, y) is an ellipse.

Similar considerations apply to the elliptic paraboloid as long as $z > 0$ and the ellipsoid. The elliptic cone is like the hyperboloid of two sheets without the 1. Therefore, z can have any value. In case $z = 0$, $(x, y) = (0, 0)$. Viewed from the side, it appears straight, not curved like the hyperboloid of two sheets.This is because if (x, y, z) is a point on the surface, then if t is a scalar, it follows (tx, ty, tz) is also on this surface.

The most interesting of these graphs is the hyperbolic paraboloid[1], $z = \frac{x^2}{a^2} - \frac{y^2}{b^2}$. If $z > 0$ this is the equation of a hyperbola which opens to the right and left while if $z < 0$ it is a hyperbola which opens up and down. As z passes from positive to negative, the hyperbola changes type and this is what yields the shape shown in the picture.

Not surprisingly, you can find intercepts and traces of quadric surfaces just as with planes.

Example 22.2.1 *Find the trace on the xy plane of the hyperbolic paraboloid, $z = x^2 - y^2$.*

This occurs when $z = 0$ and so this reduces to $y^2 = x^2$. In other words, this trace is just the two straight lines, $y = x$ and $y = -x$.

Example 22.2.2 *Find the intercepts of the ellipsoid, $x^2 + 2y^2 + 4z^2 = 9$.*

To find the intercept on the x axis, let $y = z = 0$ and this yields $x = \pm 3$. Thus there are two intercepts, $(3, 0, 0)$ and $(-3, 0, 0)$. The other intercepts are left for you to find. You can see this is an aid in graphing the quadric surface. The surface is said to be bounded if there is some number, C such that whenever, (x, y, z) is a point on the surface, $\sqrt{x^2 + y^2 + z^2} < C$. The surface is called unbounded if no such constant, C exists. Ellipsoids are bounded but the other quadric surfaces are not bounded.

Example 22.2.3 *Why is the hyperboloid of one sheet, $x^2 + 2y^2 - z^2 = 1$ unbounded?*

Let z be very large. Does there correspond (x, y) such that (x, y, z) is a point on the hyperboloid of one sheet? Certainly. Simply pick any (x, y) on the ellipse $x^2 + 2y^2 = 1 + z^2$. Then $\sqrt{x^2 + y^2 + z^2}$ is large, at lest as large as z. Thus it is unbounded.

You can also find intersections between lines and surfaces.

Example 22.2.4 *Find the points of intersection of the line $(x, y, z) = (1 + t, 1 + 2t, 1 + t)$ with the surface, $z = x^2 + y^2$.*

[1]It is traditional to refer to this as a hyperbolic paraboloid. Not a parabolic hyperboloid.

First of all, there is no guarantee there is any intersection at all. But if it exists, you have only to solve the equation for t

$$1+t=(1+t)^2+(1+2t)^2$$

This occurs at the two values of $t=-\frac{1}{2}+\frac{1}{10}\sqrt{5}, t=-\frac{1}{2}-\frac{1}{10}\sqrt{5}$. Therefore, the two points are

$$(1,1,1)+\left(-\frac{1}{2}+\frac{1}{10}\sqrt{5}\right)(1,2,1), \text{ and } (1,1,1)+\left(-\frac{1}{2}-\frac{1}{10}\sqrt{5}\right)(1,2,1)$$

That is

$$\left(\frac{1}{2}+\frac{1}{10}\sqrt{5},\frac{1}{5}\sqrt{5},\frac{1}{2}+\frac{1}{10}\sqrt{5}\right),\left(\frac{1}{2}-\frac{1}{10}\sqrt{5},-\frac{1}{5}\sqrt{5},\frac{1}{2}-\frac{1}{10}\sqrt{5}\right).$$

22.3 Exercises

1. Determine whether the lines $(1,1,2)+t(1,0,3)$ and $(4,1,3)+t(3,0,1)$ have a point of intersection. If they do, find the cosine of the angle between the two lines. If they do not intersect, explain why they do not.
2. Determine whether the lines $(1,1,2)+t(1,0,3)$ and $(4,2,3)+t(3,0,1)$ have a point of intersection. If they do, find the cosine of the angle between the two lines. If they do not intersect, explain why they do not.
3. Find where the line $(1,0,1)+t(1,2,1)$ intersects the surface $x^2+y^2+z^2=9$ if possible. If there is no intersection, explain why.
4. Find a parametric equation for the line through the points $(2,3,4,5)$ and $(-2,3,0,1)$.
5. Find the equation of a line through $(1,2,3,0)$ which has direction vector, $(2,1,3,1)$.
6. Let $(x,y)=(2\cos(t),2\sin(t))$ where $t\in[0,2\pi]$. Describe the set of points encountered as t changes.
7. Let $(x,y,z)=(2\cos(t),2\sin(t),t)$ where $t\in\mathbb{R}$. Describe the set of points encountered as t changes.
8. If there is a plane which contains the two lines, $(2t+2,1+t,3+2t)=(x,y,z)$ and $(4+t,3+2t,4+t)=(x,y,z)$ find it. If there is no such plane tell why.
9. If there is a plane which contains the two lines, $(2t+4,1+t,3+2t)=(x,y,z)$ and $(4+t,3+2t,4+t)=(x,y,z)$ find it. If there is no such plane tell why.
10. Find the equation of the plane which contains the three points $(1,-2,3),(2,3,4)$, and $(3,1,2)$.
11. Find the equation of the plane which contains the three points $(1,2,3),(2,0,4)$, and $(3,1,2)$.
12. Find the equation of the plane which contains the three points $(0,2,3),(2,3,4)$, and $(3,5,2)$.
13. Find the equation of the plane which contains the three points $(1,2,3),(0,3,4)$, and $(3,6,2)$.
14. Find the equation of the plane having a normal vector, $5\mathbf{i}+2\mathbf{j}-6\mathbf{k}$ which contains the point $(2,1,3)$.

15. Find the equation of the plane having a normal vector, $\mathbf{i}+2\mathbf{j}-4\mathbf{k}$ which contains the point $(2,0,1)$.

16. Find the equation of the plane having a normal vector, $2\mathbf{i}+\mathbf{j}-6\mathbf{k}$ which contains the point $(1,1,2)$.

17. Find the equation of the plane having a normal vector, $\mathbf{i}+2\mathbf{j}-3\mathbf{k}$ which contains the point $(1,0,3)$.

18. Find the cosine of the angle between the two planes $2x+3y-z=11$ and $3x+y+2z=9$.

19. Find the cosine of the angle between the two planes $x+3y-z=11$ and $2x+y+2z=9$.

20. Find the cosine of the angle between the two planes $2x+y-z=11$ and $3x+5y+2z=9$.

21. Find the cosine of the angle between the two planes $x+3y+z=11$ and $3x+2y+2z=9$.

22. Determine the intercepts and sketch the plane $3x-2y+z=4$.

23. Determine the intercepts and sketch the plane $x-2y+z=2$.

24. Determine the intercepts and sketch the plane $x+y+z=3$.

25. Based on an analogy with the above pictures, sketch or otherwise describe the graph of $y=\frac{x^2}{a^2}-\frac{z^2}{b^2}$.

26. Based on an analogy with the above pictures, sketch or otherwise describe the graph of $\frac{z^2}{b^2}+\frac{y^2}{c^2}=1+\frac{x^2}{a^2}$.

27. The equation of a cone is $z^2=x^2+y^2$. Suppose this cone is intersected with the plane, $z=ay+1$. Consider the projection of the intersection of the cone with this plane. Show this sometimes results in a parabola, sometimes a hyperbola, and sometimes an ellipse depending on a.

28. Find the intercepts of the quadric surface, $x^2+4y^2-z^2=4$ and sketch the surface.

29. Find the intercepts of the quadric surface, $x^2-\left(4y^2+z^2\right)=4$ and sketch the surface.

30. Find the intersection of the line $(x,y,z)=(1+t,t,3t)$ with the surface, $x^2/9+y^2/4+z^2/16=1$ if possible.

Part V

Vector Calculus

Vector Valued Functions

23.0.1 Outcomes

1. Identify the domain of a vector function.
2. Represent combinations of multivariable functions algebraically.
3. Evaluate the limit of a function of several variables or show that it does not exist.
4. Determine whether a function is continuous at a given point. Give examples of continuous functions.
5. Recall and apply the extreme value theorem.

23.1 Vector Valued Functions

Vector valued functions have values in $\mathbb{R}^p$ where p is an integer at least as large as 1. Here is a simple example which is obviously of interest.

Example 23.1.1 *A rocket is launched from the rotating earth. You could define a function having values in* $\mathbb{R}^3$ *as* $(r(t), \theta(t), \phi(t))$ *where* $r(t)$ *is the distance of the center of mass of the rocket from the center of the earth,* $\theta(t)$ *is the longitude, and* $\phi(t)$ *is the latitude of the rocket.*

Example 23.1.2 *Let* $\mathbf{f}(x,y) = \left(\sin xy, y^3 + x, x^4\right)$. *Then* $\mathbf{f}$ *is a function defined on* $\mathbb{R}^2$ *which has values in* $\mathbb{R}^3$. *For example,* $\mathbf{f}(1,2) = (\sin 2, 9, 16)$.

As usual, $D(\mathbf{f})$ denotes the domain of the function, $\mathbf{f}$ which is written in bold face because it will possibly have values in $\mathbb{R}^p$. When $D(\mathbf{f})$ is not specified, it will be understood that the domain of $\mathbf{f}$ consists of those things for which $\mathbf{f}$ makes sense.

Example 23.1.3 *Let* $\mathbf{f}(x,y,z) = \left(\frac{x+y}{z}, \sqrt{1-x^2}, y\right)$. *Then* $D(\mathbf{f})$ *would consist of the set of all* (x,y,z) *such that* $|x| \leq 1$ *and* $z \neq 0$.

There are many ways to make new functions from old ones.

Definition 23.1.4 *Let* $\mathbf{f}, \mathbf{g}$ *be functions with values in* $\mathbb{R}^p$. *Let* a, b *be elements of* $\mathbb{R}$ *(scalars). Then* $a\mathbf{f} + b\mathbf{g}$ *is the name of a function whose domain is* $D(\mathbf{f}) \cap D(\mathbf{g})$ *which is defined as*

$$(a\mathbf{f} + b\mathbf{g})(\mathbf{x}) = a\mathbf{f}(\mathbf{x}) + b\mathbf{g}(\mathbf{x}).$$

$\mathbf{f}\cdot\mathbf{g}$ *or* $(\mathbf{f},\mathbf{g})$ *is the name of a function whose domain is* $D(\mathbf{f})\cap D(\mathbf{g})$ *which is defined as*

$$(\mathbf{f},\mathbf{g})(\mathbf{x})\equiv\mathbf{f}\cdot\mathbf{g}(\mathbf{x})\equiv\mathbf{f}(\mathbf{x})\cdot\mathbf{g}(\mathbf{x}).$$

If $\mathbf{f}$ *and* $\mathbf{g}$ *have values in* $\mathbb{R}^3$*, define a new function,* $\mathbf{f}\times\mathbf{g}$ *by*

$$\mathbf{f}\times\mathbf{g}(t)\equiv\mathbf{f}(t)\times\mathbf{g}(t).$$

If $\mathbf{f}:D(\mathbf{f})\to X$ *and* $\mathbf{g}:X\to Y$*, then* $\mathbf{g}\circ\mathbf{f}$ *is the name of a function whose domain is*

$$\{\mathbf{x}\in D(\mathbf{f}):\mathbf{f}(\mathbf{x})\in D(\mathbf{g})\}$$

which is defined as

$$\mathbf{g}\circ\mathbf{f}(\mathbf{x})\equiv\mathbf{g}(\mathbf{f}(\mathbf{x})).$$

This is called the composition of the two functions.

You should note that $\mathbf{f}(\mathbf{x})$ is not a function. It is the value of the function at the point, $\mathbf{x}$. The name of the function is $\mathbf{f}$. Nevertheless, people often write $\mathbf{f}(\mathbf{x})$ to denote a function and it doesn't cause too many problems in beginning courses. When this is done, the variable, $\mathbf{x}$ should be considered as a generic variable free to be anything in $D(\mathbf{f})$. I will use this slightly sloppy abuse of notation whenever convenient.

Example 23.1.5 *Let* $\mathbf{f}(t)\equiv(t,1+t,2)$ *and* $\mathbf{g}(t)\equiv\left(t^2,t,t\right)$*. Then* $\mathbf{f}\cdot\mathbf{g}$ *is the name of the function satisfying*

$$\mathbf{f}\cdot\mathbf{g}(t)=\mathbf{f}(t)\cdot\mathbf{g}(t)=t^3+t+t^2+2t=t^3+t^2+3t$$

Note that in this case is was assumed the domains of the functions consisted of all of $\mathbb{R}$ because this was the set on which the two both made sense. Also note that $\mathbf{f}$ and $\mathbf{g}$ map $\mathbb{R}$ into $\mathbb{R}^3$ but $\mathbf{f}\cdot\mathbf{g}$ maps $\mathbb{R}$ into $\mathbb{R}$.

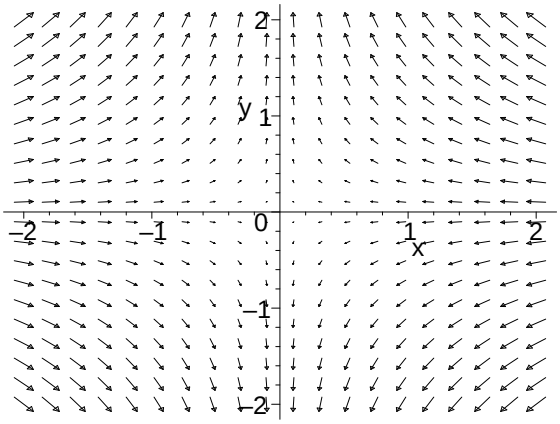
Example 23.1.6 *Suppose* $\mathbf{f}(t)=\left(2t,1+t^2\right)$ *and* $g{:}\mathbb{R}^2\to\mathbb{R}$ *is given by* $g(x,y)\equiv x+y$*. Then* $g\circ\mathbf{f}:\mathbb{R}\to\mathbb{R}$ *and*

$$g\circ\mathbf{f}(t)=g(\mathbf{f}(t))=g\left(2t,1+t^2\right)=1+2t+t^2.$$

23.2 Vector Fields

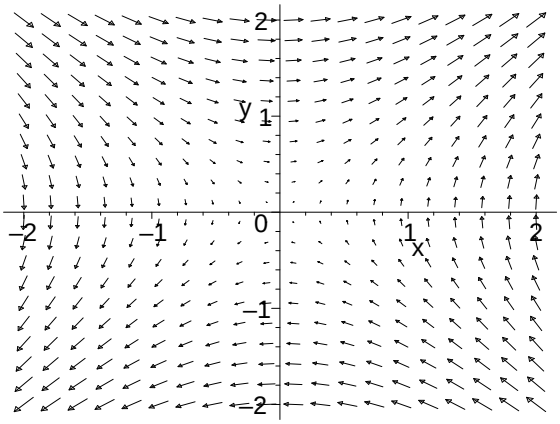
Some people find it useful to try and draw pictures to illustrate a vector valued function. This can be a very useful idea in the case where the function takes points in $D\subseteq\mathbb{R}^2$ and delivers a vector in $\mathbb{R}^2$. For many points, $(x,y)\in D$, you draw an arrow of the appropriate length and direction with its tail at (x,y). The picture of all these arrows can give you an understanding of what is happening. For example if the vector valued function gives the velocity of a fluid at the point, (x,y), the picture of these arrows can give an idea of the motion of the fluid. When they are long the fluid is moving fast, when they are short, the fluid is moving slowly the direction of these arrows is an indication of the direction of motion. The only sensible way to produce such a picture is with a computer. Otherwise, it becomes a worthless exercise in busy work. Furthermore, it is of limited usefulness in three dimensions because in three dimensions such pictures are too cluttered to convey much insight.

Example 23.2.1 *Draw a picture of the vector field,* $(-x,y)$ *which gives the velocity of a fluid flowing in two dimensions.*



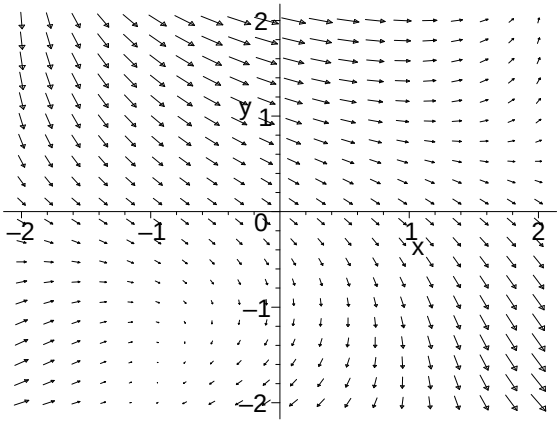
In this example, drawn by Maple, you can see how the arrows indicate the motion of this fluid.

Example 23.2.2 *Draw a picture of the vector field (y, x) for the velocity of a fluid flowing in two dimensions.*



So much for art. Get the computer to do it and it can be useful. If you try to do it, you will mainly waste time.

Example 23.2.3 *Draw a picture of the vector field $(y\cos(x)+1, x\sin(y)-1)$ for the velocity of a fluid flowing in two dimensions.*



23.3 Continuous Functions

What was done in begining calculus for scalar functions is generalized here to include the case of a vector valued function.

Definition 23.3.1 *A function* $\mathbf{f} : D(\mathbf{f}) \subseteq \mathbb{R}^p \to \mathbb{R}^q$ *is continuous at* $\mathbf{x} \in D(\mathbf{f})$ *if for each* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that whenever* $\mathbf{y} \in D(\mathbf{f})$ *and*

$$|\mathbf{y} - \mathbf{x}| < \delta$$

it follows that

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \varepsilon.$$

$\mathbf{f}$ *is continuous if it is continuous at every point of* $D(\mathbf{f})$.

Note the total similarity to the scalar valued case.

23.3.1 Sufficient Conditions For Continuity

The next theorem is a fundamental result which will allow us to worry less about the ε δ definition of continuity.

Theorem 23.3.2 *The following assertions are valid.*

1. *The function,* $a\mathbf{f} + b\mathbf{g}$ *is continuous at* $\mathbf{x}$ *whenever* $\mathbf{f}$, $\mathbf{g}$ *are continuous at* $\mathbf{x} \in D(\mathbf{f}) \cap D(\mathbf{g})$ *and* $a, b \in \mathbb{R}$.
2. *If* $\mathbf{f}$ *is continuous at* $\mathbf{x}$, $\mathbf{f}(\mathbf{x}) \in D(\mathbf{g}) \subseteq \mathbb{R}^p$, *and* $\mathbf{g}$ *is continuous at* $\mathbf{f}(\mathbf{x})$, *then* $\mathbf{g} \circ \mathbf{f}$ *is continuous at* $\mathbf{x}$.
3. *If* $\mathbf{f} = (f_1, \cdots, f_q) : D(\mathbf{f}) \to \mathbb{R}^q$, *then* $\mathbf{f}$ *is continuous if and only if each* f_k *is a continuous real valued function.*
4. *The function* $f : \mathbb{R}^p \to \mathbb{R}$, *given by* $f(\mathbf{x}) = |\mathbf{x}|$ *is continuous.*

The proof of this theorem is in the last section of this chapter. Its conclusions are not surprising. For example the first claim says that $(a\mathbf{f} + b\mathbf{g})(\mathbf{y})$ is close to $(a\mathbf{f} + b\mathbf{g})(\mathbf{x})$ when $\mathbf{y}$ is close to $\mathbf{x}$ provided the same can be said about $\mathbf{f}$ and $\mathbf{g}$. For the second claim, if $\mathbf{y}$ is close to $\mathbf{x}$, $\mathbf{f}(\mathbf{x})$ is close to $\mathbf{f}(\mathbf{y})$ and so by continuity of $\mathbf{g}$ at $\mathbf{f}(\mathbf{x})$, $\mathbf{g}(\mathbf{f}(\mathbf{y}))$ is close to $\mathbf{g}(\mathbf{f}(\mathbf{x}))$. To see the third claim is likely, note that closeness in $\mathbb{R}^p$ is the same as closeness in each coordinate. The fourth claim is immediate from the triangle inequality.

For functions defined on $\mathbb{R}^n$, there is a notion of polynomial just as there is for functions defined on $\mathbb{R}$.

Definition 23.3.3 *Let* α *be an* n *dimensional multi-index. This means*

$$\alpha = (\alpha_1, \cdots, \alpha_n)$$

where each α_i *is a natural number or zero. Also, let*

$$|\alpha| \equiv \sum_{i=1}^{n} |\alpha_i|$$

The symbol, $\mathbf{x}^\alpha$ *means*

$$\mathbf{x}^\alpha \equiv x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_3^{\alpha_n}.$$

An n dimensional polynomial of degree m is a function of the form

$$p(\mathbf{x}) = \sum_{|\alpha| \leq m} d_\alpha \mathbf{x}^\alpha.$$

where the d_α are real numbers.

The above theorem implies that polynomials are all continuous.

23.4 Limits Of A Function

As in the case of scalar valued functions of one variable, a concept closely related to continuity is that of the limit of a function. The notion of limit of a function makes sense at points, $\mathbf{x}$, which are limit points of $D(\mathbf{f})$ and this concept is defined next.

Definition 23.4.1 *Let $A \subseteq \mathbb{R}^m$ be a set. A point, $\mathbf{x}$, is a limit point of A if $B(\mathbf{x}, r)$ contains infinitely many points of A for every $r > 0$.*

Definition 23.4.2 *Let $\mathbf{f} : D(\mathbf{f}) \subseteq \mathbb{R}^p \to \mathbb{R}^q$ be a function and let $\mathbf{x}$ be a limit point of $D(\mathbf{f})$. Then*

$$\lim_{\mathbf{y} \to \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$$

if and only if the following condition holds. For all $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$0 < |\mathbf{y} - \mathbf{x}| < \delta, \text{ and } \mathbf{y} \in D(\mathbf{f})$$

then,

$$|\mathbf{L} - \mathbf{f}(\mathbf{y})| < \varepsilon.$$

Theorem 23.4.3 *If $\lim_{\mathbf{y} \to \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$ and $\lim_{y \to x} \mathbf{f}(\mathbf{y}) = \mathbf{L}_1$, then $\mathbf{L} = \mathbf{L}_1$.*

Proof: Let $\varepsilon > 0$ be given. There exists $\delta > 0$ such that if $0 < |\mathbf{y} - \mathbf{x}| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, then

$$|\mathbf{f}(\mathbf{y}) - \mathbf{L}| < \varepsilon, \ |\mathbf{f}(\mathbf{y}) - \mathbf{L}_1| < \varepsilon.$$

Pick such a $\mathbf{y}$. There exists one because $\mathbf{x}$ is a limit point of $D(\mathbf{f})$. Then

$$|\mathbf{L} - \mathbf{L}_1| \leq |\mathbf{L} - \mathbf{f}(\mathbf{y})| + |\mathbf{f}(\mathbf{y}) - \mathbf{L}_1| < \varepsilon + \varepsilon = 2\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows $\mathbf{L} = \mathbf{L}_1$.

As in the case of functions of one variable, one can define what it means for $\lim_{\mathbf{y} \to \mathbf{x}} f(\mathbf{x}) = \pm\infty$.

Definition 23.4.4 *If $f(\mathbf{x}) \in \mathbb{R}$, $\lim_{\mathbf{y} \to \mathbf{x}} f(\mathbf{x}) = \infty$ if for every number l, there exists $\delta > 0$ such that whenever $|\mathbf{y} - \mathbf{x}| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, then $f(\mathbf{x}) > l$.*

The following theorem is just like the one variable version of calculus.

Theorem 23.4.5 *Suppose $\lim_{\mathbf{y} \to \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$ and $\lim_{\mathbf{y} \to \mathbf{x}} \mathbf{g}(\mathbf{y}) = \mathbf{K}$ where $\mathbf{K}, \mathbf{L} \in \mathbb{R}^q$. Then if a, $b \in \mathbb{R}$,*

$$\lim_{\mathbf{y} \to \mathbf{x}} (a\mathbf{f}(\mathbf{y}) + b\mathbf{g}(\mathbf{y})) = a\mathbf{L} + b\mathbf{K}, \tag{23.1}$$

$$\lim_{y \to x} \mathbf{f} \cdot \mathbf{g}(y) = \mathbf{L}\mathbf{K} \tag{23.2}$$

and if g is scalar valued with $\lim_{\mathbf{y}\to\mathbf{x}} g(\mathbf{y}) = K \neq 0$,

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y})\, g(\mathbf{y}) = \mathbf{L}K. \tag{23.3}$$

Also, if $\mathbf{h}$ *is a continuous function defined near* $\mathbf{L}$*, then*

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{h}\circ\mathbf{f}(\mathbf{y}) = \mathbf{h}(\mathbf{L}). \tag{23.4}$$

Suppose $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$*. If* $|\mathbf{f}(\mathbf{y}) - \mathbf{b}| \leq r$ *for all* $\mathbf{y}$ *sufficiently close to* $\mathbf{x}$*, then* $|\mathbf{L} - \mathbf{b}| \leq r$ *also.*

Proof: The proof of (23.1) is left for you. It is like a corresponding theorem for continuous functions. Now (23.2)is to be verified. Let $\varepsilon > 0$ be given. Then by the triangle inequality,

$$\begin{aligned} |\mathbf{f}\cdot\mathbf{g}(\mathbf{y}) - \mathbf{L}\cdot\mathbf{K}| &\leq |\mathbf{f}\mathbf{g}(\mathbf{y}) - \mathbf{f}(\mathbf{y})\cdot\mathbf{K}| + |\mathbf{f}(\mathbf{y})\cdot\mathbf{K} - \mathbf{L}\cdot\mathbf{K}| \\ &\leq |\mathbf{f}(\mathbf{y})|\,|\mathbf{g}(\mathbf{y}) - \mathbf{K}| + |\mathbf{K}|\,|\mathbf{f}(\mathbf{y}) - \mathbf{L}|. \end{aligned}$$

There exists δ_1 such that if $0 < |\mathbf{y} - \mathbf{x}| < \delta_1$ and $\mathbf{y} \in D(\mathbf{f})$, then

$$|\mathbf{f}(\mathbf{y}) - \mathbf{L}| < 1,$$

and so for such $\mathbf{y}$, the triangle inequality implies, $|\mathbf{f}(\mathbf{y})| < 1 + |\mathbf{L}|$. Therefore, for $0 < |\mathbf{y} - \mathbf{x}| < \delta_1$,

$$|\mathbf{f}\cdot\mathbf{g}(\mathbf{y}) - \mathbf{L}\cdot\mathbf{K}| \leq (1 + |\mathbf{K}| + |\mathbf{L}|)\,[|\mathbf{g}(\mathbf{y}) - \mathbf{K}| + |\mathbf{f}(\mathbf{y}) - \mathbf{L}|]. \tag{23.5}$$

Now let $0 < \delta_2$ be such that if $\mathbf{y} \in D(\mathbf{f})$ and $0 < |\mathbf{x} - \mathbf{y}| < \delta_2$,

$$|\mathbf{f}(\mathbf{y}) - \mathbf{L}| < \frac{\varepsilon}{2(1 + |\mathbf{K}| + |\mathbf{L}|)},\ |\mathbf{g}(\mathbf{y}) - \mathbf{K}| < \frac{\varepsilon}{2(1 + |\mathbf{K}| + |\mathbf{L}|)}.$$

Then letting $0 < \delta \leq \min(\delta_1, \delta_2)$, it follows from (23.5) that

$$|\mathbf{f}\cdot\mathbf{g}(\mathbf{y}) - \mathbf{L}\cdot\mathbf{K}| < \varepsilon$$

and this proves (23.2).

The proof of (23.3) is left to you.

Consider (23.4). Since $\mathbf{h}$ is continuous near $\mathbf{L}$, it follows that for $\varepsilon > 0$ given, there exists $\eta > 0$ such that if $|\mathbf{y} - \mathbf{L}| < \eta$, then

$$|\mathbf{h}(\mathbf{y}) - \mathbf{h}(\mathbf{L})| < \varepsilon$$

Now since $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$, there exists $\delta > 0$ such that if $0 < |\mathbf{y} - \mathbf{x}| < \delta$, then

$$|\mathbf{f}(\mathbf{y}) - \mathbf{L}| < \eta.$$

Therefore, if $0 < |\mathbf{y} - \mathbf{x}| < \delta$,

$$|\mathbf{h}(\mathbf{f}(\mathbf{y})) - \mathbf{h}(\mathbf{L})| < \varepsilon.$$

It only remains to verify the last assertion. Assume $|\mathbf{f}(\mathbf{y}) - \mathbf{b}| \leq r$. It is required to show that $|\mathbf{L} - \mathbf{b}| \leq r$. If this is not true, then $|\mathbf{L} - \mathbf{b}| > r$. Consider $B(\mathbf{L}, |\mathbf{L} - \mathbf{b}| - r)$. Since $\mathbf{L}$ is the limit of $\mathbf{f}$, it follows $\mathbf{f}(\mathbf{y}) \in B(\mathbf{L}, |\mathbf{L} - \mathbf{b}| - r)$ whenever $\mathbf{y} \in D(\mathbf{f})$ is close enough to $\mathbf{x}$. Thus, by the triangle inequality,

$$|\mathbf{f}(\mathbf{y}) - \mathbf{L}| < |\mathbf{L} - \mathbf{b}| - r$$

and so

$$\begin{aligned} r &< |\mathbf{L}-\mathbf{b}| - |\mathbf{f}(\mathbf{y})-\mathbf{L}| \leq ||\mathbf{b}-\mathbf{L}| - |\mathbf{f}(\mathbf{y})-\mathbf{L}|| \\ &\leq |\mathbf{b}-\mathbf{f}(\mathbf{y})|, \end{aligned}$$

a contradiction to the assumption that $|\mathbf{b}-\mathbf{f}(\mathbf{y})| \leq r$.

Theorem 23.4.6 *For $\mathbf{f}: D(\mathbf{f}) \to \mathbb{R}^q$ and $\mathbf{x} \in D(\mathbf{f})$ a limit point of $D(\mathbf{f})$, $\mathbf{f}$ is continuous at $\mathbf{x}$ if and only if*

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x}).$$

Proof: First suppose $\mathbf{f}$ is continuous at $\mathbf{x}$ a limit point of $D(\mathbf{f})$. Then for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $|\mathbf{y}-\mathbf{x}| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, then $|\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{y})| < \varepsilon$. In particular, this holds if $0 < |\mathbf{x}-\mathbf{y}| < \delta$ and this is just the definition of the limit. Hence $\mathbf{f}(\mathbf{x}) = \lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y})$.

Next suppose $\mathbf{x}$ is a limit point of $D(\mathbf{f})$ and $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x})$. This means that if $\varepsilon > 0$ there exists $\delta > 0$ such that for $0 < |\mathbf{x}-\mathbf{y}| < \delta$ and $\mathbf{y} \in D(\mathbf{f})$, it follows $|\mathbf{f}(\mathbf{y})-\mathbf{f}(\mathbf{x})| < \varepsilon$. However, if $\mathbf{y} = \mathbf{x}$, then $|\mathbf{f}(\mathbf{y})-\mathbf{f}(\mathbf{x})| = |\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{x})| = 0$ and so whenever $\mathbf{y} \in D(\mathbf{f})$ and $|\mathbf{x}-\mathbf{y}| < \delta$, it follows $|\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{y})| < \varepsilon$, showing $\mathbf{f}$ is continuous at $\mathbf{x}$.

The following theorem is important.

Theorem 23.4.7 *Suppose $\mathbf{f}: D(\mathbf{f}) \to \mathbb{R}^q$. Then for $\mathbf{x}$ a limit point of $D(\mathbf{f})$,*

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L} \tag{23.6}$$

if and only if

$$\lim_{\mathbf{y}\to\mathbf{x}} f_k(\mathbf{y}) = L_k \tag{23.7}$$

where $\mathbf{f}(\mathbf{y}) \equiv (f_1(\mathbf{y}), \cdots, f_p(\mathbf{y}))$ and $\mathbf{L} \equiv (L_1, \cdots, L_p)$.

In the case where $q = 3$ and $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$ and $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{g}(\mathbf{y}) = \mathbf{K}$, then

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) \times \mathbf{g}(\mathbf{y}) = \mathbf{L} \times \mathbf{K}. \tag{23.8}$$

Proof: Suppose (23.6). Then letting $\varepsilon > 0$ be given there exists $\delta > 0$ such that if $0 < |\mathbf{y}-\mathbf{x}| < \delta$, it follows

$$|f_k(\mathbf{y}) - L_k| \leq |\mathbf{f}(\mathbf{y})-\mathbf{L}| < \varepsilon$$

which verifies (23.7).

Now suppose (23.7) holds. Then letting $\varepsilon > 0$ be given, there exists δ_k such that if $0 < |\mathbf{y}-\mathbf{x}| < \delta_k$, then

$$|f_k(\mathbf{y}) - L_k| < \frac{\varepsilon}{\sqrt{p}}.$$

Let $0 < \delta < \min(\delta_1, \cdots, \delta_p)$. Then if $0 < |\mathbf{y}-\mathbf{x}| < \delta$, it follows

$$\begin{aligned} |\mathbf{f}(\mathbf{y})-\mathbf{L}| &= \left(\sum_{k=1}^{p} |f_k(\mathbf{y}) - L_k|^2\right)^{1/2} \\ &< \left(\sum_{k=1}^{p} \frac{\varepsilon^2}{p}\right)^{1/2} = \varepsilon. \end{aligned}$$

It remains to verify (23.8). But from the first part of this theorem and the description of the cross product presented earlier in terms of the permutation symbol,

$$\begin{aligned}\lim_{\mathbf{y}\to\mathbf{x}}\left(\mathbf{f}\left(\mathbf{y}\right)\times\mathbf{g}\left(\mathbf{y}\right)\right)_i &= \lim_{\mathbf{y}\to\mathbf{x}}\varepsilon_{ijk}f_j\left(\mathbf{y}\right)g_k\left(\mathbf{y}\right)\\ &= \varepsilon_{ijk}L_jK_k=\left(\mathbf{L}\times\mathbf{K}\right)_i.\end{aligned}$$

Therefore, from the first part of this theorem, this establishes (27.5). This completes the proof.

Example 23.4.8 *Find* $\lim_{(x,y)\to(3,1)}\left(\frac{x^2-9}{x-3},y\right)$.

It is clear that $\lim_{(x,y)\to(3,1)}\frac{x^2-9}{x-3}=6$ and $\lim_{(x,y)\to(3,1)}y=1$. Therefore, this limit equals $(6,1)$.

Example 23.4.9 *Find* $\lim_{(x,y)\to(0,0)}\frac{xy}{x^2+y^2}$.

First of all observe the domain of the function is $\mathbb{R}^2\setminus\{(0,0)\}$, every point in $\mathbb{R}^2$ except the origin. Therefore, $(0,0)$ is a limit point of the domain of the function so it might make sense to take a limit. However, just as in the case of a function of one variable, the limit may not exist. In fact, this is the case here. To see this, take points on the line $y=0$. At these points, the value of the function equals 0. Now consider points on the line $y=x$ where the value of the function equals $1/2$. Since arbitrarily close to $(0,0)$ there are points where the function equals $1/2$ and points where the function has the value 0, it follows there can be no limit. Just take $\varepsilon=1/10$ for example. You can't be within $1/10$ of $1/2$ and also within $1/10$ of 0 at the same time.

Note it is necessary to rely on the definition of the limit much more than in the case of a function of one variable and there are no easy ways to do limit problems for functions of more than one variable. It is what it is and you will not deal with these concepts without suffering and anguish.

23.5 Properties Of Continuous Functions

Functions of p variables have many of the same properties as functions of one variable. First there is a version of the extreme value theorem generalizing the one dimensional case.

Theorem 23.5.1 *Let C be closed and bounded and let $f:C\to\mathbb{R}$ be continuous. Then f achieves its maximum and its minimum on C. This means there exist, $\mathbf{x}_1,\mathbf{x}_2\in C$ such that for all $\mathbf{x}\in C$,*

$$f\left(\mathbf{x}_1\right)\le f\left(\mathbf{x}\right)\le f\left(\mathbf{x}_2\right).$$

There is also the long technical theorem about sums and products of continuous functions. These theorems are proved in the next section.

Theorem 23.5.2 *The following assertions are valid*

1. *The function, $a\mathbf{f}+b\mathbf{g}$ is continuous at $\mathbf{x}$ when $\mathbf{f}$, $\mathbf{g}$ are continuous at $\mathbf{x}\in D\left(\mathbf{f}\right)\cap D\left(\mathbf{g}\right)$ and $a,b\in\mathbb{R}$.*

2. *If and f and g are each real valued functions continuous at $\mathbf{x}$, then fg is continuous at $\mathbf{x}$. If, in addition to this, $g\left(\mathbf{x}\right)\neq0$, then f/g is continuous at $\mathbf{x}$.*

3. *If* $\mathbf{f}$ *is continuous at* $\mathbf{x}$, $\mathbf{f}(\mathbf{x}) \in D(\mathbf{g}) \subseteq \mathbb{R}^p$, *and* $\mathbf{g}$ *is continuous at* $\mathbf{f}(\mathbf{x})$, *then* $\mathbf{g} \circ \mathbf{f}$ *is continuous at* $\mathbf{x}$.

4. *If* $\mathbf{f} = (f_1, \cdots, f_q) : D(\mathbf{f}) \to \mathbb{R}^q$, *then* $\mathbf{f}$ *is continuous if and only if each* f_k *is a continuous real valued function.*

5. *The function* $f : \mathbb{R}^p \to \mathbb{R}$, *given by* $f(\mathbf{x}) = |\mathbf{x}|$ *is continuous.*

23.6 Exercises

1. Let $\mathbf{f}(t) = \left(t, t^2+1, \frac{t}{t+1}\right)$ and let $\mathbf{g}(t) = \left(t+1, 1, \frac{t}{t^2+1}\right)$. Find $\mathbf{f} \cdot \mathbf{g}$.

2. Let $\mathbf{f}, \mathbf{g}$ be given in the previous problem. Find $\mathbf{f} \times \mathbf{g}$.

3. Find $D(\mathbf{f})$ if $\mathbf{f}(x,y,z,w) = \left(\frac{xy}{zw}, \sqrt{6-x^2y^2}\right)$.

4. Let $\mathbf{f}(t) = \left(t, t^2, t^3\right), \mathbf{g}(t) = \left(1, t, t^2\right)$, and $\mathbf{h}(t) = (\sin t, t, 1)$. Find the time rate of change of the volume of the parallelepiped spanned by the vectors $\mathbf{f}, \mathbf{g}$, and $\mathbf{h}$.

5. Let $\mathbf{f}(t) = (t, \sin t)$. Show f is continuous at every point t.

6. Suppose $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq K|\mathbf{x}-\mathbf{y}|$ where K is a constant. Show that $\mathbf{f}$ is everywhere continuous. Functions satisfying such an inequality are called Lipschitz functions.

7. Suppose $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq K|\mathbf{x}-\mathbf{y}|^\alpha$ where K is a constant and $\alpha \in (0,1)$. Show that $\mathbf{f}$ is everywhere continuous.

8. Suppose $f : \mathbb{R}^3 \to \mathbb{R}$ is given by $f(\mathbf{x}) = 3x_1x_2 + 2x_3^2$. Use Theorem 23.3.2 to verify that f is continuous. **Hint:** You should first verify that the function, $\pi_k : \mathbb{R}^3 \to \mathbb{R}$ given by $\pi_k(\mathbf{x}) = x_k$ is a continuous function.

9. Show that if $f : \mathbb{R}^q \to \mathbb{R}$ is a polynomial then it is continuous.

10. State and prove a theorem which involves quotients of functions encountered in the previous problem.

11. Let
$$f(x,y) \equiv \begin{cases} \frac{xy}{x^2+y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}.$$
Find $\lim_{(x,y)\to(0,0)} f(x,y)$ if it exists. If it does not exist, tell why it does not exist. **Hint:** Consider along the line $y = x$ and along the line $y = 0$.

12. Find the following limits if possible

    (a) $\lim_{(x,y)\to(0,0)} \frac{x^2-y^2}{x^2+y^2}$

    (b) $\lim_{(x,y)\to(0,0)} \frac{x\left(x^2-y^2\right)}{(x^2+y^2)}$

    (c) $\lim_{(x,y)\to(0,0)} \frac{\left(x^2-y^4\right)^2}{(x^2+y^4)^2}$ **Hint:** Consider along $y = 0$ and along $x = y^2$.

    (d) $\lim_{(x,y)\to(0,0)} x \sin\left(\frac{1}{x^2+y^2}\right)$

    (e) $\lim_{(x,y)\to(1,2)} \frac{-2yx^2+8yx+34y+3y^3-18y^2+6x^2-13x-20-xy^2-x^3}{-y^2+4y-5-x^2+2x}$. **Hint:** It might help to write this in terms of the variables $(s,t) = (x-1, y-2)$.

13. In the definition of limit, why must $\mathbf{x}$ be a limit point of $D(\mathbf{f})$? **Hint:** If $\mathbf{x}$ were not a limit point of $D(\mathbf{f})$, show there exists $\delta > 0$ such that $B(\mathbf{x}, \delta)$ contains no points of $D(\mathbf{f})$ other than possibly $\mathbf{x}$ itself. Argue that 33.3 is a limit and that so is 22 and 7 and 11. In other words the concept is totally worthless.

14. Suppose $\lim_{x\to 0} f(x,0) = 0 = \lim_{y\to 0} f(0,y)$. Does it follow that $\lim_{(x,y)\to(0,0)} f(x,y) = 0$? Prove or give counter example.

15. $\mathbf{f} : D \subseteq \mathbb{R}^p \to \mathbb{R}^q$ is Lipschitz continuous or just Lipschitz for short if there exists a constant, K such that
$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq K |\mathbf{x} - \mathbf{y}|$$
for all $\mathbf{x}, \mathbf{y} \in D$. Show every Lipschitz function is uniformly continuous which means that given $\varepsilon > 0$ there exists $\delta > 0$ independent of $\mathbf{x}$ such that if $|\mathbf{x} - \mathbf{y}| < \delta$, then $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \varepsilon$.

16. If $\mathbf{f}$ is uniformly continuous, does it follow that $|\mathbf{f}|$ is also uniformly continuous? If $|\mathbf{f}|$ is uniformly continuous does it follow that $\mathbf{f}$ is uniformly continuous? Answer the same questions with "uniformly continuous" replaced with "continuous". Explain why.

17. Let f be defined on the positive integers. Thus $D(f) = \mathbb{N}$. Show that f is automatically continuous at every point of $D(f)$. Is it also uniformly continuous? What does this mean about the concept of continuous functions being those which can be graphed without taking the pencil off the paper?

18. In Problem 12c show $\lim_{t\to 0} f(tx, ty) = 1$ for any choice of (x, y). Using Problem 12c what does this tell you about limits existing just because the limit along any line exists.

19. Let $f(x,y,z) = x^2 y + \sin(xyz)$. Does f achieve a maximum on the set
$$\left\{(x,y,z) : x^2 + y^2 + 2z^2 \leq 8\right\}?$$
Explain why.

20. Suppose $\mathbf{x}$ is defined to be a limit point of a set, A if and only if for all $r > 0$, $B(\mathbf{x}, r)$ contains a point of A different than $\mathbf{x}$. Show this is equivalent to the above definition of limit point.

21. Give an example of a set of points in $\mathbb{R}^3$ which has no limit points. Show that if $D(\mathbf{f})$ equals this set, then $\mathbf{f}$ is continuous. Show that more generally, if $\mathbf{f}$ is any function for which $D(\mathbf{f})$ has no limit points, then $\mathbf{f}$ is continuous.

22. Let $\{\mathbf{x}_k\}_{k=1}^n$ be any finite set of points in $\mathbb{R}^p$. Show this set has no limit points.

23. Suppose S is any set of points such that every pair of points is at least as far apart as 1. Show S has no limit points.

24. Find $\lim_{\mathbf{x}\to\mathbf{0}} \frac{\sin(|\mathbf{x}|)}{|\mathbf{x}|}$ and prove your answer from the definition of limit.

25. Suppose $\mathbf{g}$ is a continuous vector valued function of one variable defined on $[0, \infty)$. Prove
$$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{g}(|\mathbf{x}|) = \mathbf{g}(|\mathbf{x}_0|).$$

26. Give some examples of limit problems for functions of many variables which have limits and prove your assertions.

23.7 Some Fundamentals

This section contains the proofs of the theorems which were stated without proof along with some other significant topics which will be useful later. These topics are of fundamental significance but are difficult.

Theorem 23.7.1 *The following assertions are valid*

1. *The function,* $a\mathbf{f}+b\mathbf{g}$ *is continuous at* $\mathbf{x}$ *when* $\mathbf{f}$, $\mathbf{g}$ *are continuous at* $\mathbf{x} \in D(\mathbf{f}) \cap D(\mathbf{g})$ *and* $a, b \in \mathbb{R}$.
2. *If and* f *and* g *are each real valued functions continuous at* $\mathbf{x}$, *then* fg *is continuous at* $\mathbf{x}$. *If, in addition to this,* $g(\mathbf{x}) \neq 0$, *then* f/g *is continuous at* $\mathbf{x}$.
3. *If* $\mathbf{f}$ *is continuous at* $\mathbf{x}$, $\mathbf{f}(\mathbf{x}) \in D(\mathbf{g}) \subseteq \mathbb{R}^p$, *and* $\mathbf{g}$ *is continuous at* $\mathbf{f}(\mathbf{x})$, *then* $\mathbf{g} \circ \mathbf{f}$ *is continuous at* $\mathbf{x}$.
4. *If* $\mathbf{f} = (f_1, \cdots, f_q) : D(\mathbf{f}) \to \mathbb{R}^q$, *then* $\mathbf{f}$ *is continuous if and only if each* f_k *is a continuous real valued function.*
5. *The function* $f : \mathbb{R}^p \to \mathbb{R}$, *given by* $f(\mathbf{x}) = |\mathbf{x}|$ *is continuous.*

Proof: Begin with 1.) Let $\varepsilon > 0$ be given. By assumption, there exist $\delta_1 > 0$ such that whenever $|\mathbf{x}-\mathbf{y}| < \delta_1$, it follows $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \frac{\varepsilon}{2(|a|+|b|+1)}$ and there exists $\delta_2 > 0$ such that whenever $|\mathbf{x}-\mathbf{y}| < \delta_2$, it follows that $|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})| < \frac{\varepsilon}{2(|a|+|b|+1)}$. Then let $0 < \delta \leq \min(\delta_1, \delta_2)$. If $|\mathbf{x}-\mathbf{y}| < \delta$, then everything happens at once. Therefore, using the triangle inequality

$$|a\mathbf{f}(\mathbf{x}) + b\mathbf{f}(\mathbf{x}) - (a\mathbf{g}(\mathbf{y}) + b\mathbf{g}(\mathbf{y}))|$$

$$\begin{aligned} &\leq |a|\,|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| + |b|\,|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})| \\ &< |a|\left(\frac{\varepsilon}{2(|a|+|b|+1)}\right) + |b|\left(\frac{\varepsilon}{2(|a|+|b|+1)}\right) < \varepsilon. \end{aligned}$$

Now begin on 2.) There exists $\delta_1 > 0$ such that if $|\mathbf{y}-\mathbf{x}| < \delta_1$, then $|f(\mathbf{x}) - f(\mathbf{y})| < 1$. Therefore, for such $\mathbf{y}$,

$$|f(\mathbf{y})| < 1 + |f(\mathbf{x})|.$$

It follows that for such $\mathbf{y}$,

$$|fg(\mathbf{x}) - fg(\mathbf{y})| \leq |f(\mathbf{x})\,g(\mathbf{x}) - g(\mathbf{x})\,f(\mathbf{y})| + |g(\mathbf{x})\,f(\mathbf{y}) - f(\mathbf{y})\,g(\mathbf{y})|$$

$$\begin{aligned} &\leq |g(\mathbf{x})|\,|f(\mathbf{x}) - f(\mathbf{y})| + |f(\mathbf{y})|\,|g(\mathbf{x}) - g(\mathbf{y})| \\ &\leq (1 + |g(\mathbf{x})| + |f(\mathbf{y})|)\,[|g(\mathbf{x}) - g(\mathbf{y})| + |f(\mathbf{x}) - f(\mathbf{y})|]. \end{aligned}$$

Now let $\varepsilon > 0$ be given. There exists δ_2 such that if $|\mathbf{x}-\mathbf{y}| < \delta_2$, then

$$|g(\mathbf{x}) - g(\mathbf{y})| < \frac{\varepsilon}{2(1+|g(\mathbf{x})|+|f(\mathbf{y})|)},$$

and there exists δ_3 such that if $|\mathbf{x}-\mathbf{y}| < \delta_3$, then

$$|f(\mathbf{x}) - f(\mathbf{y})| < \frac{\varepsilon}{2(1+|g(\mathbf{x})|+|f(\mathbf{y})|)}$$

Now let $0 < \delta \leq \min(\delta_1, \delta_2, \delta_3)$. Then if $|\mathbf{x}-\mathbf{y}| < \delta$, all the above hold at once and

$$|fg(\mathbf{x}) - fg(\mathbf{y})| \leq$$

$$\begin{aligned}&(1+|g(\mathbf{x})|+|f(\mathbf{y})|)\left[|g(\mathbf{x})-g(\mathbf{y})|+|f(\mathbf{x})-f(\mathbf{y})|\right]\\ &< (1+|g(\mathbf{x})|+|f(\mathbf{y})|)\left(\frac{\varepsilon}{2(1+|g(\mathbf{x})|+|f(\mathbf{y})|)}+\frac{\varepsilon}{2(1+|g(\mathbf{x})|+|f(\mathbf{y})|)}\right)=\varepsilon.\end{aligned}$$

This proves the first part of 2.) To obtain the second part, let δ_1 be as described above and let $\delta_0 > 0$ be such that for $|\mathbf{x}-\mathbf{y}| < \delta_0$,

$$|g(\mathbf{x}) - g(\mathbf{y})| < |g(\mathbf{x})|/2$$

and so by the triangle inequality,

$$-|g(\mathbf{x})|/2 \leq |g(\mathbf{y})| - |g(\mathbf{x})| \leq |g(\mathbf{x})|/2$$

which implies $|g(\mathbf{y})| \geq |g(\mathbf{x})|/2$, and $|g(\mathbf{y})| < 3|g(\mathbf{x})|/2$.

Then if $|\mathbf{x}-\mathbf{y}| < \min(\delta_0, \delta_1)$,

$$\begin{aligned}\left|\frac{f(\mathbf{x})}{g(\mathbf{x})} - \frac{f(\mathbf{y})}{g(\mathbf{y})}\right| &= \left|\frac{f(\mathbf{x})g(\mathbf{y}) - f(\mathbf{y})g(\mathbf{x})}{g(\mathbf{x})g(\mathbf{y})}\right|\\ &\leq \frac{|f(\mathbf{x})g(\mathbf{y}) - f(\mathbf{y})g(\mathbf{x})|}{\left(\frac{|g(\mathbf{x})|^2}{2}\right)}\\ &= \frac{2|f(\mathbf{x})g(\mathbf{y}) - f(\mathbf{y})g(\mathbf{x})|}{|g(\mathbf{x})|^2}\end{aligned}$$

$$\begin{aligned}&\leq \frac{2}{|g(\mathbf{x})|^2}\left[|f(\mathbf{x})g(\mathbf{y}) - f(\mathbf{y})g(\mathbf{y}) + f(\mathbf{y})g(\mathbf{y}) - f(\mathbf{y})g(\mathbf{x})|\right]\\ &\leq \frac{2}{|g(\mathbf{x})|^2}\left[|g(\mathbf{y})||f(\mathbf{x})-f(\mathbf{y})| + |f(\mathbf{y})||g(\mathbf{y})-g(\mathbf{x})|\right]\\ &\leq \frac{2}{|g(\mathbf{x})|^2}\left[\frac{3}{2}|\mathbf{g}(\mathbf{x})||f(\mathbf{x})-f(\mathbf{y})| + (1+|f(\mathbf{x})|)|g(\mathbf{y})-g(\mathbf{x})|\right]\\ &\leq \frac{2}{|g(\mathbf{x})|^2}(1+2|f(\mathbf{x})|+2|g(\mathbf{x})|)\left[|f(\mathbf{x})-f(\mathbf{y})|+|g(\mathbf{y})-g(\mathbf{x})|\right]\\ &\equiv M\left[|f(\mathbf{x})-f(\mathbf{y})|+|g(\mathbf{y})-g(\mathbf{x})|\right]\end{aligned}$$

where

$$M \equiv \frac{2}{|g(\mathbf{x})|^2}(1+2|f(\mathbf{x})|+2|g(\mathbf{x})|)$$

Now let δ_2 be such that if $|\mathbf{x}-\mathbf{y}|<\delta_2$, then

$$|f(\mathbf{x})-f(\mathbf{y})|<\frac{\varepsilon}{2}M^{-1}$$

and let δ_3 be such that if $|\mathbf{x}-\mathbf{y}|<\delta_3$, then

$$|g(\mathbf{y})-g(\mathbf{x})|<\frac{\varepsilon}{2}M^{-1}.$$

Then if $0<\delta\leq\min(\delta_0,\delta_1,\delta_2,\delta_3)$, and $|\mathbf{x}-\mathbf{y}|<\delta$, everything holds and

$$\begin{aligned}\left|\frac{f(\mathbf{x})}{g(\mathbf{x})}-\frac{f(\mathbf{y})}{g(\mathbf{y})}\right| &\leq M\left[|f(\mathbf{x})-f(\mathbf{y})|+|g(\mathbf{y})-g(\mathbf{x})|\right]\\ &< M\left[\frac{\varepsilon}{2}M^{-1}+\frac{\varepsilon}{2}M^{-1}\right]=\varepsilon.\end{aligned}$$

This completes the proof of the second part of 2.) Note that in these proofs no effort is made to find some sort of "best" δ. The problem is one which has a yes or a no answer. Either is it or it is not continuous.

Now begin on 3.). If $\mathbf{f}$ is continuous at $\mathbf{x}$, $\mathbf{f}(\mathbf{x})\in D(\mathbf{g})\subseteq\mathbb{R}^p$, and $\mathbf{g}$ is continuous at $\mathbf{f}(\mathbf{x})$,then $\mathbf{g}\circ\mathbf{f}$ is continuous at $\mathbf{x}$. Let $\varepsilon>0$ be given. Then there exists $\eta>0$ such that if $|\mathbf{y}-\mathbf{f}(\mathbf{x})|<\eta$ and $\mathbf{y}\in D(\mathbf{g})$, it follows that $|\mathbf{g}(\mathbf{y})-\mathbf{g}(\mathbf{f}(\mathbf{x}))|<\varepsilon$. It follows from continuity of $\mathbf{f}$ at $\mathbf{x}$ that there exists $\delta>0$ such that if $|\mathbf{x}-\mathbf{z}|<\delta$ and $\mathbf{z}\in D(\mathbf{f})$, then $|\mathbf{f}(\mathbf{z})-\mathbf{f}(\mathbf{x})|<\eta$. Then if $|\mathbf{x}-\mathbf{z}|<\delta$ and $\mathbf{z}\in D(\mathbf{g}\circ\mathbf{f})\subseteq D(\mathbf{f})$, all the above hold and so

$$|\mathbf{g}(\mathbf{f}(\mathbf{z}))-\mathbf{g}(\mathbf{f}(\mathbf{x}))|<\varepsilon.$$

This proves part 3.)

Part 4.) says: If $\mathbf{f}=(f_1,\cdots,f_q):D(\mathbf{f})\to\mathbb{R}^q$, then $\mathbf{f}$ is continuous if and only if each f_k is a continuous real valued function. Then

$$\begin{aligned}|f_k(\mathbf{x})-f_k(\mathbf{y})| &\leq |\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{y})|\\ &\equiv\left(\sum_{i=1}^{q}|f_i(\mathbf{x})-f_i(\mathbf{y})|^2\right)^{1/2}\\ &\leq\sum_{i=1}^{q}|f_i(\mathbf{x})-f_i(\mathbf{y})|.\end{aligned}\tag{23.9}$$

Suppose first that $\mathbf{f}$ is continuous at $\mathbf{x}$. Then there exists $\delta>0$ such that if $|\mathbf{x}-\mathbf{y}|<\delta$, then $|\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{y})|<\varepsilon$. The first part of the above inequality then shows that for each $k=1,\cdots,q$, $|f_k(\mathbf{x})-f_k(\mathbf{y})|<\varepsilon$. This shows the only if part. Now suppose each function, f_k is continuous. Then if $\varepsilon>0$ is given, there exists $\delta_k>0$ such that whenever $|\mathbf{x}-\mathbf{y}|<\delta_k$

$$|f_k(\mathbf{x})-f_k(\mathbf{y})|<\varepsilon/q.$$

Now let $0<\delta\leq\min(\delta_1,\cdots,\delta_q)$. For $|\mathbf{x}-\mathbf{y}|<\delta$, the above inequality holds for all k and so the last part of (23.9) implies

$$\begin{aligned}|\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{y})| &\leq\sum_{i=1}^{q}|f_i(\mathbf{x})-f_i(\mathbf{y})|\\ &<\sum_{i=1}^{q}\frac{\varepsilon}{q}=\varepsilon.\end{aligned}$$

23.7.2 The Extreme Value Theorem

Definition 23.7.4 *A set, $C \subseteq \mathbb{R}^p$ is said to be **bounded** if $C \subseteq \prod_{i=1}^p [a_i, b_i]$ for some choice of intervals, $[a_i, b_i]$ where $-\infty < a_i < b_i < \infty$. The **diameter** of a set, S, is defined as*

$$\operatorname{diam}(S) \equiv \sup\{|\mathbf{x} - \mathbf{y}| : \mathbf{x}, \mathbf{y} \in S\}.$$

A function, $\mathbf{f}$ having values in $\mathbb{R}^p$ is said to be bounded if the set of values of $\mathbf{f}$ is a bounded set.

Thus $\operatorname{diam}(S)$ is just a careful description of what you would think of as the diameter. It measures how stretched out the set is.

Lemma 23.7.5 *Let $C \subseteq \mathbb{R}^p$ be closed and bounded and let $f : C \to \mathbb{R}$ be continuous. Then f is bounded.*

Proof: Suppose not. Since C is bounded, it follows $C \subseteq \prod_{i=1}^p [a_i, b_i] \equiv I_0$ for some closed intervals, $[a_i, b_i]$. Consider all sets of the form $\prod_{i=1}^p [c_i, d_i]$ where $[c_i, d_i]$ equals either $\left[a_i, \frac{a_i+b_i}{2}\right]$ or $[c_i, d_i] = \left[\frac{a_i+b_i}{2}, b_i\right]$. Thus there are 2^p of these sets because there are two choices for the i^{th} slot for $i = 1, \cdots, p$. Also, if $\mathbf{x}$ and $\mathbf{y}$ are two points in one of these sets,

$$|x_i - y_i| \leq 2^{-1} |b_i - a_i|.$$

Observe that $\operatorname{diam}(I_0) = \left(\sum_{i=1}^p |b_i - a_i|^2\right)^{1/2}$ because for $\mathbf{x}, \mathbf{y} \in I_0, |x_i - y_i| \leq |a_i - b_i|$ for each $i = 1, \cdots, p$,

$$\begin{aligned} |\mathbf{x} - \mathbf{y}| &= \left(\sum_{i=1}^p |x_i - y_i|^2\right)^{1/2} \\ &\leq 2^{-1} \left(\sum_{i=1}^p |b_i - a_i|^2\right)^{1/2} \equiv 2^{-1} \operatorname{diam}(I_0). \end{aligned}$$

Denote by $\{J_1, \cdots, J_{2^p}\}$ these sets determined above. It follows the diameter of each set is no larger than $2^{-1} \operatorname{diam}(I_0)$. In particular, since $\mathbf{d} \equiv (d_1, \cdots, d_p)$ and $\mathbf{c} \equiv (c_1, \cdots, c_p)$ are two such points, for each J_k,

$$\operatorname{diam}(J_k) \equiv \left(\sum_{i=1}^p |d_i - c_i|^2\right)^{1/2} \leq 2^{-1} \operatorname{diam}(I_0)$$

Since the union of these sets equals all of I_0, it follows

$$C = \cup_{k=1}^{2^p} J_k \cap C.$$

If f is not bounded on C, it follows that for some k, f is not bounded on $J_k \cap C$. Let $I_1 \equiv J_k$ and let $C_1 = C \cap I_1$. Now do to I_1 and C_1 what was done to I_0 and C to obtain $I_2 \subseteq I_1$, and for $\mathbf{x}, \mathbf{y} \in I_2$,

$$|\mathbf{x} - \mathbf{y}| \leq 2^{-1} \operatorname{diam}(I_1) \leq 2^{-2} \operatorname{diam}(I_2),$$

and f is unbounded on $I_2 \cap C_1 \equiv C_2$. Continue in this way obtaining sets, I_k such that $I_k \supseteq I_{k+1}$ and $\operatorname{diam}(I_k) \leq 2^{-k} \operatorname{diam}(I_0)$ and f is unbounded on $I_k \cap C$. By the nested interval lemma, there exists a point, $\mathbf{c}$ which is contained in each I_k.

Claim: $\mathbf{c} \in C$.

Proof of claim: Suppose $\mathbf{c} \notin C$. Since C is a closed set, there exists $r > 0$ such that $B(\mathbf{c}, r)$ is contained completely in $\mathbb{R}^p \setminus C$. In other words, $B(\mathbf{c}, r)$ contains no points of C. Let k be so large that $\operatorname{diam}(I_0) 2^{-k} < r$. Then since $\mathbf{c} \in I_k$, and any two points of I_k are closer than $\operatorname{diam}(I_0) 2^{-k}$, I_k must be contained in $B(\mathbf{c}, r)$ and so has no points of C in it, contrary to the manner in which the I_k are defined in which f is unbounded on $I_k \cap C$. Therefore, $\mathbf{c} \in C$ as claimed.

Now for k large enough, and $\mathbf{x} \in C \cap I_k$, the continuity of f implies $|f(\mathbf{c}) - f(\mathbf{x})| < 1$ contradicting the manner in which I_k was chosen since this inequality implies f is bounded on $I_k \cap C$. This proves the theorem.

Here is a proof of the extreme value theorem.

Theorem 23.7.6 *Let C be closed and bounded and let $f : C \to \mathbb{R}$ be continuous. Then f achieves its maximum and its minimum on C. This means there exist, $\mathbf{x}_1, \mathbf{x}_2 \in C$ such that for all $\mathbf{x} \in C$,*

$$f(\mathbf{x}_1) \leq f(\mathbf{x}) \leq f(\mathbf{x}_2).$$

Proof: Let $M = \sup\{f(\mathbf{x}) : \mathbf{x} \in C\}$. Then by Lemma 23.7.5, M is a finite number. Is $f(\mathbf{x}_2) = M$ for some x_2? if not, you could consider the function,

$$g(\mathbf{x}) \equiv \frac{1}{M - f(\mathbf{x})}$$

and g would be a continuous and unbounded function defined on C, contrary to Lemma 23.7.5. Therefore, there exists $\mathbf{x}_2 \in C$ such that $f(\mathbf{x}_2) = M$. A similar argument applies to show the existence of $\mathbf{x}_1 \in C$ such that

$$f(\mathbf{x}_1) = \inf\{f(\mathbf{x}) : \mathbf{x} \in C\}.$$

This proves the theorem.

23.7.3 Sequences And Completeness

Definition 23.7.7 *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 $\{f_i\}_{i=k}^{\infty}$ or just $\{f_i\}$ for short. The letter used for the name of the sequence is not important. Thus it is all right to let a be the name of a sequence or to refer to it as $\{a_i\}$. When the sequence has values in $\mathbb{R}^p$, it is customary to write it in bold face. Thus $\{\mathbf{a}_i\}$ would refer to a sequence having values in $\mathbb{R}^p$ for some $p > 1$.*

Example 23.7.8 *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 to us 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 23.7.9 *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.

Example 23.7.10 *Let* $\mathbf{a}_k = (k, \sin(k))$. *Thus this sequence has values in* $\mathbb{R}^2$.

Definition 23.7.11 *Let* $\{\mathbf{a}_n\}$ *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 index used to define the sequence* $\{\mathbf{a}_n\}$. *Then if* $\mathbf{b}_k \equiv \mathbf{a}_{n_k}$, $\{\mathbf{b}_k\}$ *is called a subsequence of* $\{\mathbf{a}_n\}$.

For example, suppose $a_n = \left(n^2 + 1\right)$. Thus $a_1 = 2$, $a_3 = 10$, etc. If

$$n_1 = 1, n_2 = 3, n_3 = 5, \cdots, n_k = 2k - 1,$$

then letting $b_k = a_{n_k}$, it follows

$$b_k = \left(\left(2k-1\right)^2 + 1\right) = 4k^2 - 4k + 2.$$

Definition 23.7.12 *A sequence,* $\{\mathbf{a}_k\}$ *is said to* ***converge*** *to* $\mathbf{a}$ *if for every* $\varepsilon > 0$ *there exists* n_ε *such that if* $n > n_\varepsilon$, *then* $|\mathbf{a} - \mathbf{a}_\varepsilon| < \varepsilon$. *The usual notation for this is* $\lim_{n\to\infty} \mathbf{a}_n = \mathbf{a}$ *although it is often written as* $\mathbf{a}_n \to \mathbf{a}$.

The following theorem says the limit, if it exists, is unique.

Theorem 23.7.13 *If a sequence,* $\{\mathbf{a}_n\}$ *converges to* $\mathbf{a}$ *and to* $\mathbf{b}$ *then* $\mathbf{a} = \mathbf{b}$.

Proof: There exists n_ε such that if $n > n_\varepsilon$ then $|\mathbf{a}_n - \mathbf{a}| < \frac{\varepsilon}{2}$ and if $n > n_\varepsilon$, then $|\mathbf{a}_n - \mathbf{b}| < \frac{\varepsilon}{2}$. Then pick such an n.

$$|\mathbf{a} - \mathbf{b}| < |\mathbf{a} - \mathbf{a}_n| + |\mathbf{a}_n - \mathbf{b}| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since ε is arbitrary, this proves the theorem.

The following is the definition of a Cauchy sequencein $\mathbb{R}^p$.

Definition 23.7.14 $\{\mathbf{a}_n\}$ *is a Cauchy sequence if for all* $\varepsilon > 0$, *there exists* n_ε *such that whenever* $n, m \geq n_\varepsilon$,

$$|\mathbf{a}_n - \mathbf{a}_m| < \varepsilon.$$

A sequence is Cauchy means the terms are "bunching up to each other" as m, n get large.

Theorem 23.7.15 *The set of terms in a Cauchy sequence in* $\mathbb{R}^p$ *is bounded in the sense that for all* n, $|\mathbf{a}_n| < M$ *for some* $M < \infty$.

Proof: Let $\varepsilon = 1$ in the definition of a Cauchy sequence and let $n > n_1$. Then from the definition,

$$|\mathbf{a}_n - \mathbf{a}_{n_1}| < 1.$$

It follows that for all $n > n_1$,

$$|\mathbf{a}_n| < 1 + |\mathbf{a}_{n_1}|.$$

Therefore, for all n,

$$|\mathbf{a}_n| \leq 1 + |\mathbf{a}_{n_1}| + \sum_{k=1}^{n_1} |\mathbf{a}_k|.$$

This proves the theorem.

Theorem 23.7.16 *If a sequence $\{\mathbf{a}_n\}$ in $\mathbb{R}^p$ converges, then the sequence is a Cauchy sequence. Also, if some subsequence of a Cauchy sequence converges, then the original sequence converges.*

Proof: Let $\varepsilon > 0$ be given and suppose $\mathbf{a}_n \to \mathbf{a}$. Then from the definition of convergence, there exists n_ε such that if $n > n_\varepsilon$, it follows that

$$|\mathbf{a}_n - \mathbf{a}| < \frac{\varepsilon}{2}$$

Therefore, if $m, n \geq n_\varepsilon + 1$, it follows that

$$|\mathbf{a}_n - \mathbf{a}_m| \leq |\mathbf{a}_n - \mathbf{a}| + |\mathbf{a} - \mathbf{a}_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

showing that, since $\varepsilon > 0$ is arbitrary, $\{\mathbf{a}_n\}$ is a Cauchy sequence. It remains to show the last claim. Suppose then that $\{\mathbf{a}_n\}$ is a Cauchy sequence and $\mathbf{a} = \lim_{k\to\infty} \mathbf{a}_{n_k}$ where $\{\mathbf{a}_{n_k}\}_{k=1}^{\infty}$ is a subsequence. Let $\varepsilon > 0$ be given. Then there exists K such that if $k, l \geq K$, then $|\mathbf{a}_k - \mathbf{a}_l| < \frac{\varepsilon}{2}$. Then if $k > K$, it follows $n_k > K$ because $n_1, n_2, n_3, \cdots$ is strictly increasing as the subscript increases. Also, there exists K_1 such that if $k > K_1, |\mathbf{a}_{n_k} - \mathbf{a}| < \frac{\varepsilon}{2}$. Then letting $n > \max(K, K_1)$, pick $k > \max(K, K_1)$. Then

$$|\mathbf{a} - \mathbf{a}_n| \leq |\mathbf{a} - \mathbf{a}_{n_k}| + |\mathbf{a}_{n_k} - \mathbf{a}_n| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This proves the theorem.

Definition 23.7.17 *A set, K in $\mathbb{R}^p$ is said to be sequentially compact if every sequence in K has a subsequence which converges to a point of K.*

Theorem 23.7.18 *If $I_0 = \prod_{i=1}^p [a_i, b_i]$ where $a_i \leq b_i$, then I_0 is sequentially compact.*

Proof: Let $\{\mathbf{a}_i\}_{i=1}^{\infty} \subseteq I_0$ and consider all sets of the form $\prod_{i=1}^p [c_i, d_i]$ where $[c_i, d_i]$ equals either $\left[a_i, \frac{a_i+b_i}{2}\right]$ or $[c_i, d_i] = \left[\frac{a_i+b_i}{2}, b_i\right]$. Thus there are 2^p of these sets because there are two choices for the i^{th} slot for $i = 1, \cdots, p$. Also, if $\mathbf{x}$ and $\mathbf{y}$ are two points in one of these sets,

$$|x_i - y_i| \leq 2^{-1} |b_i - a_i|.$$

$\operatorname{diam}(I_0) = \left(\sum_{i=1}^p |b_i - a_i|^2\right)^{1/2}$,

$$\begin{aligned} |\mathbf{x} - \mathbf{y}| &= \left(\sum_{i=1}^p |x_i - y_i|^2\right)^{1/2} \\ &\leq 2^{-1} \left(\sum_{i=1}^p |b_i - a_i|^2\right)^{1/2} \equiv 2^{-1} \operatorname{diam}(I_0). \end{aligned}$$

In particular, since $\mathbf{d} \equiv (d_1, \cdots, d_p)$ and $\mathbf{c} \equiv (c_1, \cdots, c_p)$ are two such points,

$$D_1 \equiv \left(\sum_{i=1}^p |d_i - c_i|^2\right)^{1/2} \leq 2^{-1} \operatorname{diam}(I_0)$$

Denote by $\{J_1, \cdots, J_{2^p}\}$ these sets determined above. Since the union of these sets equals all of $I_0 \equiv I$, it follows that for some J_k, the sequence, $\{\mathbf{a}_i\}$ is contained in J_k for infinitely many k. Let that one be called I_1. Next do for I_1 what was done for I_0 to get $I_2 \subseteq I_1$ such

that the diameter is half that of I_1 and I_2 contains $\{\mathbf{a}_k\}$ for infinitely many values of k. Continue in this way obtaining a nested sequence of intervals, $\{I_k\}$ such that $I_k \supseteq I_{k+1}$, and if $\mathbf{x}, \mathbf{y} \in I_k$, then $|\mathbf{x} - \mathbf{y}| \leq 2^{-k} \operatorname{diam}(I_0)$, and I_n contains $\{\mathbf{a}_k\}$ for infinitely many values of k for each n. Then by the nested interval lemma, there exists $\mathbf{c}$ such that $\mathbf{c}$ is contained in each I_k. Pick $\mathbf{a}_{n_1} \in I_1$. Next pick $n_2 > n_1$ such that $\mathbf{a}_{n_2} \in I_2$. If $\mathbf{a}_{n_1}, \cdots, \mathbf{a}_{n_k}$ have been chosen, let $\mathbf{a}_{n_{k+1}} \in I_{k+1}$ and $n_{k+1} > n_k$. This can be done because in the construction, I_n contains $\{\mathbf{a}_k\}$ for infinitely many k. Thus the distance between $\mathbf{a}_{n_k}$ and $\mathbf{c}$ is no larger than $2^{-k} \operatorname{diam}(I_0)$ and so $\lim_{k\to\infty} \mathbf{a}_{n_k} = \mathbf{c} \in I_0$. This proves the theorem.

Theorem 23.7.19 *Every Cauchy sequence in $\mathbb{R}^p$ converges.*

Proof: Let $\{\mathbf{a}_k\}$ be a Cauchy sequence. By Theorem 23.7.15 there is some interval, $\prod_{i=1}^p [a_i, b_i]$ containing all the terms of $\{\mathbf{a}_k\}$. Therefore, by Theorem 23.7.18 a subsequence converges to a point of this interval. By Theorem 23.7.16 the original sequence converges. This proves the theorem.

23.7.4 Continuity And The Limit Of A Sequence

Just as in the case of a function of one variable, there is a very useful way of thinking of continuity in terms of limits of sequences found in the following theorem. In words, it says a function is continuous if it takes convergent sequences to convergent sequences whenever possible.

Theorem 23.7.20 *A function $\mathbf{f} : D(\mathbf{f}) \to \mathbb{R}^q$ is continuous at $\mathbf{x} \in D(\mathbf{f})$ if and only if, whenever $\mathbf{x}_n \to \mathbf{x}$ with $\mathbf{x}_n \in D(\mathbf{f})$, it follows $\mathbf{f}(\mathbf{x}_n) \to \mathbf{f}(\mathbf{x})$.*

Proof: Suppose first that $\mathbf{f}$ is continuous at $\mathbf{x}$ and let $\mathbf{x}_n \to \mathbf{x}$. Let $\varepsilon > 0$ be given. By continuity, there exists $\delta > 0$ such that if $|\mathbf{y} - \mathbf{x}| < \delta$, then $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \varepsilon$. However, there exists n_δ such that if $n \geq n_\delta$, then $|\mathbf{x}_n - \mathbf{x}| < \delta$ and so for all n this large,

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_n)| < \varepsilon$$

which shows $\mathbf{f}(\mathbf{x}_n) \to \mathbf{f}(\mathbf{x})$.

Now suppose the condition about taking convergent sequences to convergent sequences holds at $\mathbf{x}$. Suppose $\mathbf{f}$ fails to be continuous at $\mathbf{x}$. Then there exists $\varepsilon > 0$ and $\mathbf{x}_n \in D(f)$ such that $|\mathbf{x} - \mathbf{x}_n| < \frac{1}{n}$, yet

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_n)| \geq \varepsilon.$$

But this is clearly a contradiction because, although $\mathbf{x}_n \to \mathbf{x}$, $\mathbf{f}(\mathbf{x}_n)$ fails to converge to $\mathbf{f}(\mathbf{x})$. It follows $\mathbf{f}$ must be continuous after all. This proves the theorem.

23.8 Exercises

1. Suppose $\{\mathbf{x}_n\}$ is a sequence contained in a closed set, C which converges to $\mathbf{x}$. Show that $\mathbf{x} \in C$. **Hint:** Recall that a set is closed if and only if the complement of the set is open. That is if and only if $\mathbb{R}^n \setminus C$ is open.

2. Show using Problem 1 and Theorem 23.7.18 that every closed and bounded set is sequentially compact. **Hint:** If C is such a set, then $C \subseteq I_0 \equiv \prod_{i=1}^n [a_i, b_i]$. Now if $\{\mathbf{x}_n\}$ is a sequence in C, it must also be a sequence in I_0. Apply Problem 1 and Theorem 23.7.18.

3. Prove the extreme value theorem, a continuous function achieves its maximum and minimum on any closed and bounded set, C, using the result of Problem 2. **Hint:** Suppose $\lambda = \sup\{f(\mathbf{x}) : \mathbf{x} \in C\}$. Then there exists $\{\mathbf{x}_n\} \subseteq C$ such that $f(\mathbf{x}_n) \to \lambda$. Now select a convergent subsequence using Problem 2. Do the same for the minimum.

4. Let C be a closed and bounded set and suppose $\mathbf{f} : C \to \mathbb{R}^m$ is continuous. Show that $\mathbf{f}$ must also be **uniformly continuous.** This means: For every $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $\mathbf{x}, \mathbf{y} \in C$ and $|\mathbf{x} - \mathbf{y}| < \delta$, it follows $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| < \varepsilon$. This is a good time to review the definition of continuity so you will see the difference. **Hint:** Suppose it is not so. Then there exists $\varepsilon > 0$ and $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$ such that $|\mathbf{x}_k - \mathbf{y}_k| < \frac{1}{k}$ but $|\mathbf{f}(\mathbf{x}_k) - \mathbf{f}(\mathbf{y}_k)| \geq \varepsilon$. Now use Problem 2 to obtain a convergent subsequence.

5. Suppose every Cauchy sequence converges in $\mathbb{R}$. Show this implies the least upper bound axiom which is the usual way to state completeness for $\mathbb{R}$. Explain why the convergence of Cauchy sequences is equivalent to every nonempty set which is bounded above has a least upper bound in $\mathbb{R}$.

6. From Problem 2 every closed and bounded set is sequentially compact. Are these the only sets which are sequentially compact? Explain.

7. A set whose elements are open sets, $\mathcal{C}$ is called an **open cover** of H if $\cup\mathcal{C} \supseteq H$. In other words, $\mathcal{C}$ is an open cover of H if every point of H is in at least one set of $\mathcal{C}$. Show that if $\mathcal{C}$ is an open cover of a closed and bounded set H then there exists $\delta > 0$ such that whenever $\mathbf{x} \in H$, $B(\mathbf{x}, \delta)$ is contained in some set of $\mathcal{C}$. This number, δ is called a **Lebesgue number**. **Hint:** If there is no Lebesgue number for H, let $H \subseteq I = \prod_{i=1}^{n} [a_i, b_i]$. Use the process of chopping the intervals in half to get a sequence of nested intervals, I_k contained in I where $\operatorname{diam}(I_k) \leq 2^{-k} \operatorname{diam}(I)$ and there is no Lebesgue number for the open cover on $H_k \equiv H \cap I_k$. Now use the nested interval theorem to get $\mathbf{c}$ in all these H_k. For some $r > 0$ it follows $B(\mathbf{c}, r)$ is contained in some open set of U. But for large k, it must be that $H_k \subseteq B(\mathbf{c}, r)$ which contradicts the construction. You fill in the details.

8. A set is **compact** if for every open cover of the set, there exists a finite subset of the open cover which also covers the set. Show every closed and bounded set in $\mathbb{R}^p$ is compact. Next show that if a set in $\mathbb{R}^p$ is compact, then it must be closed and bounded. This is called the Heine Borel theorem.

9. Suppose S is a nonempty set in $\mathbb{R}^p$. Define

$$\operatorname{dist}(\mathbf{x},S) \equiv \inf\{|\mathbf{x} - \mathbf{y}| : \mathbf{y} \in S\}.$$

   Show that

$$|\operatorname{dist}(\mathbf{x},S) - \operatorname{dist}(\mathbf{y},S)| \leq |\mathbf{x} - \mathbf{y}|.$$

   Hint: Suppose $\operatorname{dist}(\mathbf{x}, S) < \operatorname{dist}(\mathbf{y}, S)$. If these are equal there is nothing to show. Explain why there exists $\mathbf{z} \in S$ such that $|\mathbf{x} - \mathbf{z}| < \operatorname{dist}(\mathbf{x},S) + \varepsilon$. Now explain why

$$|\operatorname{dist}(\mathbf{x},S) - \operatorname{dist}(\mathbf{y},S)| = \operatorname{dist}(\mathbf{y},S) - \operatorname{dist}(\mathbf{x},S) \leq |\mathbf{y} - \mathbf{z}| - (|\mathbf{x} - \mathbf{z}| - \varepsilon)$$

   Now use the triangle inequality and observe that ε is arbitrary.

10. Suppose H is a closed set and $H \subseteq U \subseteq \mathbb{R}^p$, an open set. Show there exists a continuous function defined on $\mathbb{R}^p$, f such that $f(\mathbb{R}^p) \subseteq [0, 1]$, $f(\mathbf{x}) = 0$ if $\mathbf{x} \notin U$ and

$f(\mathbf{x}) = 1$ if $\mathbf{x} \in H$. **Hint:** Try something like

$$\frac{\text{dist}\left(\mathbf{x}, U^C\right)}{\text{dist}\left(\mathbf{x}, U^C\right) + \text{dist}\left(\mathbf{x}, H\right)},$$

where $U^C \equiv \mathbb{R}^p \setminus U$, a closed set. You need to explain why the denominator is never equal to zero. The rest is supplied by Problem 9. This is a special case of a major theorem called Urysohn's lemma.

Vector Valued Functions Of One Variable

24.0.1 Outcomes

1. Identify a curve given its parameterization.
2. Determine combinations of vector functions such as sums, vector products, and scalar products.
3. Define limit, derivative, and integral for vector functions.
4. Evaluate limits, derivatives and integrals of vector functions.
5. Find the line tangent to a curve at a given point.
6. Recall, derive and apply rules to combinations of vector functions for the following:
   (a) limits
   (b) differentiation
   (c) integration
7. Describe what is meant by arc length.
8. Evaluate the arc length of a curve.
9. Evaluate the work done by a varying force over a curved path.

24.1 Limits Of A Vector Valued Function Of One Variable

The above discussion considered expressions like

$$\frac{\mathbf{f}(t_0+h)-\mathbf{f}(t_0)}{h}$$

and determined what they get close to as h gets small. In other words it is desired to consider

$$\lim_{h\to 0}\frac{\mathbf{f}(t_0+h)-\mathbf{f}(t_0)}{h}$$

Specializing to functions of one variable, one can give a meaning to

$$\lim_{s\to t+}\mathbf{f}(s),\ \lim_{s\to t-}\mathbf{f}(s),\ \lim_{s\to\infty}\mathbf{f}(s),$$

and

$$\lim_{s-\infty}\mathbf{f}(s).$$

Definition 24.1.1 *In the case where $D(\mathbf{f})$ is only assumed to satisfy $D(\mathbf{f})\supseteq(t,t+r)$,*

$$\lim_{s\to t+}\mathbf{f}(s)=\mathbf{L}$$

if and only if for all $\varepsilon>0$ there exists $\delta>0$ such that if

$$0<s-t<\delta,$$

then

$$|\mathbf{f}(s)-\mathbf{L}|<\varepsilon.$$

In the case where $D(\mathbf{f})$ is only assumed to satisfy $D(\mathbf{f})\supseteq(t-r,t)$,

$$\lim_{s\to t-}\mathbf{f}(s)=\mathbf{L}$$

if and only if for all $\varepsilon>0$ there exists $\delta>0$ such that if

$$0<t-s<\delta,$$

then

$$|\mathbf{f}(s)-\mathbf{L}|<\varepsilon.$$

One can also consider limits as a variable "approaches" infinity. Of course nothing is "close" to infinity and so this requires a slightly different definition.

$$\lim_{t\to\infty}\mathbf{f}(t)=\mathbf{L}$$

if for every $\varepsilon>0$ there exists l such that whenever $t>l$,

$$|\mathbf{f}(t)-\mathbf{L}|<\varepsilon \tag{24.1}$$

and

$$\lim_{t\to-\infty}\mathbf{f}(t)=\mathbf{L}$$

if for every $\varepsilon>0$ there exists l such that whenever $t<l$, (24.1) holds.

Note that in all of this the definitions are identical to the case of scalar valued functions. The only difference is that here $|\cdot|$ refers to the norm or length in $\mathbb{R}^p$ where maybe $p>1$.

Example 24.1.2 *Let $\mathbf{f}(t)=\left(\cos t,\sin t,t^2+1,\ln(t)\right)$. Find $\lim_{t\to\pi/2}\mathbf{f}(t)$.*

Use Theorem 23.4.7 on Page 545 and the continuity of the functions to write this limit equals

$$\begin{aligned}&\left(\lim_{t\to\pi/2}\cos t,\lim_{t\to\pi/2}\sin t,\lim_{t\to\pi/2}\left(t^2+1\right),\lim_{t\to\pi/2}\ln(t)\right)\\ =\ &\left(0,1,\ln\left(\frac{\pi^2}{4}+1\right),\ln\left(\frac{\pi}{2}\right)\right).\end{aligned}$$

Example 24.1.3 *Let $\mathbf{f}(t)=\left(\frac{\sin t}{t},t^2,t+1\right)$. Find $\lim_{t\to0}\mathbf{f}(t)$.*

Recall that $\lim_{t\to0}\frac{\sin t}{t}=1$. Then from Theorem 23.4.7 on Page 545, $\lim_{t\to0}\mathbf{f}(t)=(1,0,1)$.

24.2 The Derivative And Integral

The following definition is on the derivative and integral of a vector valued function of one variable.

Definition 24.2.1 *The derivative of a function,* $\mathbf{f}'(t)$, *is defined as the following limit whenever the limit exists. If the limit does not exist, then neither does* $\mathbf{f}'(t)$.

$$\lim_{h\to 0}\frac{\mathbf{f}(t+h)-\mathbf{f}(x)}{h}\equiv \mathbf{f}'(t)$$

The function of h *on the left is called the difference quotient just as it was for a scalar valued function. If* $\mathbf{f}(t)=(f_1(t),\cdots,f_p(t))$ *and* $\int_a^b f_i(t)\,dt$ *exists for each* $i=1,\cdots,p$, *then* $\int_a^b \mathbf{f}(t)\,dt$ *is defined as the vector,*

$$\left(\int_a^b f_1(t)\,dt,\cdots,\int_a^b f_p(t)\,dt\right).$$

This is what is meant by saying $\mathbf{f}\in R([a,b])$.

This is exactly like the definition for a scalar valued function. As before,

$$\mathbf{f}'(x)=\lim_{y\to x}\frac{\mathbf{f}(y)-\mathbf{f}(x)}{y-x}.$$

As in the case of a scalar valued function, differentiability implies continuity but not the other way around.

Theorem 24.2.2 *If* $\mathbf{f}'(t)$ *exists, then* $\mathbf{f}$ *is continuous at* t.

Proof: Suppose $\varepsilon>0$ is given and choose $\delta_1>0$ such that if $|h|<\delta_1$,

$$\left|\frac{\mathbf{f}(t+h)-\mathbf{f}(t)}{h}-\mathbf{f}'(t)\right|<1.$$

then for such h, the triangle inequality implies

$$|\mathbf{f}(t+h)-\mathbf{f}(t)|<|h|+|\mathbf{f}'(t)|\,|h|.$$

Now letting $\delta<\min\left(\delta_1,\frac{\varepsilon}{1+|\mathbf{f}'(x)|}\right)$ it follows if $|h|<\delta$, then

$$|\mathbf{f}(t+h)-\mathbf{f}(t)|<\varepsilon.$$

Letting $y=h+t$, this shows that if $|y-t|<\delta$,

$$|\mathbf{f}(y)-\mathbf{f}(t)|<\varepsilon$$

which proves $\mathbf{f}$ is continuous at t. This proves the theorem.

As in the scalar case, there is a fundamental theorem of calculus.

Theorem 24.2.3 *If* $\mathbf{f}\in R([a,b])$ *and if* $\mathbf{f}$ *is continuous at* $t\in(a,b)$, *then*

$$\frac{d}{dt}\left(\int_a^t \mathbf{f}(s)\,ds\right)=\mathbf{f}(t).$$

Proof: Say $\mathbf{f}(t) = (f_1(t), \cdots, f_p(t))$. Then it follows

$$\frac{1}{h}\int_a^{t+h} \mathbf{f}(s)\, ds - \frac{1}{h}\int_a^t \mathbf{f}(s)\, ds = \left(\frac{1}{h}\int_t^{t+h} f_1(s)\, ds, \cdots, \frac{1}{h}\int_t^{t+h} f_p(s)\, ds\right)$$

and $\lim_{h\to 0} \frac{1}{h}\int_t^{t+h} f_i(s)\, ds = f_i(t)$ for each $i = 1, \cdots, p$ from the fundamental theorem of calculus for scalar valued functions. Therefore,

$$\lim_{h\to 0} \frac{1}{h}\int_a^{t+h} \mathbf{f}(s)\, ds - \frac{1}{h}\int_a^t \mathbf{f}(s)\, ds = (f_1(t), \cdots, f_p(t)) = \mathbf{f}(t)$$

and this proves the claim.

Example 24.2.4 *Let* $\mathbf{f}(x) = \mathbf{c}$ *where* $\mathbf{c}$ *is a constant. Find* $\mathbf{f}'(x)$.

The difference quotient,

$$\frac{\mathbf{f}(x+h) - \mathbf{f}(x)}{h} = \frac{\mathbf{c} - \mathbf{c}}{h} = \mathbf{0}$$

Therefore,

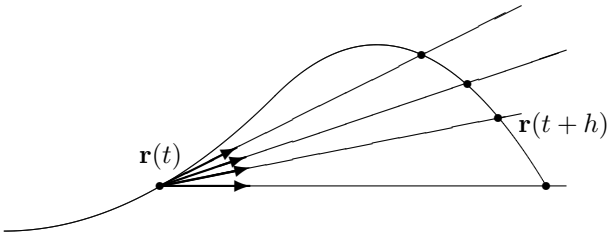
$$\lim_{h\to 0} \frac{\mathbf{f}(x+h) - \mathbf{f}(x)}{h} = \lim_{h\to 0} \mathbf{0} = \mathbf{0}$$

Example 24.2.5 *Let* $\mathbf{f}(t) = (at, bt)$ *where* a, b *are constants. Find* $\mathbf{f}'(t)$.

From the above discussion this derivative is just the vector valued functions whose components consist of the derivatives of the components of $\mathbf{f}$. Thus $\mathbf{f}'(t) = (a, b)$.

24.2.1 Geometric And Physical Significance Of The Derivative

Suppose $\mathbf{r}$ is a vector valued function of a parameter, t not necessarily time and consider the following picture of the points traced out by $\mathbf{r}$.



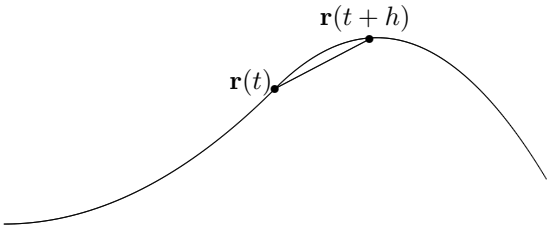
In this picture there are unit vectors in the direction of the vector from $\mathbf{r}(t)$ to $\mathbf{r}(t+h)$. You can see that it is reasonable to suppose these unit vectors, if they converge, converge to a unit vector, $\mathbf{T}$ which is tangent to the curve at the point $\mathbf{r}(t)$. Now each of these unit vectors is of the form

$$\frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{|\mathbf{r}(t+h) - \mathbf{r}(t)|} \equiv \mathbf{T}_h.$$

Thus $\mathbf{T}_h \to \mathbf{T}$, a unit tangent vector to the curve at the point $\mathbf{r}(t)$. Therefore,

$$\begin{aligned} \mathbf{r}'(t) &\equiv \lim_{h\to 0} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{h} = \lim_{h\to 0} \frac{|\mathbf{r}(t+h) - \mathbf{r}(t)|}{h} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{|\mathbf{r}(t+h) - \mathbf{r}(t)|} \\ &= \lim_{h\to 0} \frac{|\mathbf{r}(t+h) - \mathbf{r}(t)|}{h} \mathbf{T}_h = |\mathbf{r}'(t)|\, \mathbf{T}. \end{aligned}$$

In the case that t is time, the expression $|\mathbf{r}(t+h)-\mathbf{r}(t)|$ is a good approximation for the distance traveled by the object on the time interval $[t,t+h]$. The real distance would be the length of the curve joining the two points but if h is very small, this is essentially equal to $|\mathbf{r}(t+h)-\mathbf{r}(t)|$ as suggested by the picture below.



Therefore,

$$\frac{|\mathbf{r}(t+h)-\mathbf{r}(t)|}{h}$$

gives for small h, the approximate distance travelled on the time interval, $[t,t+h]$ divided by the length of time, h. Therefore, this expression is really the average speed of the object on this small time interval and so the limit as $h\to 0$, deserves to be called the instantaneous speed of the object. Thus $|\mathbf{r}'(t)|\,\mathbf{T}$ represents the speed times a unit direction vector, $\mathbf{T}$ which defines the direction in which the object is moving. Thus $\mathbf{r}'(t)$ is the velocity of the object. This is the physical significance of the derivative when t is time.

How do you go about computing $\mathbf{r}'(t)$? Letting $\mathbf{r}(t)=(r_1(t),\cdots,r_q(t))$, the expression

$$\frac{\mathbf{r}(t_0+h)-\mathbf{r}(t_0)}{h} \tag{24.2}$$

is equal to

$$\left(\frac{r_1(t_0+h)-r_1(t_0)}{h},\cdots,\frac{r_q(t_0+h)-r_q(t_0)}{h}\right).$$

Then as h converges to 0, (24.2) converges to

$$\mathbf{v}\equiv(v_1,\cdots,v_q)$$

where $v_k=r_k'(t)$. This by Theorem 23.4.7 on Page 545, which says that the term in (24.2) gets close to a vector, $\mathbf{v}$ if and only if all the coordinate functions of the term in (24.2) get close to the corresponding coordinate functions of $\mathbf{v}$.

In the case where t is time, this simply says the velocity vector equals the vector whose components are the derivatives of the components of the displacement vector, $\mathbf{r}(t)$.

In any case, the vector, $\mathbf{T}$ determines a direction vector which is tangent to the curve at the point, $\mathbf{r}(t)$ and so it is possible to find parametric equations for the line tangent to the curve at various points.

Example 24.2.6 *Let* $\mathbf{r}(t)=(\sin t,t^2,t+1)$ *for* $t\in[0,5]$. *Find a tangent line to the curve parameterized by* $\mathbf{r}$ *at the point* $\mathbf{r}(2)$.

From the above discussion, a direction vector has the same direction as $\mathbf{r}'(2)$. Therefore, it suffices to simply use $\mathbf{r}'(2)$ as a direction vector for the line. $\mathbf{r}'(2)=(\cos 2,4,1)$. Therefore, a parametric equation for the tangent line is

$$(\sin 2,4,3)+t(\cos 2,4,1)=(x,y,z).$$

Example 24.2.7 *Let* $\mathbf{r}(t) = \left(\sin t, t^2, t+1\right)$ *for* $t \in [0,5]$. *Find the velocity vector when* $t = 1$.

From the above discussion, this is simply $\mathbf{r}'(1) = (\cos 1, 2, 1)$.

24.2.2 Differentiation Rules

There are rules which relate the derivative to the various operations done with vectors such as the dot product, the cross product, and vector addition and scalar multiplication.

Theorem 24.2.8 *Let* $a, b \in \mathbb{R}$ *and suppose* $\mathbf{f}'(t)$ *and* $\mathbf{g}'(t)$ *exist. Then the following formulas are obtained.*

$$(a\mathbf{f} + b\mathbf{g})'(t) = a\mathbf{f}'(t) + b\mathbf{g}'(t). \tag{24.3}$$

$$(\mathbf{f}\cdot\mathbf{g})'(t) = \mathbf{f}'(t)\cdot\mathbf{g}(t) + \mathbf{f}(t)\cdot\mathbf{g}'(t) \tag{24.4}$$

If $\mathbf{f}, \mathbf{g}$ *have values in* $\mathbb{R}^3$, *then*

$$(\mathbf{f}\times\mathbf{g})'(t) = \mathbf{f}(t)\times\mathbf{g}'(t) + \mathbf{f}'(t)\times\mathbf{g}(t) \tag{24.5}$$

The formulas, (24.4), and (24.5) are referred to as the product rule.

Proof: The first formula is left for you to prove. Consider the second, (24.4).

$$\lim_{h\to 0}\frac{\mathbf{f}\cdot\mathbf{g}(t+h) - \mathbf{f}\mathbf{g}(t)}{h}$$

$$\begin{aligned}
&= \lim_{h\to 0}\frac{\mathbf{f}(t+h)\cdot\mathbf{g}(t+h) - \mathbf{f}(t+h)\cdot\mathbf{g}(t)}{h} + \frac{\mathbf{f}(t+h)\cdot\mathbf{g}(t) - \mathbf{f}(t)\cdot\mathbf{g}(t)}{h}\\
&= \lim_{h\to 0}\left(\mathbf{f}(t+h)\cdot\frac{(\mathbf{g}(t+h) - \mathbf{g}(t))}{h} + \frac{(\mathbf{f}(t+h) - \mathbf{f}(t))}{h}\cdot\mathbf{g}(t)\right)\\
&= \lim_{h\to 0}\sum_{k=1}^{n} f_k(t+h)\frac{(g_k(t+h) - g_k(t))}{h} + \sum_{k=1}^{n}\frac{(f_k(t+h) - f_k(t))}{h}g_k(t)\\
&= \sum_{k=1}^{n} f_k(t)\,g_k'(t) + \sum_{k=1}^{n} f_k'(t)\,g_k(t)\\
&= \mathbf{f}'(t)\cdot\mathbf{g}(t) + \mathbf{f}(t)\cdot\mathbf{g}'(t).
\end{aligned}$$

Formula (24.5) is left as an exercise which follows from the product rule and the definition of the cross product in terms of components given on Page 471.

Example 24.2.9 *Let*

$$\mathbf{r}(t) = \left(t^2, \sin t, \cos t\right)$$

and let $\mathbf{p}(t) = (t, \ln(t+1), 2t)$. *Find* $(\mathbf{r}(t)\times\mathbf{p}(t))'$.

From (24.5) this equals$(2t, \cos t, -\sin t)\times(t, \ln(t+1), 2t) + \left(t^2, \sin t, \cos t\right)\times\left(1, \frac{1}{t+1}, 2\right)$.

Example 24.2.10 *Let* $\mathbf{r}(t) = \left(t^2, \sin t, \cos t\right)$ *Find* $\int_0^\pi \mathbf{r}(t)\,dt$.

This equals $\left(\int_0^\pi t^2\,dt, \int_0^\pi \sin t\,\,dt, \int_0^\pi \cos t\,dt\right) = \left(\frac{1}{3}\pi^3, 2, 0\right)$.

Example 24.2.11 *An object has position* $\mathbf{r}(t) = \left(t^3, \frac{t}{1+1}, \sqrt{t^2+2}\right)$ *kilometers where* t *is given in hours. Find the velocity of the object in kilometers per hour when* $t = 1$.

Recall the velocity at time t was $\mathbf{r}'(t)$. Therefore, find $\mathbf{r}'(t)$ and plug in $t = 1$ to find the velocity.

$$\begin{aligned}\mathbf{r}'(t) &= \left(3t^2, \frac{1(1+t)-t}{(1+t)^2}, \frac{1}{2}\left(t^2+2\right)^{-1/2}2t\right)\\ &= \left(3t^2, \frac{1}{(1+t)^2}, \frac{1}{\sqrt{(t^2+2)}}t\right)\end{aligned}$$

When $t = 1$, the velocity is

$$\mathbf{r}'(1) = \left(3, \frac{1}{4}, \frac{1}{\sqrt{3}}\right) \text{ kilometers per hour.}$$

Obviously, this can be continued. That is, you can consider the possibility of taking the derivative of the derivative and then the derivative of that and so forth. The main thing to consider about this is the notation and it is exactly like it was in the case of a scalar valued function presented earlier. Thus $\mathbf{r}''(t)$ denotes the second derivative.

When you are given a vector valued function of one variable, sometimes it is possible to give a simple description of the curve which results. Usually it is not possible to do this!

Example 24.2.12 *Describe the curve which results from the vector valued function,* $\mathbf{r}(t) = (\cos 2t, \sin 2t, t)$ *where* $t \in \mathbb{R}$.

The first two components indicate that for $\mathbf{r}(t) = (x(t), y(t), z(t))$, the pair, $(x(t), y(t))$ traces out a circle. While it is doing so, $z(t)$ is moving at a steady rate in the positive direction. Therefore, the curve which results is a cork skrew shaped thing called a helix.

As an application of the theorems for differentiating curves, here is an interesting application. It is also a situation where the curve can be identified as something familiar.

Example 24.2.13 *Sound waves have the angle of incidence equal to the angle of reflection. Suppose you are in a large room and you make a sound. The sound waves spread out and you would expect your sound to be inaudible very far away. But what if the room were shaped so that the sound is reflected off the wall toward a single point, possibly far away from you? Then you might have the interesting phenomenon of someone far away hearing what you said quite clearly. How should the room be designed?*

Suppose you are located at the point $\mathbf{P}_0$ and the point where your sound is to be reflected is $\mathbf{P}_1$. Consider a plane which contains the two points and let $\mathbf{r}(t)$ denote a parameterization of the intersection of this plane with the walls of the room. Then the condition that the angle of reflection equals the angle of incidence reduces to saying the angle between $\mathbf{P}_0 - \mathbf{r}(t)$ and $-\mathbf{r}'(t)$ equals the angle between $\mathbf{P}_1 - \mathbf{r}(t)$ and $\mathbf{r}'(t)$. Draw a picture to see this. Therefore,

$$\frac{(\mathbf{P}_0 - \mathbf{r}(t)) \cdot (-\mathbf{r}'(t))}{|\mathbf{P}_0 - \mathbf{r}(t)|\,|\mathbf{r}'(t)|} = \frac{(\mathbf{P}_1 - \mathbf{r}(t)) \cdot (\mathbf{r}'(t))}{|\mathbf{P}_1 - \mathbf{r}(t)|\,|\mathbf{r}'(t)|}.$$

This reduces to

$$\frac{(\mathbf{r}(t) - \mathbf{P}_0) \cdot (-\mathbf{r}'(t))}{|\mathbf{r}(t) - \mathbf{P}_0|} = \frac{(\mathbf{r}(t) - \mathbf{P}_1) \cdot (\mathbf{r}'(t))}{|\mathbf{r}(t) - \mathbf{P}_1|} \tag{24.6}$$

Now

$$\frac{(\mathbf{r}(t)-\mathbf{P}_1)\cdot(\mathbf{r}'(t))}{|\mathbf{r}(t)-\mathbf{P}_1|}=\frac{d}{dt}|\mathbf{r}(t)-\mathbf{P}_1|$$

and a similar formula holds for $\mathbf{P}_1$ replaced with $\mathbf{P}_0$. This is because

$$|\mathbf{r}(t)-\mathbf{P}_1|=\sqrt{(\mathbf{r}(t)-\mathbf{P}_1)\cdot(\mathbf{r}(t)-\mathbf{P}_1)}$$

and so using the chain rule and product rule,

$$\begin{aligned}\frac{d}{dt}|\mathbf{r}(t)-\mathbf{P}_1| &= \frac{1}{2}((\mathbf{r}(t)-\mathbf{P}_1)\cdot(\mathbf{r}(t)-\mathbf{P}_1))^{-1/2}\,2\,((\mathbf{r}(t)-\mathbf{P}_1)\cdot\mathbf{r}'(t))\\ &= \frac{(\mathbf{r}(t)-\mathbf{P}_1)\cdot(\mathbf{r}'(t))}{|\mathbf{r}(t)-\mathbf{P}_1|}.\end{aligned}$$

Therefore, from (24.6),

$$\frac{d}{dt}(|\mathbf{r}(t)-\mathbf{P}_1|)+\frac{d}{dt}(|\mathbf{r}(t)-\mathbf{P}_0|)=0$$

showing that $|\mathbf{r}(t)-\mathbf{P}_1|+|\mathbf{r}(t)-\mathbf{P}_0|=C$ for some constant, C.This implies the curve of intersection of the plane with the room is an ellipse having $\mathbf{P}_0$ and $\mathbf{P}_1$ as the foci.

24.2.3 Leibniz's Notation

Leibniz's notation also generalizes routinely. For example, $\frac{d\mathbf{y}}{dt}=\mathbf{y}'(t)$ with other similar notations holding.

24.3 Product Rule For Matrices

Here is the concept of the product rule extended to matrix multiplication.

Definition 24.3.1 *Let $A(t)$ be an $m\times n$ matrix. Say $A(t)=(A_{ij}(t))$. Suppose also that $A_{ij}(t)$ is a differentiable function for all i,j. Then define $A'(t)\equiv\left(A'_{ij}(t)\right)$. That is, $A'(t)$ is the matrix which consists of replacing each entry by its derivative. Such an $m\times n$ matrix in which the entries are differentiable functions is called a differentiable matrix.*

The next lemma is just a version of the product rule.

Lemma 24.3.2 *Let $A(t)$ be an $m\times n$ matrix and let $B(t)$ be an $n\times p$ matrix with the property that all the entries of these matrices are differentiable functions. Then*

$$(A(t)B(t))'=A'(t)B(t)+A(t)B'(t).$$

Proof: $(A(t)B(t))'=\left(C'_{ij}(t)\right)$ where $C_{ij}(t)=A_{ik}(t)B_{kj}(t)$ and the repeated index summation convention is being used. Therefore,

$$\begin{aligned}C'_{ij}(t) &= A'_{ik}(t)B_{kj}(t)+A_{ik}(t)B'_{kj}(t)\\ &= (A'(t)B(t))_{ij}+(A(t)B'(t))_{ij}\\ &= (A'(t)B(t)+A(t)B'(t))_{ij}\end{aligned}$$

Therefore, the ij^{th} entry of $A(t)B(t)$ equals the ij^{th} entry of $A'(t)B(t)+A(t)B'(t)$ and this proves the lemma.

24.4 Moving Coordinate Systems

Let $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ be a right handed[1] orthonormal basis of vectors for each t. It is assumed these vectors are C^1 functions of t. Letting the positive x axis extend in the direction of $\mathbf{i}(t)$, the positive y axis extend in the direction of $\mathbf{j}(t)$, and the positive z axis extend in the direction of $\mathbf{k}(t)$, yields a moving coordinate system. Now let $\mathbf{u} = (u_1, u_2, u_3) \in \mathbb{R}^3$ and let t_0 be some reference time. For example you could let $t_0 = 0$. Then define the components of $\mathbf{u}$ with respect to these vectors, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ at time t_0 as

$$\mathbf{u} \equiv u_1 \mathbf{i}(t_0) + u_2 \mathbf{j}(t_0) + u_3 \mathbf{k}(t_0).$$

Let $\mathbf{u}(t)$ be defined as the vector which has the same components with respect to $\mathbf{i}, \mathbf{j}, \mathbf{k}$ but at time t. Thus

$$\mathbf{u}(t) \equiv u_1 \mathbf{i}(t) + u_2 \mathbf{j}(t) + u_3 \mathbf{k}(t).$$

and the vector has changed although the components have not.

For example, this is exactly the situation in the case of apparently fixed basis vectors on the earth if $\mathbf{u}$ is a position vector from the given spot on the earth's surface to a point regarded as fixed with the earth due to its keeping the same coordinates relative to coordinate axes which are fixed with the earth.

Now define a linear transformation $Q(t)$ mapping $\mathbb{R}^3$ to $\mathbb{R}^3$ by

$$Q(t)\mathbf{u} \equiv u_1 \mathbf{i}(t) + u_2 \mathbf{j}(t) + u_3 \mathbf{k}(t)$$

where

$$\mathbf{u} \equiv u_1 \mathbf{i}(t_0) + u_2 \mathbf{j}(t_0) + u_3 \mathbf{k}(t_0)$$

Thus letting $\mathbf{v}, \mathbf{u} \in \mathbb{R}^3$ be vectors and α, β, scalars,

$$Q(t)(\alpha\mathbf{u} + \beta\mathbf{v}) \equiv (\alpha u_1 + \beta v_1)\mathbf{i}(t) + (\alpha u_2 + \beta v_2)\mathbf{j}(t) + (\alpha u_3 + \beta v_3)\mathbf{k}(t)$$

$$\begin{aligned} &= (\alpha u_1 \mathbf{i}(t) + \alpha u_2 \mathbf{j}(t) + \alpha u_3 \mathbf{k}(t)) + (\beta v_1 \mathbf{i}(t) + \beta v_2 \mathbf{j}(t) + \beta v_3 \mathbf{k}(t)) \\ &= \alpha(u_1 \mathbf{i}(t) + u_2 \mathbf{j}(t) + u_3 \mathbf{k}(t)) + \beta(v_1 \mathbf{i}(t) + v_2 \mathbf{j}(t) + v_3 \mathbf{k}(t)) \\ &\equiv \alpha Q(t)\mathbf{u} + \beta Q(t)\mathbf{v} \end{aligned}$$

showing that $Q(t)$ is a linear transformation. Also, $Q(t)$ preserves all distances because, since the vectors, $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ form an orthonormal set,

$$|Q(t)\mathbf{u}| = \left(\sum_{i=1}^{3} \left(u^i\right)^2 \right)^{1/2} = |\mathbf{u}|.$$

For simplicity, let $\mathbf{i}(t) = \mathbf{e}_1(t), \mathbf{j}(t) = \mathbf{e}_2(t), \mathbf{k}(t) = \mathbf{e}_3(t)$ and $\mathbf{i}(t_0) = \mathbf{e}_1(t_0), \mathbf{j}(t_0) = \mathbf{e}_2(t_0), \mathbf{k}(t_0) = \mathbf{e}_3(t_0)$. Then using the repeated index summation convention,

$$\mathbf{u}(t) = u_j \mathbf{e}_j(t) = u_j \mathbf{e}_j(t) \cdot \mathbf{e}_i(t_0)\, \mathbf{e}_i(t_0)$$

and so with respect to the basis, $\mathbf{i}(t_0) = \mathbf{e}_1(t_0), \mathbf{j}(t_0) = \mathbf{e}_2(t_0), \mathbf{k}(t_0) = \mathbf{e}_3(t_0)$, the matrix of $Q(t)$ is

$$Q_{ij}(t) = \mathbf{e}_i(t_0) \cdot \mathbf{e}_j(t)$$

Recall this means you take a vector, $\mathbf{u} \in \mathbb{R}^3$ which is a list of the components of $\mathbf{u}$ with respect to $\mathbf{i}(t_0), \mathbf{j}(t_0), \mathbf{k}(t_0)$ and when you multiply by $Q(t)$ you get the components of $\mathbf{u}(t)$ with respect to $\mathbf{i}(t_0), \mathbf{j}(t_0), \mathbf{k}(t_0)$. I will refer to this matrix as $Q(t)$ to save notation.

[1] Recall that right handed implies $\mathbf{i} \times \mathbf{j} = \mathbf{k}$.

Lemma 24.4.1 *Suppose $Q(t)$ is a real, differentiable $n \times n$ matrix which preserves distances. Then $Q(t) Q(t)^T = Q(t)^T Q(t) = I$. Also, if $\mathbf{u}(t) \equiv Q(t)\mathbf{u}$, then there exists a vector, $\mathbf{\Omega}(t)$ such that*

$$\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t).$$

Proof: Recall that $(\mathbf{z} \cdot \mathbf{w}) = \frac{1}{4}\left(|\mathbf{z}+\mathbf{w}|^2 - |\mathbf{z}-\mathbf{w}|^2\right)$. Therefore,

$$\begin{aligned}(Q(t)\mathbf{u} \cdot Q(t)\mathbf{w}) &= \frac{1}{4}\left(|Q(t)(\mathbf{u}+\mathbf{w})|^2 - |Q(t)(\mathbf{u}-\mathbf{w})|^2\right) \\ &= \frac{1}{4}\left(|\mathbf{u}+\mathbf{w}|^2 - |\mathbf{u}-\mathbf{w}|^2\right) \\ &= (\mathbf{u}\cdot\mathbf{w}).\end{aligned}$$

This implies

$$\left(Q(t)^T Q(t)\mathbf{u} \cdot \mathbf{w}\right) = (\mathbf{u}\cdot\mathbf{w})$$

for all $\mathbf{u}, \mathbf{w}$. Therefore, $Q(t)^T Q(t)\mathbf{u} = \mathbf{u}$ and so $Q(t)^T Q(t) = Q(t) Q(t)^T = I$. This proves the first part of the lemma.

It follows from the product rule, Lemma 24.3.2 that

$$Q'(t) Q(t)^T + Q(t) Q'(t)^T = 0$$

and so

$$Q'(t) Q(t)^T = -\left(Q'(t) Q(t)^T\right)^T. \tag{24.7}$$

From the definition, $Q(t)\mathbf{u} = \mathbf{u}(t)$,

$$\mathbf{u}'(t) = Q'(t)\mathbf{u} = Q'(t)\overbrace{Q(t)^T\mathbf{u}(t)}^{=\mathbf{u}}.$$

Then writing the matrix of $Q'(t) Q(t)^T$ with respect to $\mathbf{i}(t_0), \mathbf{j}(t_0), \mathbf{k}(t_0)$, it follows from (24.7) that the matrix of $Q'(t) Q(t)^T$ is of the form

$$\begin{pmatrix} 0 & -\omega_3(t) & \omega_2(t) \\ \omega_3(t) & 0 & -\omega_1(t) \\ -\omega_2(t) & \omega_1(t) & 0 \end{pmatrix}$$

for some time dependent scalars, ω_i. Therefore,

$$\begin{aligned}\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}'(t) &= \begin{pmatrix} 0 & -\omega_3(t) & \omega_2(t) \\ \omega_3(t) & 0 & -\omega_1(t) \\ -\omega_2(t) & \omega_1(t) & 0 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}(t) \\ &= \begin{pmatrix} w_2(t)u_3(t) - w_3(t)u_2(t) \\ w_3(t)u_1(t) - w_1(t)u_3(t) \\ w_1(t)u_2(t) - w_2(t)u_1(t) \end{pmatrix}\end{aligned}$$

where the u_i are the components of the vector $\mathbf{u}(t)$ in terms of the fixed vectors $\mathbf{i}(t_0), \mathbf{j}(t_0), \mathbf{k}(t_0)$. Therefore,

$$\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t) = Q'(t) Q(t)^T \mathbf{u}(t) \tag{24.8}$$

where

$$\mathbf{\Omega}(t) = \omega_1(t)\mathbf{i}(t_0) + \omega_2(t)\mathbf{j}(t_0) + \omega_3(t)\mathbf{k}(t_0).$$

because

$$\boldsymbol{\Omega}(t)\times\mathbf{u}(t)\equiv\begin{vmatrix}\mathbf{i}(t_0) & \mathbf{j}(t_0) & \mathbf{k}(t_0)\\ w_1 & w_2 & w_3\\ u_1 & u_2 & u_3\end{vmatrix}\equiv$$

$$\mathbf{i}(t_0)(w_2u_3-w_3u_2)+\mathbf{j}(t_0)(w_3u_1-w_1u_3)+\mathbf{k}(t_0)(w_1u_2-w_2u_1).$$

This proves the lemma and yields the existence part of the following theorem.

Theorem 24.4.2 *Let $\mathbf{i}(t),\mathbf{j}(t),\mathbf{k}(t)$ be as described. Then there exists a unique vector $\boldsymbol{\Omega}(t)$ such that if $\mathbf{u}(t)$ is a vector whose components are constant with respect to $\mathbf{i}(t),\mathbf{j}(t),\mathbf{k}(t)$, then*

$$\mathbf{u}'(t)=\boldsymbol{\Omega}(t)\times\mathbf{u}(t).$$

Proof: It only remains to prove uniqueness. Suppose $\boldsymbol{\Omega}_1$ also works. Then $\mathbf{u}(t)=Q(t)\mathbf{u}$ and so $\mathbf{u}'(t)=Q'(t)\mathbf{u}$ and

$$Q'(t)\mathbf{u}=\boldsymbol{\Omega}\times Q(t)\mathbf{u}=\boldsymbol{\Omega}_1\times Q(t)\mathbf{u}$$

for all $\mathbf{u}$. Therefore,

$$(\boldsymbol{\Omega}-\boldsymbol{\Omega}_1)\times Q(t)\mathbf{u}=\mathbf{0}$$

for all $\mathbf{u}$ and since $Q(t)$ is one to one and onto, this implies $(\boldsymbol{\Omega}-\boldsymbol{\Omega}_1)\times\mathbf{w}=\mathbf{0}$ for all $\mathbf{w}$ and thus $\boldsymbol{\Omega}-\boldsymbol{\Omega}_1=\mathbf{0}$. This proves the theorem.

Definition 24.4.3 *A **rigid body** in $\mathbb{R}^3$ has a moving coordinate system with the property that for an observer on the rigid body, the vectors, $\mathbf{i}(t),\mathbf{j}(t),\mathbf{k}(t)$ are constant. More generally, a vector $\mathbf{u}(t)$ is said to be fixed with the body if to a person on the body, the vector appears to have the same magnitude and same direction independent of t. Thus $\mathbf{u}(t)$ is fixed with the body if $\mathbf{u}(t)=u_1\mathbf{i}(t)+u_2\mathbf{j}(t)+u_3\mathbf{k}(t)$.*

The following comes from the above discussion.

Theorem 24.4.4 *Let $B(t)$ be the set of points in three dimensions occupied by a rigid body. Then there exists a vector $\boldsymbol{\Omega}(t)$ such that whenever $\mathbf{u}(t)$ is fixed with the rigid body,*

$$\mathbf{u}'(t)=\boldsymbol{\Omega}(t)\times\mathbf{u}(t).$$

24.5 Exercises

1. Find the following limits if possible
   (a) $\lim_{x\to 0+}\left(\frac{|x|}{x},\sin x/x,\cos x\right)$
   (b) $\lim_{x\to 0+}\left(\frac{x}{|x|},\sec x,e^x\right)$
   (c) $\lim_{x\to 4}\left(\frac{x^2-16}{x+4},x+7,\frac{\tan 4x}{5x}\right)$
   (d) $\lim_{x\to\infty}\left(\frac{x}{1+x^2},\frac{x^2}{1+x^2},\frac{\sin x^2}{x}\right)$
2. Find $\lim_{x\to 2}\left(\frac{x^2-4}{x+2},x^2+2x-1,\frac{x^2-4}{x-2}\right)$.

3. Prove from the definition that $\lim_{x\to a}(\sqrt[3]{x}, x+1) = (\sqrt[3]{a}, a+1)$ for all $a \in \mathbb{R}$. **Hint:** You might want to use the formula for the difference of two cubes,
$$a^3 - b^3 = (a-b)\left(a^2 + ab + b^2\right).$$

4. Let $\mathbf{r}(t) = \left(4 + (t-1)^2, \sqrt{t^2+1}\,(t-1)^3, \frac{(t-1)^3}{t^5}\right)$ describe the position of an object in $\mathbb{R}^3$ as a function of t where t is measured in seconds and $\mathbf{r}(t)$ is measured in meters. Is the velocity of this object ever equal to zero? If so, find the value of t at which this occurs and the point in $\mathbb{R}^3$ at which the velocity is zero.

5. Let $\mathbf{r}(t) = \left(\sin 2t, t^2, 2t+1\right)$ for $t \in [0,4]$. Find a tangent line to the curve parameterized by $\mathbf{r}$ at the point $\mathbf{r}(2)$.

6. Let $\mathbf{r}(t) = \left(t, \sin t^2, t+1\right)$ for $t \in [0,5]$. Find a tangent line to the curve parameterized by $\mathbf{r}$ at the point $\mathbf{r}(2)$.

7. Let $\mathbf{r}(t) = \left(\sin t, t^2, \cos\left(t^2\right)\right)$ for $t \in [0,5]$. Find a tangent line to the curve parameterized by $\mathbf{r}$ at the point $\mathbf{r}(2)$.

8. Let $\mathbf{r}(t) = \left(\sin t, \cos\left(t^2\right), t+1\right)$ for $t \in [0,5]$. Find the velocity when $t = 3$.

9. Let $\mathbf{r}(t) = \left(\sin t, t^2, t+1\right)$ for $t \in [0,5]$. Find the velocity when $t = 3$.

10. Let $\mathbf{r}(t) = \left(t, \ln\left(t^2+1\right), t+1\right)$ for $t \in [0,5]$. Find the velocity when $t = 3$.

11. Suppose an object has position $\mathbf{r}(t) \in \mathbb{R}^3$ where $\mathbf{r}$ is differentiable and suppose also that $|\mathbf{r}(t)| = c$ where c is a constant.

    (a) Show first that this condition does not require $\mathbf{r}(t)$ to be a constant. **Hint:** You can do this either mathematically or by giving a physical example.

    (b) Show that you can conclude that $\mathbf{r}'(t) \cdot \mathbf{r}(t) = 0$. That is, the velocity is always perpendicular to the displacement.

12. Prove (24.5) from the component description of the cross product.

13. Prove (24.5) from the formula $(\mathbf{f} \times \mathbf{g})_i = \varepsilon_{ijk} f_j g_k$.

14. Prove (24.5) directly from the definition of the derivative without considering components.

15. A bezier curve in $\mathbb{R}^n$ is a vector valued function of the form
$$\mathbf{y}(t) = \sum_{k=0}^{n} \binom{n}{k} \mathbf{x}_k (1-t)^{n-k} t^k$$
where here the $\binom{n}{k}$ are the binomial coefficients and $\mathbf{x}_k$ are $n+1$ points in $\mathbb{R}^n$. Show that $\mathbf{y}(0) = \mathbf{x}_0$, $\mathbf{y}(1) = \mathbf{x}_n$, and find $\mathbf{y}'(0)$ and $\mathbf{y}'(1)$. Recall that $\binom{n}{0} = \binom{n}{n} = 1$ and $\binom{n}{n-1} = \binom{n}{1} = n$. Curves of this sort are important in various computer programs.

16. Suppose $\mathbf{r}(t)$, $\mathbf{s}(t)$, and $\mathbf{p}(t)$ are three differentiable functions of t which have values in $\mathbb{R}^3$. Find a formula for $(\mathbf{r}(t) \times \mathbf{s}(t) \cdot \mathbf{p}(t))'$.

17. If $\mathbf{r}'(t) = \mathbf{0}$ for all $t \in (a,b)$, show there exists a constant vector, $\mathbf{c}$ such that $\mathbf{r}(t) = \mathbf{c}$ for all $t \in (a,b)$.

18. If $\mathbf{F}'(t) = \mathbf{f}(t)$ for all $t \in (a,b)$ and $\mathbf{F}$ is continuous on $[a,b]$, show $\int_a^b \mathbf{f}(t)\,dt = \mathbf{F}(b) - \mathbf{F}(a)$.

19. Verify that if $\mathbf{\Omega} \times \mathbf{u} = \mathbf{0}$ for all $\mathbf{u}$, then $\mathbf{\Omega} = \mathbf{0}$.

20. Verify that if $\mathbf{u} \neq \mathbf{0}$ and $\mathbf{v} \cdot \mathbf{u} = 0$ and both $\mathbf{\Omega}$ and $\mathbf{\Omega}_1$ satisfy $\mathbf{\Omega} \times \mathbf{u} = \mathbf{v}$, then $\mathbf{\Omega}_1 = \mathbf{\Omega}$.

24.6 Newton's Laws Of Motion

Definition 24.6.1 *Let $\mathbf{r}(t)$ denote the position of an object. Then the acceleration of the object is defined to be $\mathbf{r}''(t)$.*

Newton's[2] first law is: "Every body persists in its state of rest or of uniform motion in a straight line unless it is compelled to change that state by forces impressed on it."

Newton's second law is:

$$\mathbf{F} = m\mathbf{a} \tag{24.9}$$

where $\mathbf{a}$ is the acceleration and m is the mass of the object.

Newton's third law states: "To every action there is always opposed an equal reaction; or, the mutual actions of two bodies upon each other are always equal, and directed to contrary parts."

Of these laws, only the second two are independent of each other, the first law being implied by the second. The third law says roughly that if you apply a force to something, the thing applies the same force back.

The second law is the one of most interest. Note that the statement of this law depends on the concept of the derivative because the acceleration is defined as a derivative. Newton used calculus and these laws to solve profound problems involving the motion of the planets and other problems in mechanics. The next example involves the concept that if you know the force along with the initial velocity and initial position, then you can determine the position.

Example 24.6.2 *Let $\mathbf{r}(t)$ denote the position of an object of mass 2 kilogram at time t and suppose the force acting on the object is given by $\mathbf{F}(t) = \left(t, 1-t^2, 2e^{-t}\right)$. Suppose $\mathbf{r}(0) = (1,0,1)$ meters, and $\mathbf{r}'(0) = (0,1,1)$ meters/sec. Find $\mathbf{r}(t)$.*

By Newton's second law, $2\mathbf{r}''(t) = \mathbf{F}(t) = \left(t, 1-t^2, 2e^{-t}\right)$ and so

$$\mathbf{r}''(t) = \left(t/2, \left(1-t^2\right)/2, e^{-t}\right).$$

Therefore the velocity is given by

$$\mathbf{r}'(t) = \left(\frac{t^2}{4}, \frac{t - t^3/3}{2}, -e^{-t}\right) + \mathbf{c}$$

[2]Isaac Newton 1642-1727 is often credited with inventing calculus although this is not correct since most of the ideas were in existence earlier. However, he made major contributions to the subject partly in order to study physics and astronomy. He formulated the laws of gravity, made major contributions to optics, and stated the fundamental laws of mechanics listed here. He invented a version of the binomial theorem when he was only 23 years old and built a reflecting telescope. He showed that Kepler's laws for the motion of the planets came from calculus and his laws of gravitation. In 1686 he published an important book, Principia, in which many of his ideas are found. Newton was also very interested in theology and had strong views on the nature of God which were based on his study of the Bible and early Christian writings. He finished his life as Master of the Mint.

where $\mathbf{c}$ is a constant vector which must be determined from the initial condition given for the velocity. Thus letting $\mathbf{c} = (c_1, c_2, c_3)$,

$$(0,1,1) = (0,0,-1) + (c_1, c_2, c_3)$$

which requires $c_1 = 0, c_2 = 1$, and $c_3 = 2$. Therefore, the velocity is found.

$$\mathbf{r}'(t) = \left(\frac{t^2}{4}, \frac{t - t^3/3}{2} + 1, -e^{-t} + 2 \right).$$

Now from this, the displacement must equal

$$\mathbf{r}(t) = \left(\frac{t^3}{12}, \frac{t^2/2 - t^4/12}{2} + t, e^{-t} + 2t \right) + (C_1, C_2, C_3)$$

where the constant vector, (C_1, C_2, C_3) must be determined from the initial condition for the displacement. Thus

$$\mathbf{r}(0) = (1,0,1) = (0,0,1) + (C_1, C_2, C_3)$$

which means $C_1 = 1, C_2 = 0$, and $C_3 = 0$. Therefore, the displacement has also been found.

$$\mathbf{r}(t) = \left(\frac{t^3}{12} + 1, \frac{t^2/2 - t^4/12}{2} + t, e^{-t} + 2t \right) \text{ meters.}$$

Actually, in applications of this sort of thing acceleration does not usually come to you as a nice given function written in terms of simple functions you understand. Rather, it comes as measurements taken by instruments and the position is continuously being updated based on this information. Another situation which often occurs is the case when the forces on the object depend not just on time but also on the position or velocity of the object.

Example 24.6.3 *An artillery piece is fired at ground level on a level plain. The angle of elevation is $\pi/6$ radians and the speed of the shell is 400 meters per second. How far does the shell fly before hitting the ground?*

Neglect air resistance in this problem. Also let the direction of flight be along the positive x axis. Thus the initial velocity is the vector, $400\cos(\pi/6)\,\mathbf{i} + 400\sin(\pi/6)\,\mathbf{j}$ while the only force experienced by the shell after leaving the artillery piece is the force of gravity, $-mg\mathbf{j}$ where m is the mass of the shell. The acceleration of gravity equals 9.8 meters per sec^2 and so the following needs to be solved.

$$m\mathbf{r}''(t) = -mg\mathbf{j},\ \mathbf{r}(0) = (0,0), \mathbf{r}'(0) = 400\cos(\pi/6)\,\mathbf{i} + 400\sin(\pi/6)\,\mathbf{j}.$$

Denoting $\mathbf{r}(t)$ as $(x(t), y(t))$,

$$x''(t) = 0,\ y''(t) = -g.$$

Therefore, $y'(t) = -gt + C$ and from the information on the initial velocity, $C = 400\sin(\pi/6) = 200$. Thus

$$y(t) = -4.9t^2 + 200t + D.$$

$D = 0$ because the artillery piece is fired at ground level which requires both x and y to equal zero at this time. Similarly, $x'(t) = 400\cos(\pi/6)$ so $x(t) = 400\cos(\pi/6)\,t = 200\sqrt{3}t$. The shell hits the ground when $y = 0$ and this occurs when $-4.9t^2 + 200t = 0$. Thus $t = 40.816\,326\,530\,6$ seconds and so at this time,

$$x = 200\sqrt{3}\,(40.816\,326\,530\,6) = 14139.190\,265\,9 \text{ meters.}$$

The next example is more complicated because it also takes in to account air resistance. We do not live in a vacume.

Example 24.6.4 *A lump of "blue ice" escapes the lavatory of a jet flying at 600 miles per hour at an altitude of 30,000 feet. This blue ice weighs 64 pounds near the earth and experiences a force of air resistance equal to* $(-.1)\,\mathbf{r}'(t)$ *pounds. Find the position and velocity of the blue ice as a function of time measured in seconds. Also find the velocity when the lump hits the ground. Such lumps have been known to surprise people on the ground.*

The first thing needed is to obtain information which involves consistent units. The blue ice weighs 32 pounds near the earth. Thus 32 pounds is the force exerted by gravity on the lump and so its mass must be given by Newton's second law as follows.

$$64 = m \times 32.$$

Thus $m = 2$ slugs. The slug is the unit of mass in the system involving feet and pounds. The jet is flying at 600 miles per hour. I want to change this to feet per second. Thus it flies at

$$\frac{600 \times 5280}{60 \times 60} = 880 \text{ feet per second.}$$

The explanation for this is that there are 5280 feet in a mile and so it goes 600×5280 feet in one hour. There are 60×60 seconds in an hour. The position of the lump of blue ice will be computed from a point on the ground directly beneath the airplane at the instant the blue ice escapes and regard the airplane as moving in the direction of the positive x axis. Thus the initial displacement is

$$\mathbf{r}(0) = (0, 30000) \text{ feet}$$

and the initial velocity is

$$\mathbf{r}'(0) = (880, 0) \text{ feet/sec.}$$

The force of gravity is

$$(0, -64) \text{ pounds}$$

and the force due to air resistance is

$$(-.1)\,\mathbf{r}'(t) \text{ pounds.}$$

Newtons second law yields the following initial value problem for $\mathbf{r}(t) = (r_1(t), r_2(t))$.

$$\begin{aligned} &2\left(r_1''(t), r_2''(t)\right) = (-.1)\left(r_1'(t), r_2'(t)\right) + (0, -64), \ \left(r_1(0), r_2(0)\right) = (0, 30000), \\ &\left(r_1'(0), r_2'(0)\right) = (880, 0) \end{aligned}$$

Therefore,

$$\begin{aligned} 2r_1''(t) + (.1)\,r_1'(t) &= 0 \\ 2r_2''(t) + (.1)\,r_2'(t) &= -64 \\ r_1(0) &= 0 \\ r_2(0) &= 30000 \\ r_1'(0) &= 880 \\ r_2'(0) &= 0 \end{aligned} \tag{24.10}$$

To save on repetition solve

$$mr'' + kr' = c, r(0) = u, r'(0) = v.$$

Divide the differential equation by m and get

$$r'' + (k/m)\,r' = c/m.$$

Now multiply both sides by $e^{(k/m)t}$. You should check this gives

$$\frac{d}{dt}\left(e^{(k/m)t}r'\right) = (c/m)\,e^{(k/m)t}$$

Therefore,

$$e^{(k/m)t}r' = \frac{1}{k}e^{\frac{k}{m}t}c + C$$

and using the initial condition, $v = c/k + C$ and so

$$r'(t) = (c/k) + (v - (c/k))\,e^{-\frac{k}{m}t}$$

Now this implies

$$r(t) = (c/k)\,t - \frac{1}{k}me^{-\frac{k}{m}t}\left(v - \frac{c}{k}\right) + D \tag{24.11}$$

where D is a constant to be determined from the initial conditions. Thus

$$u = -\frac{m}{k}\left(v - \frac{c}{k}\right) + D$$

and so

$$r(t) = (c/k)\,t - \frac{1}{k}me^{-\frac{k}{m}t}\left(v - \frac{c}{k}\right) + \left(u + \frac{m}{k}\left(v - \frac{c}{k}\right)\right).$$

Now apply this to the system (24.10) to find

$$\begin{aligned} r_1(t) &= -\frac{1}{(.1)}2\left(\exp\left(\frac{-(.1)}{2}t\right)\right)(880) + \left(\frac{2}{(.1)}(880)\right) \\ &= -17600.0\exp(-.0\,5t) + 17600.0 \end{aligned}$$

and

$$\begin{aligned} r_2(t) &= (-64/(.1))\,t - \frac{1}{(.1)}2\left(\exp\left(-\frac{(.1)}{2}t\right)\right)\left(\frac{64}{(.1)}\right) + \left(30000 + \frac{2}{(.1)}\left(\frac{64}{(.1)}\right)\right) \\ &= -640.0t - 12800.0\exp(-.0\,5t) + 42800.0 \end{aligned}$$

This gives the coordinates of the position. What of the velocity? Using (24.11) in the same way to obtain the velocity,

$$\begin{aligned} r_1'(t) &= 880.0\exp(-.0\,5t), \\ r_2'(t) &= -640.0 + 640.0\exp(-.0\,5t). \end{aligned} \tag{24.12}$$

To determine the velocity when the blue ice hits the ground, it is necessary to find the value of t when this event takes place and then to use (24.12) to determine the velocity. It hits ground when $r_2(t) = 0$. Thus it suffices to solve the equation,

$$0 = -640.0t - 12800.0\exp(-.0\,5t) + 42800.0.$$

This is a fairly hard equation to solve using the methods of algebra. In fact, I do not have a good way to find this value of t using algebra. However if plugging in various values of t using a calculator you eventually find that when $t = 66.14$,

$$-640.0(66.14) - 12800.0\exp(-.0\,5(66.14)) + 42800.0 = 1.588 \text{ feet.}$$

This is close enough to hitting the ground and so plugging in this value for t yields the approximate velocity,

$$(880.0\exp(-.0\,5(66.14)), -640.0 + 640.0\exp(-.0\,5(66.14))) = (32.23, -616.56).$$

Notice how because of air resistance the component of velocity in the horizontal direction is only about 32 feet per second even though this component started out at 880 feet per second while the component in the vertical direction is -616 feet per second even though this component started off at 0 feet per second. You see that air resistance can be very important so it is not enough to pretend, as is often done in beginning physics courses that everything takes place in a vacuum. Actually, this problem used several physical simplifications. It was assumed the force acting on the lump of blue ice by gravity was constant. This is not really true because it actually depends on the distance between the center of mass of the earth and the center of mass of the lump. It was also assumed the air resistance is proportional to the velocity. This is an over simplification when high speeds are involved. However, increasingly correct models can be studied in a systematic way as above.

24.6.1 Kinetic Energy

Newton's second law is also the basis for the notion of kinetic energy. When a force is exerted on an object which causes the object to move, it follows that the force is doing work which manifests itself in a change of velocity of the object. How is the total work done on the object by the force related to the final velocity of the object? By Newton's second law, and letting $\mathbf{v}$ be the velocity,

$$\mathbf{F}(t) = m\mathbf{v}'(t).$$

Now in a small increment of time, $(t, t+dt)$, the work done on the object would be approximately equal to

$$dW = \mathbf{F}(t) \cdot \mathbf{v}(t)\, dt. \tag{24.13}$$

If no work has been done at time $t = 0$,then (24.13) implies

$$\frac{dW}{dt} = \mathbf{F} \cdot \mathbf{v},\ W(0) = 0.$$

Hence,

$$\frac{dW}{dt} = m\mathbf{v}'(t) \cdot \mathbf{v}(t) = \frac{m}{2}\frac{d}{dt}|\mathbf{v}(t)|^2.$$

Therefore, the total work done up to time t would be $W(t) = \frac{m}{2}|\mathbf{v}(t)|^2 - \frac{m}{2}|\mathbf{v}_0|^2$ where $|\mathbf{v}_0|$ denotes the initial speed of the object. This difference represents the change in the kinetic energy.

24.6.2 Impulse And Momentum

Work and energy involve a force acting on an object for some distance. Impulse involves a force which acts on an object for an interval of time.

Definition 24.6.5 *Let $\mathbf{F}$ be a force which acts on an object during the time interval, $[a, b]$. The impulse of this force is*

$$\int_a^b \mathbf{F}(t)\ dt.$$

This is defined as

$$\left(\int_a^b F_1(t)\ dt, \int_a^b F_2(t)\ dt, \int_a^b F_3(t)\ dt\right).$$

The linear momentum of an object of mass m and velocity $\mathbf{v}$ is defined as

$$\textit{Linear momentum} = m\mathbf{v}.$$

The notion of impulse and momentum are related in the following theorem.

Theorem 24.6.6 *Let* $\mathbf{F}$ *be a force acting on an object of mass* m*. Then the impulse equals the change in momentum. More precisely,*

$$\int_a^b \mathbf{F}(t)\, dt = m\mathbf{v}(b) - m\mathbf{v}(a).$$

Proof: This is really just the fundamental theorem of calculus and Newton's second law applied to the components of $\mathbf{F}$.

$$\int_a^b \mathbf{F}(t)\, dt = \int_a^b m\frac{d\mathbf{v}}{dt}dt = m\mathbf{v}(b) - m\mathbf{v}(a) \tag{24.14}$$

Now suppose two point masses, A and B collide. Newton's third law says the force exerted by mass A on mass B is equal in magnitude but opposite in direction to the force exerted by mass B on mass A. Letting the collision take place in the time interval, $[a,b]$ and denoting the two masses by m_A and m_B and their velocities by $\mathbf{v}_A$ and $\mathbf{v}_B$ it follows that

$$m_A\mathbf{v}_A(b) - m_A\mathbf{v}_A(a) = \int_a^b (\text{Force of } B \text{ on } A)\, dt$$

and

$$\begin{aligned} m_B\mathbf{v}_B(b) - m_B\mathbf{v}_B(a) &= \int_a^b (\text{Force of } A \text{ on } B)\, dt \\ &= -\int_a^b (\text{Force of } B \text{ on } A)\, dt \\ &= -(m_A\mathbf{v}_A(b) - m_A\mathbf{v}_A(a)) \end{aligned}$$

and this shows

$$m_B\mathbf{v}_B(b) + m_A\mathbf{v}_A(b) = m_B\mathbf{v}_B(a) + m_A\mathbf{v}_A(a).$$

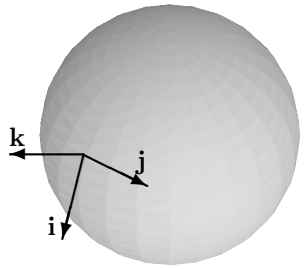
In other words, in a collision between two masses the total linear momentum before the collision equals the total linear momentum after the collision. This is known as the conservation of linear momentum.

24.7 Acceleration With Respect To Moving Coordinate Systems

The idea is you have a coordinate system which is moving and this results in strange forces experienced relative to these moving coordinates systems. A good example is what we experience every day living on a rotating ball. Relative to our supposedly fixed coordinate system, we experience forces which account for many phenomena which are observed.

24.7.1 The Coriolis Acceleration

Imagine a point on the surface of the earth. Now consider unit vectors, one pointing South, one pointing East and one pointing directly away from the center of the earth.



Denote the first as $\mathbf{i}$, the second as $\mathbf{j}$ and the third as $\mathbf{k}$. If you are standing on the earth you will consider these vectors as fixed, but of course they are not. As the earth turns, they change direction and so each is in reality a function of t. Nevertheless, it is with respect to these apparently fixed vectors that you wish to understand acceleration, velocities, and displacements.

In general, let $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ be an orthonormal basis of vectors for each t, like the vectors described in the first paragraph. It is assumed these vectors are C^1 functions of t. Letting the positive x axis extend in the direction of $\mathbf{i}(t)$, the positive y axis extend in the direction of $\mathbf{j}(t)$, and the positive z axis extend in the direction of $\mathbf{k}(t)$, yields a moving coordinate system. By Theorem 24.4.2 on Page 571, there exists an angular velocity vector, $\mathbf{\Omega}(t)$ such that if $\mathbf{u}(t)$ is any vector which has constant components with respect to $\mathbf{i}(t), \mathbf{j}(t)$, and $\mathbf{k}(t)$, then

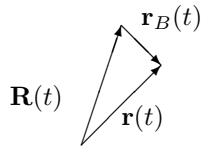
$$\mathbf{\Omega} \times \mathbf{u} = \mathbf{u}'. \tag{24.15}$$

Now let $\mathbf{R}(t)$ be a position vector of the moving coordinate system and let

$$\mathbf{r}(t) = \mathbf{R}(t) + \mathbf{r}_B(t)$$

where

$$\mathbf{r}_B(t) \equiv x(t)\,\mathbf{i}(t) + y(t)\,\mathbf{j}(t) + z(t)\,\mathbf{k}(t).$$



In the example of the earth, $\mathbf{R}(t)$ is the position vector of a point $\mathbf{p}(t)$ on the earth's surface and $\mathbf{r}_B(t)$ is the position vector of another point from $\mathbf{p}(t)$, thus regarding $\mathbf{p}(t)$ as the origin. $\mathbf{r}_B(t)$ is the position vector of a point as perceived by the observer on the earth with respect to the vectors he thinks of as fixed. Similarly, $\mathbf{v}_B(t)$ and $\mathbf{a}_B(t)$ will be the velocity and acceleration relative to $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$, and so $\mathbf{v}_B = x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k}$ and $\mathbf{a}_B = x''\mathbf{i} + y''\mathbf{j} + z''\mathbf{k}$. Then

$$\mathbf{v} \equiv \mathbf{r}' = \mathbf{R}' + x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k} + x\mathbf{i}' + y\mathbf{j}' + z\mathbf{k}'.$$

By , (24.15), if $\mathbf{e} \in \{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$, $\mathbf{e}' = \mathbf{\Omega} \times \mathbf{e}$ because the components of these vectors with respect to $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are constant. Therefore,

$$\begin{aligned} x\mathbf{i}' + y\mathbf{j}' + z\mathbf{k}' &= x\mathbf{\Omega} \times \mathbf{i} + y\mathbf{\Omega} \times \mathbf{j} + z\mathbf{\Omega} \times \mathbf{k} \\ &= \mathbf{\Omega} \times (x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) \end{aligned}$$

and consequently,

$$\mathbf{v} = \mathbf{R}' + x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k} + \mathbf{\Omega} \times \mathbf{r}_B = \mathbf{R}' + x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k} + \mathbf{\Omega} \times (x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) .$$

Now consider the acceleration. Quantities which are relative to the moving coordinate system are distinguished by using the subscript, B.

$$\begin{aligned}\mathbf{a} = \mathbf{v}' &= \mathbf{R}'' + x''\mathbf{i} + y''\mathbf{j} + z''\mathbf{k} + \overbrace{x'\mathbf{i}' + y'\mathbf{j}' + z'\mathbf{k}'}^{\mathbf{\Omega}\times\mathbf{v}_B} + \mathbf{\Omega}' \times \mathbf{r}_B \\ &\quad + \mathbf{\Omega} \times \left(\overbrace{x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k}}^{\mathbf{v}_B} + \overbrace{x\mathbf{i}' + y\mathbf{j}' + z\mathbf{k}'}^{\mathbf{\Omega}\times\mathbf{r}_B(t)} \right) \\ &= \mathbf{R}'' + \mathbf{a}_B + \mathbf{\Omega}' \times \mathbf{r}_B + 2\mathbf{\Omega} \times \mathbf{v}_B + \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B) .\end{aligned}$$

The acceleration $\mathbf{a}_B$ is that perceived by an observer for whom the moving coordinate system is fixed. The term $\mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B)$ is called the centripetal acceleration. Solving for $\mathbf{a}_B$,

$$\mathbf{a}_B = \mathbf{a} - \mathbf{R}'' - \mathbf{\Omega}' \times \mathbf{r}_B - 2\mathbf{\Omega} \times \mathbf{v}_B - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B) . \tag{24.16}$$

Here the term $-(\mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B))$ is called the centrifugal acceleration, it being an acceleration felt by the observer relative to the moving coordinate system which he regards as fixed, and the term $-2\mathbf{\Omega} \times \mathbf{v}_B$ is called the Coriolis acceleration, an acceleration experienced by the observer as he moves relative to the moving coordinate system. The mass multiplied by the Coriolis acceleration defines the Coriolis force.

There is a ride found in some amusement parks in which the victims stand next to a circular wall covered with a carpet or some rough material. Then the whole circular room begins to revolve faster and faster. At some point, the bottom drops out and the victims are held in place by friction. The force they feel which keeps them stuck to the wall is called centrifugal force and it causes centrifugal acceleration. It is not necessary to move relative to coordinates fixed with the revolving wall in order to feel this force and it is pretty predictable. However, if the nauseated victim moves relative to the rotating wall, he will feel the effects of the Coriolis force and this force is really strange. The difference between these forces is that the Coriolis force is caused by movement relative to the moving coordinate system and the centrifugal force is not.

24.7.2 The Coriolis Acceleration On The Rotating Earth

Now consider the earth. Let $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$, be the usual basis vectors attached to the rotating earth. Thus $\mathbf{k}^*$ is fixed in space with $\mathbf{k}^*$ pointing in the direction of the north pole from the center of the earth while $\mathbf{i}^*$ and $\mathbf{j}^*$ point to fixed points on the surface of the earth. Thus $\mathbf{i}^*$ and $\mathbf{j}^*$ depend on t while $\mathbf{k}^*$ does not. Let $\mathbf{i}, \mathbf{j}, \mathbf{k}$ be the unit vectors described earlier with $\mathbf{i}$ pointing South, $\mathbf{j}$ pointing East, and $\mathbf{k}$ pointing away from the center of the earth at some point of the rotating earth's surface, $\mathbf{p}$. Letting $\mathbf{R}(t)$ be the position vector of the point $\mathbf{p}$, from the center of the earth, observe the coordinates of $\mathbf{R}(t)$ are constant with respect to $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$. Also, since the earth rotates from West to East and the speed of a point on the surface of the earth relative to an observer fixed in space is $\omega |\mathbf{R}| \sin\phi$ where ω is the angular speed of the earth about an axis through the poles, it follows from the geometric definition of the cross product that

$$\mathbf{R}' = \omega \mathbf{k}^* \times \mathbf{R}$$

Therefore, $\mathbf{\Omega} = \omega \mathbf{k}^*$ and so

$$\mathbf{R}'' = \overbrace{\mathbf{\Omega}' \times \mathbf{R}}^{=\mathbf{0}} + \mathbf{\Omega} \times \mathbf{R}' = \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{R})$$

since $\mathbf{\Omega}$ does not depend on t. Formula (24.16) implies

$$\mathbf{a}_B = \mathbf{a} - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{R}) - 2\mathbf{\Omega} \times \mathbf{v}_B - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B). \tag{24.17}$$

In this formula, you can totally ignore the term $\mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r}_B)$ because it is so small whenever you are considering motion near some point on the earth's surface. To see this, note $\omega \overbrace{(24)(3600)}^{\text{seconds in a day}} = 2\pi$, and so $\omega = 7.2722 \times 10^{-5}$ in radians per second. If you are using seconds to measure time and feet to measure distance, this term is therefore, no larger than

$$\left(7.2722 \times 10^{-5}\right)^2 |\mathbf{r}_B|.$$

Clearly this is not worth considering in the presence of the acceleration due to gravity which is approximately 32 feet per second squared near the surface of the earth.

If the acceleration $\mathbf{a}$, is due to gravity, then

$$\mathbf{a}_B = \mathbf{a} - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{R}) - 2\mathbf{\Omega} \times \mathbf{v}_B =$$

$$\overbrace{-\frac{GM(\mathbf{R}+\mathbf{r}_B)}{|\mathbf{R}+\mathbf{r}_B|^3} - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{R})}^{\equiv \mathbf{g}} - 2\mathbf{\Omega} \times \mathbf{v}_B \equiv \mathbf{g} - 2\mathbf{\Omega} \times \mathbf{v_B}.$$

Note that

$$\mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{R}) = (\mathbf{\Omega} \cdot \mathbf{R})\mathbf{\Omega} - |\mathbf{\Omega}|^2 \mathbf{R}$$

and so $\mathbf{g}$, the acceleration relative to the moving coordinate system on the earth is not directed exactly toward the center of the earth except at the poles and at the equator, although the components of acceleration which are in other directions are very small when compared with the acceleration due to the force of gravity and are often neglected. Therefore, if the only force acting on an object is due to gravity, the following formula describes the acceleration relative to a coordinate system moving with the earth's surface.

$$\mathbf{a}_B = \mathbf{g} - 2(\mathbf{\Omega} \times \mathbf{v}_B)$$

While the vector, $\mathbf{\Omega}$ is quite small, if the relative velocity, $\mathbf{v}_B$ is large, the Coriolis acceleration could be significant. This is described in terms of the vectors $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ next.

Letting (ρ, θ, ϕ) be the usual spherical coordinates of the point $\mathbf{p}(t)$ on the surface taken with respect to $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$ the usual way with ϕ the polar angle, it follows the $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$ coordinates of this point are

$$\begin{pmatrix} \rho \sin(\phi) \cos(\theta) \\ \rho \sin(\phi) \sin(\theta) \\ \rho \cos(\phi) \end{pmatrix}.$$

It follows,

$$\mathbf{i} = \cos(\phi) \cos(\theta)\, \mathbf{i}^* + \cos(\phi) \sin(\theta)\, \mathbf{j}^* - \sin(\phi)\, \mathbf{k}^*$$

$$\mathbf{j} = -\sin(\theta)\, \mathbf{i}^* + \cos(\theta)\, \mathbf{j}^* + 0\mathbf{k}^*$$

and

$$\mathbf{k} = \sin(\phi) \cos(\theta)\, \mathbf{i}^* + \sin(\phi) \sin(\theta)\, \mathbf{j}^* + \cos(\phi)\, \mathbf{k}^*.$$

It is necessary to obtain $\mathbf{k}^*$ in terms of the vectors, $\mathbf{i}, \mathbf{j}, \mathbf{k}$. Thus the following equation needs to be solved for a, b, c to find $\mathbf{k}^* = a\mathbf{i}+b\mathbf{j}+c\mathbf{k}$

$$\overbrace{\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}}^{\mathbf{k}^*} = \begin{pmatrix} \cos(\phi)\cos(\theta) & -\sin(\theta) & \sin(\phi)\cos(\theta) \\ \cos(\phi)\sin(\theta) & \cos(\theta) & \sin(\phi)\sin(\theta) \\ -\sin(\phi) & 0 & \cos(\phi) \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} \tag{24.18}$$

The first column is $\mathbf{i}$, the second is $\mathbf{j}$ and the third is $\mathbf{k}$ in the above matrix. The solution is $a = -\sin(\phi), b = 0$, and $c = \cos(\phi)$.

Now the Coriolis acceleration on the earth equals

$$2(\mathbf{\Omega} \times \mathbf{v}_B) = 2\omega \left(\overbrace{-\sin(\phi)\,\mathbf{i}+0\mathbf{j}+\cos(\phi)\,\mathbf{k}}^{\mathbf{k}^*} \right) \times (x'\mathbf{i}+y'\mathbf{j}+z'\mathbf{k}).$$

This equals

$$2\omega\left[(-y'\cos\phi)\,\mathbf{i}+(x'\cos\phi+z'\sin\phi)\,\mathbf{j}-(y'\sin\phi)\,\mathbf{k}\right]. \tag{24.19}$$

Remember ϕ is fixed and pertains to the fixed point, $\mathbf{p}(t)$ on the earth's surface. Therefore, if the acceleration, $\mathbf{a}$ is due to gravity,

$$\mathbf{a}_B = \mathbf{g}-2\omega\left[(-y'\cos\phi)\,\mathbf{i}+(x'\cos\phi+z'\sin\phi)\,\mathbf{j}-(y'\sin\phi)\,\mathbf{k}\right]$$

where $\mathbf{g} = -\frac{GM(\mathbf{R}+\mathbf{r}_B)}{|\mathbf{R}+\mathbf{r}_B|^3} - \mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R})$ as explained above. The term $\mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R})$ is pretty small and so it will be neglected. However, the Coriolis force will not be neglected.

Example 24.7.1 *Suppose a rock is dropped from a tall building. Where will it strike?*

Assume $\mathbf{a} = -g\mathbf{k}$ and the $\mathbf{j}$ component of $\mathbf{a}_B$ is approximately

$$-2\omega(x'\cos\phi+z'\sin\phi).$$

The dominant term in this expression is clearly the second one because x' will be small. Also, the $\mathbf{i}$ and $\mathbf{k}$ contributions will be very small. Therefore, the following equation is descriptive of the situation.

$$\mathbf{a}_B = -g\mathbf{k}-2z'\omega\sin\phi\mathbf{j}.$$

$z' = -gt$ approximately. Therefore, considering the $\mathbf{j}$ component, this is

$$2gt\omega\sin\phi.$$

Two integrations give $\left(\omega gt^3/3\right)\sin\phi$ for the $\mathbf{j}$ component of the relative displacement at time t.

This shows the rock does not fall directly towards the center of the earth as expected but slightly to the east.

Example 24.7.2 *In 1851 Foucault set a pendulum vibrating and observed the earth rotate out from under it. It was a very long pendulum with a heavy weight at the end so that it would vibrate for a long time without stopping*[3]*. This is what allowed him to observe the earth rotate out from under it. Clearly such a pendulum will take 24 hours for the plane of vibration to appear to make one complete revolution at the north pole. It is also reasonable to expect that no such observed rotation would take place on the equator. Is it possible to predict what will take place at various latitudes?*

[3]There is such a pendulum in the Eyring building at BYU and to keep people from touching it, there is a little sign which says Warning! 1000 ohms.

Using (24.19), in (24.17),

$$\mathbf{a}_B = \mathbf{a} - \mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R})$$
$$-2\omega\left[(-y'\cos\phi)\,\mathbf{i}+(x'\cos\phi+z'\sin\phi)\,\mathbf{j}-(y'\sin\phi)\,\mathbf{k}\right].$$

Neglecting the small term, $\mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R})$, this becomes

$$=-g\mathbf{k}+\mathbf{T}/m-2\omega\left[(-y'\cos\phi)\,\mathbf{i}+(x'\cos\phi+z'\sin\phi)\,\mathbf{j}-(y'\sin\phi)\,\mathbf{k}\right]$$

where $\mathbf{T}$, the tension in the string of the pendulum, is directed towards the point at which the pendulum is supported, and m is the mass of the pendulum bob. The pendulum can be thought of as the position vector from $(0,0,l)$ to the surface of the sphere $x^2+y^2+(z-l)^2=l^2$. Therefore,

$$\mathbf{T}=-T\frac{x}{l}\mathbf{i}-T\frac{y}{l}\mathbf{j}+T\frac{l-z}{l}\mathbf{k}$$

and consequently, the differential equations of relative motion are

$$x''=-T\frac{x}{ml}+2\omega y'\cos\phi$$

$$y''=-T\frac{y}{ml}-2\omega\left(x'\cos\phi+z'\sin\phi\right)$$

and

$$z''=T\frac{l-z}{ml}-g+2\omega y'\sin\phi.$$

If the vibrations of the pendulum are small so that for practical purposes, $z''=z=0$, the last equation may be solved for T to get

$$gm-2\omega y'\sin(\phi)\,m=T.$$

Therefore, the first two equations become

$$x''=-\left(gm-2\omega my'\sin\phi\right)\frac{x}{ml}+2\omega y'\cos\phi$$

and

$$y''=-\left(gm-2\omega my'\sin\phi\right)\frac{y}{ml}-2\omega\left(x'\cos\phi+z'\sin\phi\right).$$

All terms of the form xy' or $y'y$ can be neglected because it is assumed x and y remain small. Also, the pendulum is assumed to be long with a heavy weight so that x' and y' are also small. With these simplifying assumptions, the equations of motion become

$$x''+g\frac{x}{l}=2\omega y'\cos\phi$$

and

$$y''+g\frac{y}{l}=-2\omega x'\cos\phi.$$

These equations are of the form

$$x''+a^2x=by',\ y''+a^2y=-bx' \tag{24.20}$$

where $a^2=\frac{g}{l}$ and $b=2\omega\cos\phi$. Then it is fairly tedious but routine to verify that for each constant, c,

$$x=c\sin\left(\frac{bt}{2}\right)\sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right),\ y=c\cos\left(\frac{bt}{2}\right)\sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right) \tag{24.21}$$

yields a solution to (24.20) along with the initial conditions,

$$x(0)=0, y(0)=0, x'(0)=0, y'(0)=\frac{c\sqrt{b^2+4a^2}}{2}. \tag{24.22}$$

It is clear from experiments with the pendulum that the earth does indeed rotate out from under it causing the plane of vibration of the pendulum to appear to rotate. The purpose of this discussion is not to establish these self evident facts but to predict how long it takes for the plane of vibration to make one revolution. Therefore, there will be some instant in time at which the pendulum will be vibrating in a plane determined by $\mathbf{k}$ and $\mathbf{j}$. (Recall $\mathbf{k}$ points away from the center of the earth and $\mathbf{j}$ points East.) At this instant in time, defined as $t=0$, the conditions of (24.22) will hold for some value of c and so the solution to (24.20) having these initial conditions will be those of (24.21) by uniqueness of the initial value problem. Writing these solutions differently,

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = c\begin{pmatrix} \sin\left(\frac{bt}{2}\right) \\ \cos\left(\frac{bt}{2}\right) \end{pmatrix} \sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right)$$

This is very interesting! The vector, $c\begin{pmatrix} \sin\left(\frac{bt}{2}\right) \\ \cos\left(\frac{bt}{2}\right) \end{pmatrix}$ always has magnitude equal to $|c|$ but its direction changes very slowly because b is very small. The plane of vibration is determined by this vector and the vector $\mathbf{k}$. The term $\sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right)$ changes relatively fast and takes values between -1 and 1. This is what describes the actual observed vibrations of the pendulum. Thus the plane of vibration will have made one complete revolution when $t=P$ for

$$\frac{bP}{2}\equiv 2\pi.$$

Therefore, the time it takes for the earth to turn out from under the pendulum is

$$P=\frac{4\pi}{2\omega\cos\phi}=\frac{2\pi}{\omega}\sec\phi.$$

Since ω is the angular speed of the rotating earth, it follows $\omega=\frac{2\pi}{24}=\frac{\pi}{12}$ in radians per hour. Therefore, the above formula implies

$$P=24\sec\phi.$$

I think this is really amazing. You could actually determine latitude, not by taking readings with instruments using the North Star but by doing an experiment with a big pendulum. You would set it vibrating, observe P in hours, and then solve the above equation for ϕ. Also note the pendulum would not appear to change its plane of vibration at the equator because $\lim_{\phi\to\pi/2}\sec\phi=\infty$.

The Coriolis acceleration is also responsible for the phenomenon of the next example.

Example 24.7.3 *It is known that low pressure areas rotate counterclockwise as seen from above in the Northern hemisphere but clockwise in the Southern hemisphere. Why?*

Neglect accelerations other than the Coriolis acceleration and the following acceleration which comes from an assumption that the point $\mathbf{p}(t)$ is the location of the lowest pressure.

$$\mathbf{a}=-a(r_B)\,\mathbf{r}_B$$

where $r_B=r$ will denote the distance from the fixed point $\mathbf{p}(t)$ on the earth's surface which is also the lowest pressure point. Of course the situation could be more complicated but

this will suffice to explain the above question. Then the acceleration observed by a person on the earth relative to the apparently fixed vectors, $\mathbf{i}, \mathbf{k}, \mathbf{j}$, is

$$\mathbf{a}_B = -a(r_B)(x\mathbf{i}+y\mathbf{j}+z\mathbf{k}) - 2\omega\left[-y'\cos(\phi)\,\mathbf{i}+(x'\cos(\phi)+z'\sin(\phi))\,\mathbf{j}-(y'\sin(\phi)\,\mathbf{k})\right]$$

Therefore, one obtains some differential equations from $\mathbf{a}_B = x''\mathbf{i}+y''\mathbf{j}+z''\mathbf{k}$ by matching the components. These are

$$\begin{aligned} x''+a(r_B)\,x &= 2\omega y'\cos\phi \\ y''+a(r_B)\,y &= -2\omega x'\cos\phi - 2\omega z'\sin(\phi) \\ z''+a(r_B)\,z &= 2\omega y'\sin\phi \end{aligned}$$

Now remember, the vectors, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are fixed relative to the earth and so are constant vectors. Therefore, from the properties of the determinant and the above differential equations,

$$(\mathbf{r}_B'\times\mathbf{r}_B)' = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x' & y' & z' \\ x & y & z \end{vmatrix}' = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x'' & y'' & z'' \\ x & y & z \end{vmatrix}$$

$$= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -a(r_B)\,x+2\omega y'\cos\phi & -a(r_B)\,y-2\omega x'\cos\phi-2\omega z'\sin(\phi) & -a(r_B)\,z+2\omega y'\sin\phi \\ x & y & z \end{vmatrix}$$

Then the $\mathbf{k}^{th}$ component of this cross product equals

$$\omega\cos(\phi)\left(y^2+x^2\right)' + 2\omega x z'\sin(\phi).$$

The first term will be negative because it is assumed $\mathbf{p}(t)$ is the location of low pressure causing y^2+x^2 to be a decreasing function. If it is assumed there is not a substantial motion in the $\mathbf{k}$ direction, so that z is fairly constant and the last term can be neglected, then the $\mathbf{k}^{th}$ component of $(\mathbf{r}_B'\times\mathbf{r}_B)'$ is negative provided $\phi\in\left(0,\frac{\pi}{2}\right)$ and positive if $\phi\in\left(\frac{\pi}{2},\pi\right)$. Beginning with a point at rest, this implies $\mathbf{r}_B'\times\mathbf{r}_B=\mathbf{0}$ initially and then the above implies its $\mathbf{k}^{th}$ component is negative in the upper hemisphere when $\phi<\pi/2$ and positive in the lower hemisphere when $\phi>\pi/2$. Using the right hand and the geometric definition of the cross product, this shows clockwise rotation in the lower hemisphere and counter clockwise rotation in the upper hemisphere.

Note also that as ϕ gets close to $\pi/2$ near the equator, the above reasoning tends to break down because $\cos(\phi)$ becomes close to zero. Therefore, the motion towards the low pressure has to be more pronounced in comparison with the motion in the $\mathbf{k}$ direction in order to draw this conclusion.

24.8 Exercises

1. Show the solution to $\mathbf{v}'+r\mathbf{v}=\mathbf{c}$ with the initial condition, $\mathbf{v}(0)=\mathbf{v}_0$ is $\mathbf{v}(t)=\left(\mathbf{v}_0-\frac{\mathbf{c}}{r}\right)e^{-rt}+(\mathbf{c}/r)$. If $\mathbf{v}$ is velocity and $r=k/m$ where k is a constant for air resistance and m is the mass, and $\mathbf{c}=\mathbf{f}/m$, argue from Newton's second law that this is the equation for finding the velocity, $\mathbf{v}$ of an object acted on by air resistance proportional to the velocity and a constant force, $\mathbf{f}$, possibly from gravity. Does there exist a terminal velocity? What is it?

2. Verify Formula (24.14) carefully by considering the components.

3. Suppose that the air resistance is proportional to the velocity but it is desired to find the constant of proportionality. Describe how you could find this constant.

4. Suppose an object having mass equal to 5 kilograms experiences a time dependent force, $\mathbf{F}(t) = e^{-t}\mathbf{i} + \cos(t)\,\mathbf{j} + t^2\mathbf{k}$ meters per sec^2. Suppose also that the object is at the point $(0,1,1)$ meters at time $t = 0$ and that its initial velocity at this time is $\mathbf{v} = \mathbf{i} + \mathbf{j} - \mathbf{k}$ meters per sec. Find the position of the object as a function of t.

5. Fill in the details for the derivation of kinetic energy. In particular verify that $m\mathbf{v}'(t) \cdot \mathbf{v}(t) = \frac{m}{2}\frac{d}{dt}|\mathbf{v}(t)|^2$. Also, why would $dW = \mathbf{F}(t) \cdot \mathbf{v}(t)\,dt$?

6. Suppose the force acting on an object, $\mathbf{F}$ is always perpendicular to the velocity of the object. Thus $\mathbf{F} \cdot \mathbf{v} = 0$. Show the Kinetic energy of the object is constant. Such forces are sometimes called forces of constraint because they do not contribute to the speed of the object, only its direction.

7. A cannon is fired at an angle, θ from ground level on a vast plain. The speed of the ball as it leaves the mouth of the cannon is known to be s meters per second. Neglecting air resistance, find a formula for how far the cannon ball goes before hitting the ground. Show the maximum range for the cannon ball is achieved when $\theta = \pi/4$.

8. Suppose in the context of Problem 7 that the cannon ball has mass 10 kilograms and it experiences a force of air resistance which is $.01\mathbf{v}$ Newtons where $\mathbf{v}$ is the velocity in meters per second. The acceleration of gravity is 9.8 meters per sec^2. Also suppose that the initial speed is 100 meters per second. Find a formula for the displacement, $\mathbf{r}(t)$ of the cannon ball. If the angle of elevation equals $\pi/4$, use a calculator or other means to estimate the time before the cannon ball hits the ground.

9. Show that Newton's first law can be obtained from the second law.

10. Show that if $\mathbf{v}'(t) = \mathbf{0}$, for all $t \in (a,b)$, then there exists a constant vector, $\mathbf{z}$ independent of t such that $\mathbf{v}(t) = \mathbf{z}$ for all t.

11. Suppose an object moves in three dimensional space in such a way that the only force acting on the object is directed toward a single fixed point in three dimensional space. Verify that the motion of the object takes place in a plane. **Hint:** Let $\mathbf{r}(t)$ denote the position vector of the object from the fixed point. Then the force acting on the object must be of the form $g(\mathbf{r}(t))\,\mathbf{r}(t)$ and by Newton's second law, this equals $m\mathbf{r}''(t)$. Therefore,
$$m\mathbf{r}'' \times \mathbf{r} = g(\mathbf{r})\,\mathbf{r} \times \mathbf{r} = \mathbf{0}.$$
Now argue that $\mathbf{r}'' \times \mathbf{r} = (\mathbf{r}' \times \mathbf{r})'$, showing that $(\mathbf{r}' \times \mathbf{r})$ must equal a constant vector, $\mathbf{z}$. Therefore, what can be said about $\mathbf{z}$ and $\mathbf{r}$?

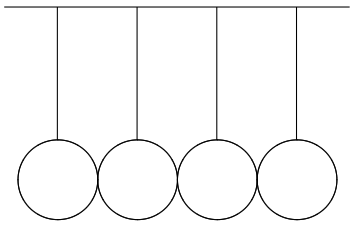
12. Suppose the only forces acting on an object are the force of gravity, $-mg\mathbf{k}$ and a force, $\mathbf{F}$ which is perpendicular to the motion of the object. Thus $\mathbf{F} \cdot \mathbf{v} = \mathbf{0}$. Show the total energy of the object,
$$E \equiv \frac{1}{2}m|\mathbf{v}|^2 + mgz$$
is constant. Here $\mathbf{v}$ is the velocity and the first term is the kinetic energy while the second is the potential energy. **Hint:** Use Newton's second law to show the time derivative of the above expression equals zero.

13. Using Problem 12, suppose an object slides down a frictionless inclined plane from a height of 100 feet. When it reaches the bottom, how fast will it be going? Assume it starts from rest.

14. The ballistic pendulum is an interesting device which is used to determine the speed of a bullet. It is a large massive block of wood hanging from a long string. A rifle is fired into the block of wood which then moves. The speed of the bullet can be determined from measuring how high the block of wood rises. Explain how this can be done and why. **Hint:** Let v be the speed of the bullet which has mass m and let the block of wood have mass M. By conservation of momentum $mv = (m+M)\,V$ where V is the speed of the block of wood immediately after the collision. Thus the energy is $\frac{1}{2}(m+M)\,V^2$ and this block of wood rises to a height of h. Now use Problem 12.

15. In the experiment of Problem 14, show the kinetic energy before the collision is greater than the kinetic energy after the collision. Thus linear momentum is conserved but energy is not. Such a collision is called inelastic.

16. There is a popular toy consisting of identical steel balls suspended from strings of equal length as illustrated in the following picture.



The ball at the right is lifted and allowed to swing. When it collides with the other balls, the ball on the left is observed to swing away from the others with the same speed the ball on the right had when it collided. Why does this happen? Why don't two or more of the stationary balls start to move, perhaps at a slower speed? This is an example of an elastic collision because energy is conserved. Of course this could change if you fixed things so the balls would stick to each other.

17. An illustration used in many beginning physics books is that of firing a rifle horizontally and dropping an identical bullet from the same height above the perfectly flat ground followed by an assertion that the two bullets will hit the ground at exactly the same time. Is this true on the rotating earth assuming the experiment takes place over a large perfectly flat field so the curvature of the earth is not an issue? Explain. What other irregularities will occur? Recall the Coriolis force is $2\omega\left[(-y'\cos\phi)\,\mathbf{i}+(x'\cos\phi+z'\sin\phi)\,\mathbf{j}-(y'\sin\phi)\,\mathbf{k}\right]$ where $\mathbf{k}$ points away from the center of the earth, $\mathbf{j}$ points East, and $\mathbf{i}$ points South.

18. Suppose you have n masses, $m_1,\cdots,m_n$. Let the position vector of the i^{th} mass be $\mathbf{r}_i(t)$. The center of mass of these is defined to be

$$\mathbf{R}(t) \equiv \frac{\sum_{i=1}^n \mathbf{r}_i m_i}{\sum_{i=1}^n m_i} \equiv \frac{\sum_{i=1}^n \mathbf{r}_i(t)\, m_i}{M}.$$

Let $\mathbf{r}_{Bi}(t) = \mathbf{r}_i(t) - \mathbf{R}(t)$. Show that $\sum_{i=1}^n m_i \mathbf{r}_i(t) - \sum_i m_i \mathbf{R}(t) = \mathbf{0}$.

19. Suppose you have n masses, $m_1,\cdots,m_n$ which make up a moving rigid body. Let $\mathbf{R}(t)$ denote the position vector of the center of mass of these n masses. Find a formula for the total kinetic energy in terms of this position vector, the angular velocity vector, and the position vector of each mass from the center of mass. **Hint:** Use Problem 18.

24.9 Line Integrals

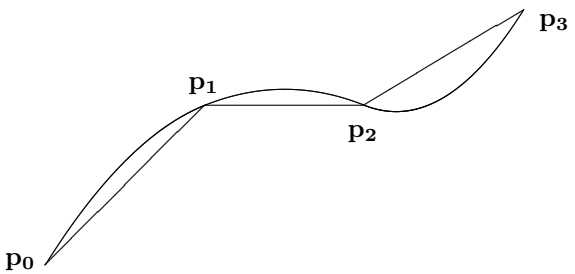
The concept of the integral can be extended to functions which are not defined on an interval of the real line but on some curve in $\mathbb{R}^n$. This is done by defining things in such a way that the more general concept reduces to the earlier notion. First it is necessary to consider what is meant by arc length.

24.9.1 Arc Length And Orientations

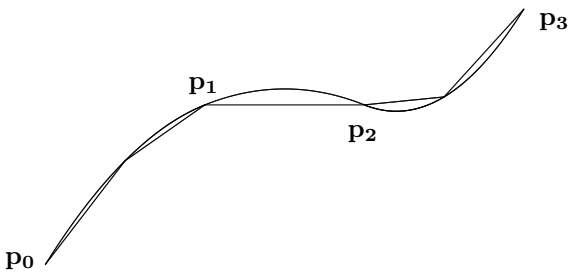
The application of the integral considered here is the concept of the length of a curve. C is a smooth curve in $\mathbb{R}^n$ if there exists an interval, $[a,b] \subseteq \mathbb{R}$ and functions $x_i : [a,b] \to \mathbb{R}$ such that the following conditions hold

1. x_i is continuous on $[a,b]$.
2. x_i' exists and is continuous and bounded on $[a,b]$, with $x_i'(a)$ defined as the derivative from the right,
$$\lim_{h \to 0+} \frac{x_i(a+h) - x_i(a)}{h},$$
and $x_i'(b)$ defined similarly as the derivative from the left.
3. For $\mathbf{p}(t) \equiv (x_1(t), \cdots, x_n(t))$, $t \to \mathbf{p}(t)$ is one to one on (a,b).
4. $|\mathbf{p}'(t)| \equiv \left(\sum_{i=1}^n |x_i'(t)|^2\right)^{1/2} \neq 0$ for all $t \in [a,b]$.
5. $C = \cup \{(x_1(t), \cdots, x_n(t)) : t \in [a,b]\}$.

The functions, $x_i(t)$, defined above are giving the coordinates of a point in $\mathbb{R}^n$ and the list of these functions is called a parameterization for the smooth curve. Note the natural direction of the interval also gives a direction for moving along the curve. Such a direction is called an orientation. The integral is used to define what is meant by the length of such a smooth curve. Consider such a smooth curve having parameterization $(x_1, \cdots, x_n)$. Forming a partition of $[a,b]$, $a = t_0 < \cdots < t_n = b$ and letting $\mathbf{p}_i = (x_1(t_i), \cdots, x_n(t_i))$, you could consider the polygon formed by lines from $\mathbf{p}_0$ to $\mathbf{p}_1$ and from $\mathbf{p}_1$ to $\mathbf{p}_2$ and from $\mathbf{p}_3$ to $\mathbf{p}_4$ etc. to be an approximation to the curve, C. The following picture illustrates what is meant by this.



Now consider what happens when the partition is refined by including more points. You can see from the following picture that the polygonal approximation would appear to be even better and that as more points are added in the partition, the sum of the lengths of the line segments seems to get close to something which deserves to be defined as the length of the curve, C.



Thus the length of the curve is approximated by

$$\sum_{k=1}^{n} |\mathbf{p}(t_k) - \mathbf{p}(t_{k-1})|.$$

Since the functions in the parameterization are differentiable, it is reasonable to expect this to be close to

$$\sum_{k=1}^{n} |\mathbf{p}'(t_{k-1})| (t_k - t_{k-1})$$

which is seen to be a Riemann sum for the integral

$$\int_a^b |\mathbf{p}'(t)| \, dt$$

and it is this integral which is defined as the length of the curve.

Would the same length be obtained if another parameterization were used? This is a very important question because the length of the curve should depend only on the curve itself and not on the method used to trace out the curve. The answer to this question is that the length of the curve does not depend on parameterization. The proof is somewhat technical so is given in the last section of this chapter.

Does the definition of length given above correspond to the usual definition of length in the case when the curve is a line segment? It is easy to see that it does so by considering two points in $\mathbb{R}^n$, $\mathbf{p}$ and $\mathbf{q}$. A parameterization for the line segment joining these two points is

$$f_i(t) \equiv tp_i + (1-t) q_i, \; t \in [0,1].$$

Using the definition of length of a smooth curve just given, the length according to this definition is

$$\int_0^1 \left(\sum_{i=1}^{n} (p_i - q_i)^2 \right)^{1/2} dt = |\mathbf{p} - \mathbf{q}|.$$

Thus this new definition which is valid for smooth curves which may not be straight line segments gives the usual length for straight line segments.

The proof that curve length is well defined for a smooth curve contains a result which deserves to be stated as a corollary. It is proved in Lemma 25.4.13 on Page 608 but the proof is mathematically fairly advanced so it is presented later.

Corollary 24.9.1 *Let C be a smooth curve and let $\mathbf{f} : [a,b] \to C$ and $\mathbf{g} : [c,d] \to C$ be two parameterizations satisfying 1 - 5. Then $\mathbf{g}^{-1} \circ \mathbf{f}$ is either strictly increasing or strictly decreasing.*

Definition 24.9.2 *If* $\mathbf{g}^{-1} \circ \mathbf{f}$ *is increasing, then* $\mathbf{f}$ *and* $\mathbf{g}$ *are said to be equivalent parameterizations and this is written as* $\mathbf{f} \sim \mathbf{g}$. *It is also said that the two parameterizations give the same orientation for the curve when* $\mathbf{f} \sim \mathbf{g}$.

When the parameterizations are equivalent, they preserve the direction, of motion along the curve and this also shows there are exactly two orientations of the curve since either $\mathbf{g}^{-1} \circ \mathbf{f}$ is increasing or it is decreasing. This is not hard to believe. In simple language, the message is that there are exactly two directions of motion along a curve. The difficulty is in proving this is actually the case.

Lemma 24.9.3 *The following hold for* $\sim$.

$$\mathbf{f} \sim \mathbf{f}, \tag{24.23}$$

$$\text{If } \mathbf{f} \sim \mathbf{g} \text{ then } \mathbf{g} \sim \mathbf{f}, \tag{24.24}$$

$$\text{If } \mathbf{f} \sim \mathbf{g} \text{ and } \mathbf{g} \sim \mathbf{h}, \text{ then } \mathbf{f} \sim \mathbf{h}. \tag{24.25}$$

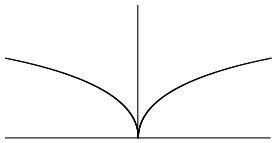
Proof: Formula (24.23) is obvious because $\mathbf{f}^{-1} \circ \mathbf{f}(t) = t$ so it is clearly an increasing function. If $\mathbf{f} \sim \mathbf{g}$ then $\mathbf{f}^{-1} \circ \mathbf{g}$ is increasing. Now $\mathbf{g}^{-1} \circ \mathbf{f}$ must also be increasing because it is the inverse of $\mathbf{f}^{-1} \circ \mathbf{g}$. This verifies (24.24). To see (24.25), $\mathbf{f}^{-1} \circ \mathbf{h} = \left(\mathbf{f}^{-1} \circ \mathbf{g}\right) \circ \left(\mathbf{g}^{-1} \circ \mathbf{h}\right)$ and so since both of these functions are increasing, it follows $\mathbf{f}^{-1} \circ \mathbf{h}$ is also increasing. This proves the lemma.

The symbol, $\sim$ is called an equivalence relation. If C is such a smooth curve just described, and if $\mathbf{f} : [a, b] \to C$ is a parameterization of C, consider $\mathbf{g}(t) \equiv \mathbf{f}((a + b) - t)$, also a parameterization of C. Now by Corollary 24.9.1, if $\mathbf{h}$ is a parameterization, then if $\mathbf{f}^{-1} \circ \mathbf{h}$ is not increasing, it must be the case that $\mathbf{g}^{-1} \circ \mathbf{h}$ is increasing. Consequently, either $\mathbf{h} \sim \mathbf{g}$ or $\mathbf{h} \sim \mathbf{f}$. These parameterizations, $\mathbf{h}$, which satisfy $\mathbf{h} \sim \mathbf{f}$ are called the equivalence class determined by $\mathbf{f}$ and those $\mathbf{h} \sim \mathbf{g}$ are called the equivalence class determined by $\mathbf{g}$. These two classes are called orientations of C. They give the direction of motion on C. You see that going from $\mathbf{f}$ to $\mathbf{g}$ corresponds to tracing out the curve in the opposite direction.

Sometimes people wonder why it is required, in the definition of a smooth curve that $\mathbf{p}'(t) \neq \mathbf{0}$. Imagine t is time and $\mathbf{p}(t)$ gives the location of a point in space. If $\mathbf{p}'(t)$ is allowed to equal zero, the point can stop and change directions abruptly, producing a pointy place in C. Here is an example.

Example 24.9.4 *Graph the curve* $\left(t^3, t^2\right)$ *for* $t \in [-1, 1]$.

In this case, $t = x^{1/3}$ and so $y = x^{2/3}$. Thus the graph of this curve looks like the picture below. Note the pointy place. Such a curve should not be considered smooth.



24.9.2 Line Integrals And Work

Let C be a smooth curve contained in $\mathbb{R}^p$. A curve, C is an "oriented curve" if the only parameterizations considered are those which lie in exactly one of the two equivalence classes, each of which is called an "orientation". In simple language, orientation specifies a direction over which motion along the curve is to take place. Thus, it specifies the order in which the points of C are encountered. The pair of concepts consisting of the set of points making up the curve along with a direction of motion along the curve is called an oriented curve.

Definition 24.9.5 *Suppose* $\mathbf{F}(\mathbf{x}) \in \mathbb{R}^p$ *is given for each* $\mathbf{x} \in C$ *where* C *is a smooth oriented curve and suppose* $\mathbf{x} \to \mathbf{F}(\mathbf{x})$ *is continuous. The mapping* $\mathbf{x} \to \mathbf{F}(\mathbf{x})$ *is called a vector field. In the case that* $\mathbf{F}(\mathbf{x})$ *is a force, it is called a force field.*

Next the concept of work done by a force field, $\mathbf{F}$ on an object as it moves along the curve, C, in the direction determined by the given orientation of the curve will be defined. This is new. Earlier the work done by a force which acts on an object moving in a straight line was discussed but here the object moves oveo a curve. In ordeo to define what is meant by the work, consideo the following picture.

Theorem 24.9.7 *The symbol, $\int_C \mathbf{F}\cdot d\mathbf{R}$, is well defined in the sense that every parameterization in the given orientation of C gives the same value for $\int_C \mathbf{F}\cdot d\mathbf{R}$.*

Proof: Suppose $\mathbf{g}:[c,d]\to C$ is another allowed parameterization. Thus $\mathbf{g}^{-1}\circ\mathbf{f}$ is an increasing function, ϕ. Then since ϕ is increasing,

$$\int_c^d \mathbf{F}(\mathbf{g}(s))\cdot\mathbf{g}'(s)\,ds = \int_a^b \mathbf{F}(\mathbf{g}(\phi(t)))\cdot\mathbf{g}'(\phi(t))\,\phi'(t)\,dt$$

$$= \int_a^b \mathbf{F}(\mathbf{f}(t))\cdot\frac{d}{dt}\left(\mathbf{g}\left(\mathbf{g}^{-1}\circ\mathbf{f}(t)\right)\right)\,dt = \int_a^b \mathbf{F}(\mathbf{f}(t))\cdot\mathbf{f}'(t)\,dt.$$

This proves the theorem.

Regardless the physical interpretation of $\mathbf{F}$, this is called the line integral. When $\mathbf{F}$ is interpreted as a force, the line integral measures the extent to which the motion over the curve in the indicated direction is aided by the force. If the net effect of the force on the object is to impede rather than to aid the motion, this will show up as the work being negative.

Does the concept of work as defined here coincide with the earlier concept of work when the object moves over a straight line when acted on by a constant force?

Let $\mathbf{p}$ and $\mathbf{q}$ be two points in $\mathbb{R}^n$ and suppose $\mathbf{F}$ is a constant force acting on an object which moves from $\mathbf{p}$ to $\mathbf{q}$ along the straight line joining these points. Then the work done is $\mathbf{F}\cdot(\mathbf{q}-\mathbf{p})$. Is the same thing obtained from the above definition? Let $\mathbf{x}(t)\equiv\mathbf{p}+t(\mathbf{q}-\mathbf{p}), t\in[0,1]$ be a parameterization for this oriented curve, the straight line in the direction from $\mathbf{p}$ to $\mathbf{q}$. Then $\mathbf{x}'(t)=\mathbf{q}-\mathbf{p}$ and $\mathbf{F}(\mathbf{x}(t))=\mathbf{F}$. Therefore, the above definition yields

$$\int_0^1 \mathbf{F}\cdot(\mathbf{q}-\mathbf{p})\,dt = \mathbf{F}\cdot(\mathbf{q}-\mathbf{p}).$$

Therefore, the new definition adds to but does not contradict the old one.

Example 24.9.8 *Suppose for $t\in[0,\pi]$ the position of an object is given by $\mathbf{r}(t)=t\mathbf{i}+\cos(2t)\,\mathbf{j}+\sin(2t)\,\mathbf{k}$. Also suppose there is a force field defined on $\mathbb{R}^3$, $\mathbf{F}(x,y,z)\equiv 2xy\mathbf{i}+x^2\mathbf{j}+\mathbf{k}$. Find*

$$\int_C \mathbf{F}\cdot d\mathbf{R}$$

where C is the curve traced out by this object which has the orientation determined by the direction of increasing t.

To find this line integral use the above definition and write

$$\int_C \mathbf{F}\cdot d\mathbf{R} = \int_0^\pi \left(2t(\cos(2t)),t^2,1\right)\cdot (1,-2\sin(2t),2\cos(2t))\,dt$$

In evaluating this replace the x in the formula for $\mathbf{F}$ with t, the y in the formula for $\mathbf{F}$ with $\cos(2t)$ and the z in the formula for $\mathbf{F}$ with $\sin(2t)$ because these are the values of these variables which correspond to the value of t. Taking the dot product, this equals the following integral.

$$\int_0^\pi \left(2t\cos 2t - 2(\sin 2t)\,t^2 + 2\cos 2t\right)\,dt = \pi^2$$

Example 24.9.9 *Let C denote the oriented curve obtained by $\mathbf{r}(t) = \left(t, \sin t, t^3\right)$ where the orientation is determined by increasing t for $t \in [0,2]$. Also let $\mathbf{F} = (x, y, xz + z)$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$.*

You use the definition.

$$\begin{aligned}\int_C \mathbf{F} \cdot d\mathbf{R} &= \int_0^2 \left(t, \sin(t), (t+1)t^3\right) \cdot \left(1, \cos(t), 3t^2\right) dt \\ &= \int_0^2 \left(t + \sin(t)\cos(t) + 3(t+1)t^5\right) dt \\ &= \frac{1251}{14} - \frac{1}{2}\cos^2(2).\end{aligned}$$

Suppose you have a curve specified by $\mathbf{r}(s) = (x(s), y(s), z(s))$ and it has the property that $|\mathbf{r}'(s)| = 1$ for all $s \in [0,b]$. Then the length of this curve for s between 0 and s_1 is

$$\int_0^{s_1} |\mathbf{r}'(s)|\, ds = \int_0^{s_1} 1 ds = s_1.$$

This parameter is therefore called arc length because the length of the curve up to s equals s. Now you can always change the parameter to be arc length.

Proposition 24.9.10 *Suppose C is an oriented smooth curve parameterized by $\mathbf{r}(t)$ for $t \in [a,b]$. Then letting l denote the total length of C, there exists $\mathbf{R}(s)$, $s \in [0,l]$ another parameterization for this curve which preserves the orientation and such that $|\mathbf{R}'(s)| = 1$ so that s is arc length.*

Prove: Let $\phi(t) \equiv \int_a^t |\mathbf{r}'(\tau)|\, d\tau \equiv s$. Then s is an increasing function of t because

$$\frac{ds}{dt} = \phi'(t) = |\mathbf{r}'(t)| > 0.$$

Now define $\mathbf{R}(s) \equiv \mathbf{r}\left(\phi^{-1}(s)\right)$. Then

$$\begin{aligned}\mathbf{R}'(s) &= \mathbf{r}'\left(\phi^{-1}(s)\right)\left(\phi^{-1}\right)'(s) \\ &= \frac{\mathbf{r}'\left(\phi^{-1}(s)\right)}{\left|\mathbf{r}'\left(\phi^{-1}(s)\right)\right|}\end{aligned}$$

and so $|\mathbf{R}'(s)| = 1$ as claimed. $\mathbf{R}(l) = \mathbf{r}\left(\phi^{-1}(l)\right) = \mathbf{r}\left(\phi^{-1}\left(\int_a^b |\mathbf{r}'(\tau)|\, d\tau\right)\right) = \mathbf{r}(b)$ and $\mathbf{R}(0) = \mathbf{r}\left(\phi^{-1}(0)\right) = \mathbf{r}(a)$ and $\mathbf{R}$ delivers the same set of points in the same order as $\mathbf{r}$ because $\frac{ds}{dt} > 0$.

The arc length parameter is just like any other parameter in so far as considerations of line integrals are concerned because it was shown above that line integrals are independent of parameterization. However, when things are defined in terms of the arc length parameterization, it is clear they depend only on geometric properties of the curve itself and for this reason, the arc length parameterization is important in differential geometry.

24.9.3 Another Notation For Line Integrals

Definition 24.9.11 *Let $\mathbf{F}(x,y,z) = (P(x,y,z), Q(x,y,z), R(x,y,z))$ and let C be an oriented curve. Then another way to write $\int_C \mathbf{F} \cdot d\mathbf{R}$ is*

$$\int_C P dx + Q dy + R dz$$

This last is referred to as the integral of a differential form, $Pdx + Qdy + Rdz$. The study of differential forms is important. Formally, $d\mathbf{R} = (dx, dy, dz)$ and so the integrand in the above is formally $\mathbf{F} \cdot d\mathbf{R}$. Other occurances of this notation are handled similarly in 2 or higher dimensions.

24.10 Exercises

1. Suppose for $t \in [0, 2\pi]$ the position of an object is given by $\mathbf{r}(t) = t\mathbf{i} + \cos(2t)\,\mathbf{j} + \sin(2t)\,\mathbf{k}$. Also suppose there is a force field defined on $\mathbb{R}^3, \mathbf{F}(x, y, z) \equiv 2xy\mathbf{i} + \left(x^2 + 2zy\right)\mathbf{j} + y^2\mathbf{k}$. Find the work,
$$\int_C \mathbf{F} \cdot d\mathbf{R}$$
where C is the curve traced out by this object which has the orientation determined by the direction of increasing t.

2. Here is a vector field, $\left(y, x + z^2, 2yz\right)$ and here is the parameterization of a curve, C. $\mathbf{R}(t) = (\cos 2t, 2\sin 2t, t)$ where t goes from 0 to $\pi/4$. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$.

3. If f and g are both increasing functions, show $f \circ g$ is an increasing function also. Assume anything you like about the domains of the functions.

4. Suppose for $t \in [0, 3]$ the position of an object is given by $\mathbf{r}(t) = t\mathbf{i} + t\mathbf{j} + t\mathbf{k}$. Also suppose there is a force field defined on $\mathbb{R}^3, \mathbf{F}(x, y, z) \equiv yz\mathbf{i} + xz\mathbf{j} + xy\mathbf{k}$. Find
$$\int_C \mathbf{F} \cdot d\mathbf{R}$$
where C is the curve traced out by this object which has the orientation determined by the direction of increasing t. Repeat the problem for $\mathbf{r}(t) = t\mathbf{i} + t^2\mathbf{j} + t\mathbf{k}$.

5. Suppose for $t \in [0, 1]$ the position of an object is given by $\mathbf{r}(t) = t\mathbf{i} + t\mathbf{j} + t\mathbf{k}$. Also suppose there is a force field defined on $\mathbb{R}^3, \mathbf{F}(x, y, z) \equiv z\mathbf{i} + xz\mathbf{j} + xy\mathbf{k}$. Find
$$\int_C \mathbf{F} \cdot d\mathbf{R}$$
where C is the curve traced out by this object which has the orientation determined by the direction of increasing t. Repeat the problem for $\mathbf{r}(t) = t\mathbf{i} + t^2\mathbf{j} + t\mathbf{k}$.

6. Let $\mathbf{F}(x, y, z)$ be a given force field and suppose it acts on an object having mass, m on a curve with parameterization, $(x(t), y(t), z(t))$ for $t \in [a, b]$. Show directly that the work done equals the difference in the kinetic energy. **Hint:**
$$\int_a^b \mathbf{F}(x(t), y(t), z(t)) \cdot (x'(t), y'(t), z'(t))\, dt =$$
$$\int_a^b m(x''(t), y''(t), z''(t)) \cdot (x'(t), y'(t), z'(t))\, dt,$$
etc.

Motion On A Space Curve

25.0.1 Outcomes

1. Recall the definitions of unit tangent, unit normal, and osculating plane.
2. Calculate the curvature for a space curve.
3. Given the position vector function of a moving object, calculate the velocity, speed, and acceleration of the object and write the acceleration in terms of its tangential and normal components.
4. Derive formulas for the curvature of a parameterized curve and the curvature of a plane curve given as a function.

25.1 Space Curves

A fly buzzing around the room, a person riding a roller coaster, and a satellite orbiting the earth all have something in common. They are moving over some sort of curve in three dimensions.

Denote by $\mathbf{R}(t)$ the function which takes t to a point on this curve where t is time. Thus $\mathbf{R}(t)$ equals the point on the curve which occurs at time t. Assume that $\mathbf{R}', \mathbf{R}''$ exist and is continuous. Thus $\mathbf{R}' = \mathbf{v}$, the velocity and $\mathbf{R}'' = \mathbf{a}$ is the acceleration.

Lemma 25.1.1 *Define* $\mathbf{T}(t) \equiv \mathbf{R}'(t) / |\mathbf{R}'(t)|$. *Then* $|\mathbf{T}(t)| = 1$ *and if* $\mathbf{T}'(t) \neq 0$, *then there exists a unit vector,* $\mathbf{N}(t)$ *called the principle normal perpendicular to* $\mathbf{T}(t)$ *and a scalar valued function,* $\kappa(t)$ *with* $\mathbf{T}'(t) = \kappa(t) |\mathbf{v}| \mathbf{N}(t)$.

Proof: It follows from the definition that $|\mathbf{T}| = 1$. Therefore, $\mathbf{T} \cdot \mathbf{T} = 1$ and so, upon differentiating both sides,

$$\mathbf{T}' \cdot \mathbf{T} + \mathbf{T} \cdot \mathbf{T}' = 2\mathbf{T}' \cdot \mathbf{T} = 0.$$

Therefore, $\mathbf{T}'$ is perpendicular to $\mathbf{T}$. Let

$$\mathbf{N}(t) \equiv \frac{\mathbf{T}'}{|\mathbf{T}'|}.$$

Then letting $|\mathbf{T}'| \equiv \kappa(t) |\mathbf{v}(t)|$, it follows

$$\mathbf{T}'(t) = \kappa(t) |\mathbf{v}(t)| \mathbf{N}(t).$$

This proves the lemma.

The plane determined by the two vectors, $\mathbf{T}$ and $\mathbf{N}$ is called the osculating[1] plane. It identifies a particular plane which is in a sense tangent to this space curve. In the case where $|\mathbf{T}'(t)| = 0$ near the point of interest, $\mathbf{T}(t)$ equals a constant and so the space curve is a straight line which it would be supposed has no curvature. Also, the principal normal is undefined in this case. This makes sense because if there is no curving going on, there is no special direction normal to the curve at such points which could be distinguished from any other direction normal to the curve. In the case where $|\mathbf{T}'(t)| = 0$, $\kappa(t) = 0$ and the radius of curvature would be considered infinite.

Definition 25.1.2 *The vector, $\mathbf{T}(t)$ is called the unit tangent vector and the vector, $\mathbf{N}(t)$ is called the principal normal. The function, $\kappa(t)$ in the above lemma is called the curvature. The radius of curvature is defined as $\rho = 1/\kappa$.*

The important thing about this is that it is possible to write the acceleration as the sum of two vectors, one perpendicular to the direction of motion and the other in the direction of motion.

Theorem 25.1.3 *For $\mathbf{R}(t)$ the position vector of a space curve, the acceleration is given by the formula*

$$\begin{aligned} \mathbf{a} &= \frac{d|\mathbf{v}|}{dt}\mathbf{T} + \kappa|\mathbf{v}|^2\mathbf{N} \\ &\equiv a_T\mathbf{T} + a_N\mathbf{N}. \end{aligned} \tag{25.1}$$

Furthermore, $a_T^2 + a_N^2 = |\mathbf{a}|^2$.

Proof:

$$\begin{aligned} \mathbf{a} &= \frac{d\mathbf{v}}{dt} = \frac{d}{dt}(\mathbf{R}') = \frac{d}{dt}(|\mathbf{v}|\mathbf{T}) \\ &= \frac{d|\mathbf{v}|}{dt}\mathbf{T} + |\mathbf{v}|\mathbf{T}' \\ &= \frac{d|\mathbf{v}|}{dt}\mathbf{T} + |\mathbf{v}|^2\kappa\mathbf{N}. \end{aligned}$$

This proves the first part.

For the second part,

$$\begin{aligned} |\mathbf{a}|^2 &= (a_T\mathbf{T} + a_N\mathbf{N}) \cdot (a_T\mathbf{T} + a_N\mathbf{N}) \\ &= a_T^2\mathbf{T}\cdot\mathbf{T} + 2a_Na_T\mathbf{T}\cdot\mathbf{N} + a_N^2\mathbf{N}\cdot\mathbf{N} \\ &= a_T^2 + a_N^2 \end{aligned}$$

because $\mathbf{T}\cdot\mathbf{N} = 0$. This proves the theorem.

Finally, it is well to point out that the curvature is a property of the curve itself, and does not depend on the paramterization of the curve. If the curve is given by two different vector valued functions, $\mathbf{R}(t)$ and $\mathbf{R}(\tau)$, then from the formula above for the curvature,

$$\kappa(t) = \frac{|\mathbf{T}'(t)|}{|\mathbf{v}(t)|} = \frac{\left|\frac{d\mathbf{T}}{d\tau}\frac{d\tau}{dt}\right|}{\left|\frac{d\mathbf{R}}{d\tau}\frac{d\tau}{dt}\right|} = \frac{\left|\frac{d\mathbf{T}}{d\tau}\right|}{\left|\frac{d\mathbf{R}}{d\tau}\right|} \equiv \kappa(\tau).$$

From this, it is possible to give an important formula from physics. Suppose an object orbits a point at constant speed, v. In the above notation, $|\mathbf{v}| = v$. What is the centripetal

[1] To osculate means to kiss. Thus this plane could be called the kissing plane. However, that does not sound formal enough so we call it the osculating plane.

acceleration of this object? You may know from a physics class that the answer is v^2/r where r is the radius. This follows from the above quite easily. The parameterization of the object which is as described is

$$\mathbf{R}(t) = \left(r\cos\left(\frac{v}{r}t\right), r\sin\left(\frac{v}{r}t\right)\right).$$

Therefore, $\mathbf{T} = \left(-\sin\left(\frac{v}{r}t\right), \cos\left(\frac{v}{r}t\right)\right)$ and $\mathbf{T}' = \left(-\frac{v}{r}\cos\left(\frac{v}{r}t\right), -\frac{v}{r}\sin\left(\frac{v}{r}t\right)\right)$. Thus, $\kappa = |\mathbf{T}'(t)|/v = \frac{1}{r}$. It follows

$$\mathbf{a} = \frac{dv}{dt}\mathbf{T} + v^2\kappa\mathbf{N} = \frac{v^2}{r}\mathbf{N}.$$

The vector, $\mathbf{N}$ points from the object toward the center of the circle because it is a positive multiple of the vector, $\left(-\frac{v}{r}\cos\left(\frac{v}{r}t\right), -\frac{v}{r}\sin\left(\frac{v}{r}t\right)\right)$.

Formula (25.1) also yields an easy way to find the curvature. Take the cross product of both sides with $\mathbf{v}$, the velocity. Then

$$\begin{aligned}\mathbf{a}\times\mathbf{v} &= \frac{d|\mathbf{v}|}{dt}\mathbf{T}\times\mathbf{v} + |\mathbf{v}|^2\kappa\mathbf{N}\times\mathbf{v}\\ &= \frac{d|\mathbf{v}|}{dt}\mathbf{T}\times\mathbf{v} + |\mathbf{v}|^3\kappa\mathbf{N}\times\mathbf{T}\end{aligned}$$

Now $\mathbf{T}$ and $\mathbf{v}$ have the same direction so the first term on the right equals zero. Taking the magnitude of both sides, and using the fact that $\mathbf{N}$ and $\mathbf{T}$ are two perpendicular unit vectors,

$$|\mathbf{a}\times\mathbf{v}| = |\mathbf{v}|^3\kappa$$

and so

$$\kappa = \frac{|\mathbf{a}\times\mathbf{v}|}{|\mathbf{v}|^3}. \tag{25.2}$$

Example 25.1.4 *Let $\mathbf{R}(t) = \left(\cos(t), t, t^2\right)$ for $t \in [0,3]$. Find the speed, velocity, curvature, and write the acceleration in terms of normal and tangential components.*

First of all $\mathbf{v}(t) = (-\sin t, 1, 2t)$ and so the speed is given by

$$|\mathbf{v}| = \sqrt{\sin^2(t) + 1 + 4t^2}.$$

Therefore,

$$a_T = \frac{d}{dt}\left(\sqrt{\sin^2(t) + 1 + 4t^2}\right) = \frac{\sin(t)\cos(t) + 4t}{\sqrt{(2 + 4t^2 - \cos^2 t)}}.$$

It remains to find a_N. To do this, you can find the curvature first if you like.

$$\mathbf{a}(t) = \mathbf{R}''(t) = (-\cos t, 0, 2).$$

Then

$$\begin{aligned}\kappa &= \frac{|(-\cos t, 0, 2)\times(-\sin t, 1, 2t)|}{\left(\sqrt{\sin^2(t) + 1 + 4t^2}\right)^3}\\ &= \frac{\sqrt{4 + (-2\sin(t) + 2(\cos(t))t)^2 + \cos^2(t)}}{\left(\sqrt{\sin^2(t) + 1 + 4t^2}\right)^3}\end{aligned}$$

Then

$$\begin{aligned} a_N &= \kappa |\mathbf{v}|^2 \\ &= \frac{\sqrt{4+\left(-2\sin(t)+2(\cos(t))t\right)^2+\cos^2(t)}}{\left(\sqrt{\sin^2(t)+1+4t^2}\right)^3}\left(\sin^2(t)+1+4t^2\right) \\ &= \frac{\sqrt{4+\left(-2\sin(t)+2(\cos(t))t\right)^2+\cos^2(t)}}{\sqrt{\sin^2(t)+1+4t^2}}. \end{aligned}$$

You can observe the formula $a_N^2 + a_T^2 = |\mathbf{a}|^2$ holds. Indeed $a_N^2 + a_T^2 =$

$$\begin{aligned} &\left(\frac{\sqrt{4+\left(-2\sin(t)+2(\cos(t))t\right)^2+\cos^2(t)}}{\sqrt{\sin^2(t)+1+4t^2}}\right)^2+\left(\frac{\sin(t)\cos(t)+4t}{\sqrt{(2+4t^2-\cos^2 t)}}\right)^2 \\ = \;& \frac{4+\left(-2\sin t+2(\cos t)t\right)^2+\cos^2 t}{\sin^2 t+1+4t^2}+\frac{(\sin t\cos t+4t)^2}{2+4t^2-\cos^2 t} = \cos^2 t+4 = |\mathbf{a}|^2 \end{aligned}$$

Example 25.1.5 *Find a formula for the curvature of the curve given by the graph of* $y = f(x)$ *for* $x \in [a,b]$. *Assume whatever you like about smoothness of* f.

You need to write this as a parametric curve. This is most easily accomplished by letting $t = x$. Thus a parameterization is

$$(t, f(t), 0) : t \in [a,b].$$

Then you can use the formula given above. The acceleration is $(0, f''(t), 0)$ and the velocity is $(1, f'(t), 0)$. Therefore,

$$\mathbf{a} \times \mathbf{v} = (0, f''(t), 0) \times (1, f'(t), 0) = (0, 0, -f''(t)).$$

Therefore, the curvature is given by

$$\frac{|\mathbf{a} \times \mathbf{v}|}{|\mathbf{v}|^3} = \frac{|f''(t)|}{\left(1+f'(t)^2\right)^{3/2}}.$$

Sometimes curves don't come to you parametrically. This is unfortunate when it occurs but you can sometimes find a parametric description of such curves. It should be emphasized that it is only sometimes when you can actually find a parameterization. General systems of nonlinear equations cannot be solved using algebra.

Example 25.1.6 *Find a parameterization for the intersection of the surfaces* $y + 3z = 2x^2 + 4$ *and* $y + 2z = x + 1$.

You need to solve for x and y in terms of x. This yields

$$z = 2x^2 - x + 3, \; y = -4x^2 + 3x - 5.$$

Therefore, letting $t = x$, the parameterization is $(x, y, z) = \left(t, -4t^2 - 5 + 3t, -t + 3 + 2t^2\right)$.

Example 25.1.7 *Find a parametrization for the straight line joining* $(3,2,4)$ *and* $(1,10,5)$.

$(x, y, z) = (3,2,4) + t(-2,8,1) = (3-2t, 2+8t, 4+t)$ where $t \in [0,1]$. Note where this came from. The vector, $(-2,8,1)$ is obtained from $(1,10,5) - (3,2,4)$. Now you should check to see this works.

25.2 Geometry Of Space Curves*

If you are interested in more on space curves, you should read this section. Otherwise, procede to the exercises. Denote by $\mathbf{R}(s)$ the function which takes s to a point on this curve where s is arc length. Thus $\mathbf{R}(s)$ equals the point on the curve which occurs when you have traveled a distance of s along the curve from one end. This is known as the parameterization of the curve in terms of arc length. Note also that it incorporates an orientation on the curve because there are exactly two ends you could begin measuring length from. In this section, assume anything about smoothness and continuity to make the following manipulations valid. In particular, assume that $\mathbf{R}'$ exists and is continuous.

Lemma 25.2.1 *Define* $\mathbf{T}(s) \equiv \mathbf{R}'(s)$. *Then* $|\mathbf{T}(s)| = 1$ *and if* $\mathbf{T}'(s) \neq 0$, *then there exists a unit vector,* $\mathbf{N}(s)$ *perpendicular to* $\mathbf{T}(s)$ *and a scalar valued function,* $\kappa(s)$ *with* $\mathbf{T}'(s) = \kappa(s)\mathbf{N}(s)$.

Proof: First, $s = \int_0^s |\mathbf{R}'(r)|\, dr$ because of the definition of arc length. Therefore, from the fundamental theorem of calculus, $1 = |\mathbf{R}'(s)| = |\mathbf{T}(s)|$. Therefore, $\mathbf{T} \cdot \mathbf{T} = 1$ and so upon differentiating this on both sides, yields $\mathbf{T}' \cdot \mathbf{T} + \mathbf{T} \cdot \mathbf{T}' = 0$ which shows $\mathbf{T} \cdot \mathbf{T}' = 0$. Therefore, the vector, $\mathbf{T}'$ is perpendicular to the vector, $\mathbf{T}$. In case $\mathbf{T}'(s) \neq \mathbf{0}$, let $\mathbf{N}(s) = \frac{\mathbf{T}'(s)}{|\mathbf{T}'(s)|}$ and so $\mathbf{T}'(s) = |\mathbf{T}'(s)|\mathbf{N}(s)$, showing the scalar valued function is $\kappa(s) = |\mathbf{T}'(s)|$. This proves the lemma.

The radius of curvature is defined as $\rho = \frac{1}{\kappa}$. Thus at points where there is a lot of curvature, the radius of curvature is small and at points where the curvature is small, the radius of curvature is large. The plane determined by the two vectors, $\mathbf{T}$ and $\mathbf{N}$ is called the osculating plane. It identifies a particular plane which is in a sense tangent to this space curve. In the case where $|\mathbf{T}'(s)| = 0$ near the point of interest, $\mathbf{T}(s)$ equals a constant and so the space curve is a straight line which it would be supposed has no curvature. Also, the principal normal is undefined in this case. This makes sense because if there is no curving going on, there is no special direction normal to the curve at such points which could be distinguished from any other direction normal to the curve. In the case where $|\mathbf{T}'(s)| = 0$, $\kappa(s) = 0$ and the radius of curvature would be considered infinite.

Definition 25.2.2 *The vector,* $\mathbf{T}(s)$ *is called the unit tangent vector and the vector,* $\mathbf{N}(s)$ *is called the principal normal. The function,* $\kappa(s)$ *in the above lemma is called the curvature. When* $\mathbf{T}'(s) \neq 0$ *so the principal normal is defined, the vector,* $\mathbf{B}(s) \equiv \mathbf{T}(s) \times \mathbf{N}(s)$ *is called the binormal.*

The binormal is normal to the osculating plane and $\mathbf{B}'$ tells how fast this vector changes. Thus it measures the rate at which the curve twists.

Lemma 25.2.3 *Let* $\mathbf{R}(s)$ *be a parameterization of a space curve with respect to arc length and let the vectors,* $\mathbf{T}, \mathbf{N}$, *and* $\mathbf{B}$ *be as defined above. Then* $\mathbf{B}' = \mathbf{T} \times \mathbf{N}'$ *and there exists a scalar function,* $\tau(s)$ *such that* $\mathbf{B}' = \tau\mathbf{N}$.

Proof: From the definition of $\mathbf{B} = \mathbf{T} \times \mathbf{N}$, and you can differentiate both sides and get $\mathbf{B}' = \mathbf{T}' \times \mathbf{N} + \mathbf{T} \times \mathbf{N}'$. Now recall that $\mathbf{T}'$ is a multiple called curvature multiplied by $\mathbf{N}$ so the vectors, $\mathbf{T}'$ and $\mathbf{N}$ have the same direction and $\mathbf{B}' = \mathbf{T} \times \mathbf{N}'$. Therefore, $\mathbf{B}'$ is either zero or is perpendicular to $\mathbf{T}$. But also, from the definition of $\mathbf{B}, \mathbf{B}$ is a unit vector and so $\mathbf{B}(s) \cdot \mathbf{B}(s) = 0$. Differentiating this,$\mathbf{B}'(s) \cdot \mathbf{B}(s) + \mathbf{B}(s) \cdot \mathbf{B}'(s) = 0$ showing that $\mathbf{B}'$ is perpendicular to $\mathbf{B}$ also. Therefore, $\mathbf{B}'$ is a vector which is perpendicular to both vectors, $\mathbf{T}$ and $\mathbf{B}$ and since this is in three dimensions, $\mathbf{B}'$ must be some scalar multiple of $\mathbf{N}$ and it is this multiple called τ. Thus $\mathbf{B}' = \tau\mathbf{N}$ as claimed.

Lets go over this last claim a little more. Later it will be done algebraically but for now, the following situation is obtained. There are two vectors, $\mathbf{T}$ and $\mathbf{B}$ which are perpendicular to each other and both $\mathbf{B}'$ and $\mathbf{N}$ are perpendicular to these two vectors, hence perpendicular to the plane determined by them. Therefore, $\mathbf{B}'$ must be a multiple of $\mathbf{N}$. Take a piece of paper, draw two unit vectors on it which are perpendicular. Then you can see that any two vectors which are perpendicular to this plane must be multiples of each other.

The scalar function, τ is called the torsion. In case $\mathbf{T}' = 0$, none of this is defined because in this case there is not a well defined osculating plane. The conclusion of the following theorem is called the Serret Frenet formulas.

Theorem 25.2.4 *(Serret Frenet) Let $\mathbf{R}(s)$ be the parameterization with respect to arc length of a space curve and $\mathbf{T}(s) = \mathbf{R}'(s)$ is the unit tangent vector. Suppose $|\mathbf{T}'(s)| \neq 0$ so the principal normal, $\mathbf{N}(s) = \frac{\mathbf{T}'(s)}{|\mathbf{T}'(s)|}$ is defined. The binormal is the vector $\mathbf{B} \equiv \mathbf{T} \times \mathbf{N}$ so $\mathbf{T}, \mathbf{N}, \mathbf{B}$ forms a right handed system of unit vectors each of which is perpendicular to every other. Then the following system of differential equations holds in $\mathbb{R}^9$.*

$$\mathbf{B}' = \tau\mathbf{N},\ \mathbf{T}' = \kappa\mathbf{N},\ \mathbf{N}' = -\kappa\mathbf{T} - \tau\mathbf{B}$$

where κ is the curvature and is nonnegative and τ is the torsion.

Proof: $\kappa \geq 0$ because $\kappa = |\mathbf{T}'(s)|$. The first two equations are already established. To get the third, note that $\mathbf{B} \times \mathbf{T} = \mathbf{N}$ which follows because $\mathbf{T}, \mathbf{N}, \mathbf{B}$ is given to form a right handed system of unit vectors each perpendicular to the others. (Use your right hand.) Now take the derivative of this expression. thus

$$\begin{aligned} \mathbf{N}' &= \mathbf{B}' \times \mathbf{T} + \mathbf{B} \times \mathbf{T}' \\ &= \tau\mathbf{N} \times \mathbf{T} + \kappa\mathbf{B} \times \mathbf{N}. \end{aligned}$$

Now recall again that $\mathbf{T}, \mathbf{N}, \mathbf{B}$ is a right hand system. Thus $\mathbf{N} \times \mathbf{T} = -\mathbf{B}$ and $\mathbf{B} \times \mathbf{N} = -\mathbf{T}$. This establishes the Frenet Serret formulas.

This is an important example of a system of differential equations in $\mathbb{R}^9$. It is a remarkable result because it says that from knowledge of the two scalar functions, τ and κ, and initial values for $\mathbf{B}, \mathbf{T}$, and $\mathbf{N}$ when $s = 0$ you can obtain the binormal, unit tangent, and principal normal vectors. It is just the solution of an initial value problem of the sort discussed earlier. Having done this, you can reconstruct the entire space curve starting at some point, $\mathbf{R}_0$ because $\mathbf{R}'(s) = \mathbf{T}(s)$ and so $\mathbf{R}(s) = \mathbf{R}_0 + \int_0^s \mathbf{T}'(r)\, dr$.

The vectors, $\mathbf{B}, \mathbf{T}$, and $\mathbf{N}$ are vectors which are functions of position on the space curve. Often, especially in applications, you deal with a space curve which is parameterized by a function of t where t is time. Thus a value of t would correspond to a point on this curve and you could let $\mathbf{B}(t), \mathbf{T}(t)$, and $\mathbf{N}(t)$ be the binormal, unit tangent, and principal normal at this point of the curve. The following example is typical.

Example 25.2.5 *Given the circular helix, $\mathbf{R}(t) = (a\cos t)\,\mathbf{i} + (a\sin t)\,\mathbf{j} + (bt)\,\mathbf{k}$, find the arc length, $s(t)$, the unit tangent vector, $\mathbf{T}(t)$, the principal normal, $\mathbf{N}(t)$, the binormal, $\mathbf{B}(t)$, the curvature, $\kappa(t)$, and the torsion, $\tau(t)$. Here $t \in [0, T]$.*

The arc length is $s(t) = \int_0^t \left(\sqrt{a^2 + b^2}\right) dr = \left(\sqrt{a^2 + b^2}\right) t$. Now the tangent vector is obtained using the chain rule as

$$\begin{aligned} \mathbf{T} &= \frac{d\mathbf{R}}{ds} = \frac{d\mathbf{R}}{dt}\frac{dt}{ds} = \frac{1}{\sqrt{a^2 + b^2}}\mathbf{R}'(t) \\ &= \frac{1}{\sqrt{a^2 + b^2}}\left((-a\sin t)\,\mathbf{i} + (a\cos t)\,\mathbf{j} + b\mathbf{k}\right) \end{aligned}$$

The principal normal:

$$\begin{aligned}\frac{d\mathbf{T}}{ds} &= \frac{d\mathbf{T}}{dt}\frac{dt}{ds} \\ &= \frac{1}{a^2+b^2}\left(\left(-a\cos t\right)\mathbf{i}+\left(-a\sin t\right)\mathbf{j}+0\mathbf{k}\right)\end{aligned}$$

and so

$$\mathbf{N} = \frac{d\mathbf{T}}{ds}\Big/\left|\frac{d\mathbf{T}}{ds}\right| = -\left(\left(\cos t\right)\mathbf{i}+\left(\sin t\right)\mathbf{j}\right)$$

The binormal:

$$\begin{aligned}\mathbf{B} &= \frac{1}{\sqrt{a^2+b^2}}\begin{vmatrix}\mathbf{i} & \mathbf{j} & \mathbf{k}\\ -a\sin t & a\cos t & b\\ -\cos t & -\sin t & 0\end{vmatrix} \\ &= \frac{1}{\sqrt{a^2+b^2}}\left(\left(b\sin t\right)\mathbf{i}-b\cos t\mathbf{j}+a\mathbf{k}\right)\end{aligned}$$

Now the curvature, $\kappa\left(t\right)=\left|\frac{d\mathbf{T}}{ds}\right|=\sqrt{\left(\frac{a\cos t}{a^2+b^2}\right)^2+\left(\frac{a\sin t}{a^2+b^2}\right)^2}=\frac{a}{a^2+b^2}$. Note the curvature is constant in this example. The final task is to find the torsion. Recall that $\mathbf{B}'=\tau\mathbf{N}$ where the derivative on $\mathbf{B}$ is taken with respect to arc length. Therefore, remembering that t is a function of s,

$$\begin{aligned}\mathbf{B}'\left(s\right) &= \frac{1}{\sqrt{a^2+b^2}}\left(\left(b\cos t\right)\mathbf{i}+\left(b\sin t\right)\mathbf{j}\right)\frac{dt}{ds} \\ &= \frac{1}{a^2+b^2}\left(\left(b\cos t\right)\mathbf{i}+\left(b\sin t\right)\mathbf{j}\right) \\ &= \tau\left(-\left(\cos t\right)\mathbf{i}-\left(\sin t\right)\mathbf{j}\right)=\tau\mathbf{N}\end{aligned}$$

and it follows $-b/\left(a^2+b^2\right)=\tau$.

An important application of the usefulness of these ideas involves the decomposition of the acceleration in terms of these vectors of an object moving over a space curve.

Corollary 25.2.6 *Let $\mathbf{R}\left(t\right)$ be a space curve and denote by $\mathbf{v}\left(t\right)$ the velocity, $\mathbf{v}\left(t\right)=\mathbf{R}'\left(t\right)$ and let $v\left(t\right)\equiv\left|\mathbf{v}\left(t\right)\right|$ denote the speed and let $\mathbf{a}\left(t\right)$ denote the acceleration. Then $\mathbf{v}=v\mathbf{T}$ and $\mathbf{a}=\frac{dv}{dt}\mathbf{T}+\kappa v^2\mathbf{N}$.*

Proof: $\mathbf{T}=\frac{d\mathbf{R}}{ds}=\frac{d\mathbf{R}}{dt}\frac{dt}{ds}=\mathbf{v}\frac{dt}{ds}$. Also, $s=\int_0^t v\left(r\right)dr$ and so $\frac{ds}{dt}=v$ which implies $\frac{dt}{ds}=\frac{1}{v}$. Therefore, $\mathbf{T}=\mathbf{v}/v$ which implies $\mathbf{v}=v\mathbf{T}$ as claimed.

Now the acceleration is just the derivative of the velocity and so by the Serrat Frenet formulas,

$$\begin{aligned}\mathbf{a} &= \frac{dv}{dt}\mathbf{T}+v\frac{d\mathbf{T}}{dt} \\ &= \frac{dv}{dt}\mathbf{T}+v\frac{d\mathbf{T}}{ds}v=\frac{dv}{dt}\mathbf{T}+v^2\kappa\mathbf{N}\end{aligned}$$

Note how this decomposes the acceleration into a component tangent to the curve and one which is normal to it. Also note that from the above, $v\left|\mathbf{T}'\right|\frac{\mathbf{T}'(t)}{\left|\mathbf{T}'\right|}=v^2\kappa\mathbf{N}$ and so $\frac{\left|\mathbf{T}'\right|}{v}=\kappa$ and $\mathbf{N}=\frac{\mathbf{T}'(t)}{\left|\mathbf{T}'\right|}$

From this, it is possible to give an important formula from physics. Suppose an object orbits a point at constant speed, v. What is the centripetal acceleration of this object? You

may know from a physics class that the answer is v^2/r where r is the radius. This follows from the above quite easily. The parameterization of the object which is as described is

$$\mathbf{R}(t) = \left(r\cos\left(\frac{v}{r}t\right), r\sin\left(\frac{v}{r}t\right)\right).$$

Therefore, $\mathbf{T} = \left(-\sin\left(\frac{v}{r}t\right), \cos\left(\frac{v}{r}t\right)\right)$ and $\mathbf{T}' = \left(-\frac{v}{r}\cos\left(\frac{v}{r}t\right), -\frac{v}{r}\sin\left(\frac{v}{r}t\right)\right)$. Thus, $\kappa = |\mathbf{T}'(t)|/v = \frac{1}{r}$. It follows

$$\mathbf{a} = \frac{dv}{dt}\mathbf{T} + v^2\kappa\mathbf{N} = \frac{v^2}{r}\mathbf{N}.$$

The vector, $\mathbf{N}$ points from the object toward the center of the circle because it is a positive multiple of the vector, $\left(-\frac{v}{r}\cos\left(\frac{v}{r}t\right), -\frac{v}{r}\sin\left(\frac{v}{r}t\right)\right)$.

25.3 Exercises

1. Find a parametrization for the intersection of the planes $2x + y + 3z = -2$ and $3x - 2y + z = -4$.

2. Find a parametrization for the intersection of the plane $3x + y + z = -3$ and the circular cylinder $x^2 + y^2 = 1$.

3. Find a parametrization for the intersection of the plane $4x + 2y + 3z = 2$ and the elliptic cylinder $x^2 + 4z^2 = 9$.

4. Find a parametrization for the straight line joining $(1, 2, 1)$ and $(-1, 4, 4)$.

5. Find a parametrization for the intersection of the surfaces $3y + 3z = 3x^2 + 2$ and $3y + 2z = 3$.

6. Find a formula for the curvature of the curve, $y = \sin x$ in the xy plane.

7. Find a formula for the curvature of the space curve in $\mathbb{R}^2$, $(x(t), y(t))$.

8. An object moves over the helix, $(\cos 3t, \sin 3t, 5t)$. Find the normal and tangential components of the acceleration of this object as a function of t and write the acceleration in the form $a_T\mathbf{T} + a_N\mathbf{N}$.

9. An object moves over the helix, $(\cos t, \sin t, t)$. Find the normal and tangential components of the acceleration of this object as a function of t and write the acceleration in the form $a_T\mathbf{T} + a_N\mathbf{N}$.

10. An object moves in $\mathbb{R}^3$ according to the formula, $\left(\cos 3t, \sin 3t, t^2\right)$. Find the normal and tangential components of the acceleration of this object as a function of t and write the acceleration in the form $a_T\mathbf{T} + a_N\mathbf{N}$.

11. An object moves over the helix, $(\cos t, \sin t, 2t)$. Find the osculating plane at the point of the curve corresponding to $t = \pi/4$.

12. An object moves over a circle of radius r according to the formula,

$$\mathbf{r}(t) = (r\cos(\omega t), r\sin(\omega t))$$

where $v = r\omega$. Show that the speed of the object is constant and equals to v. Tell why $a_T = 0$ and find a_N, $\mathbf{N}$. This yields the formula for centripetal acceleration from beginning physics classes.

13. Suppose $|\mathbf{R}(t)| = c$ where c is a constant and $\mathbf{R}(t)$ is the position vector of an object. Show the velocity, $\mathbf{R}'(t)$ is always perpendicular to $\mathbf{R}(t)$.

14. An object moves in three dimensions and the only force on the object is a central force. This means that if $\mathbf{r}(t)$ is the position of the object, $\mathbf{a}(t) = k(\mathbf{r}(t))\mathbf{r}(t)$ where k is some function. Show that if this happens, then the motion of the object must be in a plane. **Hint:** First argue that $\mathbf{a} \times \mathbf{r} = \mathbf{0}$. Next show that $(\mathbf{a} \times \mathbf{r}) = (\mathbf{v} \times \mathbf{r})'$. Therefore, $(\mathbf{v} \times \mathbf{r})' = \mathbf{0}$. Explain why this requires $\mathbf{v} \times \mathbf{r} = \mathbf{c}$ for some vector, $\mathbf{c}$ which does not depend on t. Then explain why $\mathbf{c} \cdot \mathbf{r} = 0$. This implies the motion is in a plane. Why? What are some examples of central forces?

15. Let $\mathbf{R}(t) = (\cos t)\mathbf{i} + (\cos t)\mathbf{j} + \left(\sqrt{2}\sin t\right)\mathbf{k}$. Find the arc length, s as a function of the parameter, t, if $t = 0$ is taken to correspond to $s = 0$.

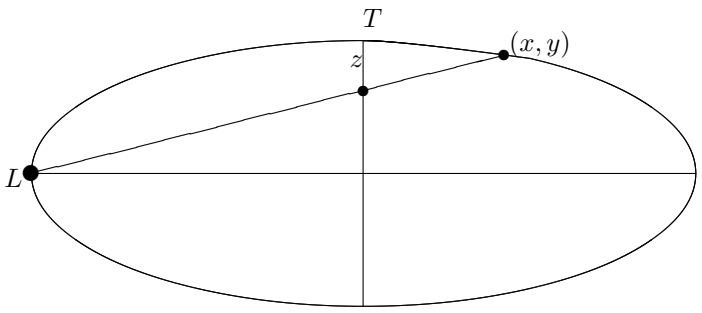
16. Let $\mathbf{R}(t) = 2\mathbf{i} + (4t+2)\mathbf{j} + 4t\mathbf{k}$. Find the arc length, s as a function of the parameter, t, if $t = 0$ is taken to correspond to $s = 0$.

17. Let $\mathbf{R}(t) = e^{5t}\mathbf{i} + e^{-5t}\mathbf{j} + 5\sqrt{2}t\mathbf{k}$. Find the arc length, s as a function of the parameter, t, if $t = 0$ is taken to correspond to $s = 0$.

18. An object moves along the x axis toward $(0,0)$ and then along the curve $y = x^2$ in the direction of increasing x at constant speed. Is the force acting on the object a continuous function? Explain. Is there any physically reasonable way to make this force continuous by relaxing the requirement that the object move at constant speed? If the curve were part of a railroad track, what would happen at the point where $x = 0$?

19. An object of mass m moving over a space curve is acted on by a force, $\mathbf{F}$. Show the work done by this force equals ma_T (length of the curve). In other words, it is only the tangential component of the force which does work.

20. The edge of an elliptical skating rink represented in the following picture has a light at its left end and satisfies the equation $\frac{x^2}{900} + \frac{y^2}{256} = 1$. (Distances measured in yards.)



A hockey puck slides from the point, T towards the center of the rink at the rate of 2 yards per second. What is the speed of its shadow along the wall when $z = 8$? **Hint:** You need to find $\sqrt{x'^2 + y'^2}$ at the instant described.

25.4 Independence Of Parameterization*

Recall that if $\mathbf{p}(t) : t \in [a,b]$ was a parameterization of a smooth curve, C, the length of C is defined as

$$\int_a^b |\mathbf{p}'(t)|\, dt$$

If some other parameterization were used to trace out C, would the same answer be obtained? To answer this question in a satisfactory manner requires some hard calculus.

25.4.1 Hard Calculus

Definition 25.4.1 *A sequence* $\{a_n\}_{n=1}^{\infty}$ *converges to* a,

$$\lim_{n\to\infty} a_n = a \text{ or } a_n \to a$$

if and only if for every $\varepsilon > 0$ *there exists* n_ε *such that whenever* $n \geq n_\varepsilon$,

$$|a_n - a| < \varepsilon.$$

In words the definition says that given any measure of closeness, ε, the terms of the sequence are eventually all this close to a. Note the similarity with the concept of limit. Here, the word "eventually" refers to n being sufficiently large. Earlier, it referred to y being sufficiently close to x on one side or another or else x being sufficiently large in either the positive or negative directions. The limit of a sequence, if it exists, is unique.

Theorem 25.4.2 *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_ε such that if $n \geq n_\varepsilon$, then $|a_n - a| < \varepsilon$ and $|a_n - a_1| < \varepsilon$. Therefore, for such n,

$$\begin{aligned} |a_1 - a| &\leq |a_1 - a_n| + |a_n - a| \\ &< \varepsilon + \varepsilon < |a_1 - a|/2 + |a_1 - a|/2 = |a_1 - a|, \end{aligned}$$

a contradiction.

Definition 25.4.3 *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 index used to define the sequence* $\{a_n\}$. *Then if* $b_k \equiv a_{n_k}$, $\{b_k\}$ *is called a subsequence of* $\{a_n\}$.

Theorem 25.4.4 *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_ε 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.

There is a very useful way of thinking of continuity in terms of limits of sequences found in the following theorem. In words, it says a function is continuous if it takes convergent sequences to convergent sequences whenever possible.

Theorem 25.4.5 *A function $f : D(f) \to \mathbb{R}$ is continuous at $x \in D(f)$ if and only if, whenever $x_n \to x$ with $x_n \in D(f)$, it follows $f(x_n) \to f(x)$.*

Proof: Suppose first that f is continuous at x and let $x_n \to x$. Let $\varepsilon > 0$ be given. By continuity, there exists $\delta > 0$ such that if $|y - x| < \delta$, then $|f(x) - f(y)| < \varepsilon$. However, there exists n_δ such that if $n \geq n_\delta$, then $|x_n - x| < \delta$ and so for all n this large,

$$|f(x) - f(x_n)| < \varepsilon$$

which shows $f(x_n) \to f(x)$.

Now suppose the condition about taking convergent sequences to convergent sequences holds at x. Suppose f fails to be continuous at x. Then there exists $\varepsilon > 0$ and $x_n \in D(f)$ such that $|x - x_n| < \frac{1}{n}$, yet

$$|f(x) - f(x_n)| \geq \varepsilon.$$

But this is clearly a contradiction because, although $x_n \to x$, $f(x_n)$ fails to converge to $f(x)$. It follows f must be continuous after all. This proves the theorem.

Definition 25.4.6 *A set, $K \subseteq \mathbb{R}$ is sequentially compact if whenever $\{a_n\} \subseteq K$ is a sequence, there exists a subsequence, $\{a_{n_k}\}$ such that this subsequence converges to a point of K.*

The following theorem is part of a major advanced calculus theorem known as the Heine Borel theorem.

Theorem 25.4.7 *Every closed interval, $[a, b]$ is sequentially compact.*

Proof: Let $\{x_n\} \subseteq [a, b] \equiv I_0$. Consider the two intervals $\left[a, \frac{a+b}{2}\right]$ and $\left[\frac{a+b}{2}, b\right]$ each of which has length $(b-a)/2$. At least one of these intervals contains x_n for infinitely many values of n. Call this interval I_1. Now do for I_1 what was done for I_0. Split it in half and let I_2 be the interval which contains x_n for infinitely many values of n. Continue this way obtaining a sequence of nested intervals $I_0 \supseteq I_1 \supseteq I_2 \supseteq I_3 \cdots$ where the length of I_n is $(b-a)/2^n$. Now pick n_1 such that $x_{n_1} \in I_1$, n_2 such that $n_2 > n_1$ and $x_{n_2} \in I_2, n_3$ such that $n_3 > n_2$ and $x_{n_3} \in I_3$, etc. (This can be done because in each case the intervals contained x_n for infinitely many values of n.) By the nested interval lemma there exists a point, c contained in all these intervals. Furthermore,

$$|x_{n_k} - c| < (b-a)\, 2^{-k}$$

and so $\lim_{k\to\infty} x_{n_k} = c \in [a, b]$. This proves the theorem.

Lemma 25.4.8 *Let $\phi : [a, b] \to \mathbb{R}$ be a continuous function and suppose ϕ is $1-1$ on (a, b). Then ϕ is either strictly increasing or strictly decreasing on $[a, b]$. Furthermore, ϕ^{-1} is continuous.*

Proof: First it is shown that ϕ is either strictly increasing or strictly decreasing on (a,b).

If ϕ is not strictly decreasing on (a,b), then there exists $x_1 < y_1$, $x_1, y_1 \in (a,b)$ such that

$$(\phi(y_1) - \phi(x_1))(y_1 - x_1) > 0.$$

If for some other pair of points, $x_2 < y_2$ with $x_2, y_2 \in (a,b)$, the above inequality does not hold, then since ϕ is $1-1$,

$$(\phi(y_2) - \phi(x_2))(y_2 - x_2) < 0.$$

Let $x_t \equiv tx_1 + (1-t)x_2$ and $y_t \equiv ty_1 + (1-t)y_2$. Then $x_t < y_t$ for all $t \in [0,1]$ because

$$tx_1 \leq ty_1 \text{ and } (1-t)x_2 \leq (1-t)y_2$$

with strict inequality holding for at least one of these inequalities since not both t and $(1-t)$ can equal zero. Now define

$$h(t) \equiv (\phi(y_t) - \phi(x_t))(y_t - x_t).$$

Since h is continuous and $h(0) < 0$, while $h(1) > 0$, there exists $t \in (0,1)$ such that $h(t) = 0$. Therefore, both x_t and y_t are points of (a,b) and $\phi(y_t) - \phi(x_t) = 0$ contradicting the assumption that ϕ is one to one. It follows ϕ is either strictly increasing or strictly decreasing on (a,b).

This property of being either strictly increasing or strictly decreasing on (a,b) carries over to $[a,b]$ by the continuity of ϕ. Suppose ϕ is strictly increasing on (a,b), a similar argument holding for ϕ strictly decreasing on (a,b). If $x > a$, then pick $y \in (a,x)$ and from the above, $\phi(y) < \phi(x)$. Now by continuity of ϕ at a,

$$\phi(a) = \lim_{x \to a+} \phi(z) \leq \phi(y) < \phi(x).$$

Therefore, $\phi(a) < \phi(x)$ whenever $x \in (a,b)$. Similarly $\phi(b) > \phi(x)$ for all $x \in (a,b)$.

It only remains to verify ϕ^{-1} is continuous. Suppose then that $s_n \to s$ where s_n and s are points of $\phi([a,b])$. It is desired to verify that $\phi^{-1}(s_n) \to \phi^{-1}(s)$. If this does not happen, there exists $\varepsilon > 0$ and a subsequence, still denoted by s_n such that $\left|\phi^{-1}(s_n) - \phi^{-1}(s)\right| \geq \varepsilon$. Using the sequential compactness of $[a,b]$ there exists a further subsequence, still denoted by n, such that $\phi^{-1}(s_n) \to t_1 \in [a,b], t_1 \neq \phi^{-1}(s)$. Then by continuity of ϕ, it follows $s_n \to \phi(t_1)$ and so $s = \phi(t_1)$. Therefore, $t_1 = \phi^{-1}(s)$ after all. This proves the lemma.

Corollary 25.4.9 *Let $f : (a,b) \to \mathbb{R}$ be one to one and continuous. Then $f(a,b)$ is an open interval, (c,d) and $f^{-1} : (c,d) \to (a,b)$ is continuous.*

Proof: Since f is either strictly increasing or strictly decreasing, it follows that $f(a,b)$ is an open interval, (c,d). Assume f is decreasing. Now let $x \in (a,b)$. Why is f^{-1} is continuous at $f(x)$? Since f is decreasing, if $f(x) < f(y)$, then $y \equiv f^{-1}(f(y)) < x \equiv f^{-1}(f(x))$ and so f^{-1} is also decreasing. Let $\varepsilon > 0$ be given. Let $\varepsilon > \eta > 0$ and $(x-\eta, x+\eta) \subseteq (a,b)$. Then $f(x) \in (f(x+\eta), f(x-\eta))$. Let $\delta = \min(f(x) - f(x+\eta), f(x-\eta) - f(x))$. Then if

$$|f(z) - f(x)| < \delta,$$

it follows

$$z \equiv f^{-1}(f(z)) \in (x-\eta, x+\eta) \subseteq (x-\varepsilon, x+\varepsilon)$$

so

$$\left|f^{-1}(f(z)) - x\right| = \left|f^{-1}(f(z)) - f^{-1}(f(x))\right| < \varepsilon.$$

This proves the theorem in the case where f is strictly decreasing. The case where f is increasing is similar.

Theorem 25.4.10 *Let $f : [a,b] \to \mathbb{R}$ be continuous and one to one. Suppose $f'(x_1)$ exists for some $x_1 \in [a,b]$ and $f'(x_1) \neq 0$. Then $\left(f^{-1}\right)'(f(x_1))$ exists and is given by the formula, $\left(f^{-1}\right)'(f(x_1)) = \frac{1}{f'(x_1)}$.*

Proof: By Lemma 25.4.8 f is either strictly increasing or strictly decreasing and f^{-1} is continuous on $[a,b]$. Therefore there exists $\eta > 0$ such that if $0 < |f(x_1) - f(x)| < \eta$, then

$$0 < |x_1 - x| = \left|f^{-1}(f(x_1)) - f^{-1}(f(x))\right| < \delta$$

where δ is small enough that for $0 < |x_1 - x| < \delta$,

$$\left|\frac{x - x_1}{f(x) - f(x_1)} - \frac{1}{f'(x_1)}\right| < \varepsilon.$$

It follows that if $0 < |f(x_1) - f(x)| < \eta$,

$$\left|\frac{f^{-1}(f(x)) - f^{-1}(f(x_1))}{f(x) - f(x_1)} - \frac{1}{f'(x_1)}\right| = \left|\frac{x - x_1}{f(x) - f(x_1)} - \frac{1}{f'(x_1)}\right| < \varepsilon$$

Therefore, since $\varepsilon > 0$ is arbitrary,

$$\lim_{y \to f(x_1)} \frac{f^{-1}(y) - f^{-1}(f(x_1))}{y - f(x_1)} = \frac{1}{f'(x_1)}$$

and this proves the theorem.

The following obvious corollary comes from the above by not bothering with end points.

Corollary 25.4.11 *Let $f : (a,b) \to \mathbb{R}$ be continuous and one to one. Suppose $f'(x_1)$ exists for some $x_1 \in (a,b)$ and $f'(x_1) \neq 0$. Then $\left(f^{-1}\right)'(f(x_1))$ exists and is given by the formula, $\left(f^{-1}\right)'(f(x_1)) = \frac{1}{f'(x_1)}$.*

This is one of those theorems which is very easy to remember if you neglect the difficult questions and simply focus on formal manipulations. Consider the following.

$$f^{-1}(f(x)) = x.$$

Now use the chain rule on both sides to write

$$\left(f^{-1}\right)'(f(x))\, f'(x) = 1,$$

and then divide both sides by $f'(x)$ to obtain

$$\left(f^{-1}\right)'(f(x)) = \frac{1}{f'(x)}.$$

Of course this gives the conclusion of the above theorem rather effortlessly and it is formal manipulations like this which aid many of us in remembering formulas such as the one given in the theorem.

25.4.2 Independence Of Parameterization

Theorem 25.4.12 *Let $\phi : [a,b] \to [c,d]$ be one to one and suppose ϕ' exists and is continuous on $[a,b]$. Then if f is a continuous function defined on $[a,b]$ which is Riemann integrable*[2]*,*

$$\int_c^d f(s)\, ds = \int_a^b f(\phi(t)) \left|\phi'(t)\right| dt$$

[2]Recall that all continuous functions of this sort are Riemann integrable.

Proof: Let $F'(s) = f(s)$. (For example, let $F(s) = \int_a^s f(r)\, dr$.) Then the first integral equals $F(d) - F(c)$ by the fundamental theorem of calculus. By Lemma 5.15.3, ϕ is either strictly increasing or strictly decreasing. Suppose ϕ is strictly decreasing. Then $\phi(a) = d$ and $\phi(b) = c$. Therefore, $\phi' \leq 0$ and the second integral equals

$$-\int_a^b f(\phi(t))\,\phi'(t)\, dt = \int_b^a \frac{d}{dt}\left(F(\phi(t))\right) dt$$

$$= F(\phi(a)) - F(\phi(b)) = F(d) - F(c).$$

The case when ϕ is increasing is similar. This proves the theorem.

Lemma 25.4.13 *Let $\mathbf{f}:[a,b] \to C$, $\mathbf{g}:[c,d] \to C$ be parameterizations of a smooth curve which satisfy conditions 1 - 5. Then $\boldsymbol{\phi}(t) \equiv \mathbf{g}^{-1} \circ \mathbf{f}(t)$ is $1-1$ on (a,b), continuous on $[a,b]$, and either strictly increasing or strictly decreasing on $[a,b]$.*

Proof: It is obvious ϕ is $1-1$ on (a,b) from the conditions $\mathbf{f}$ and $\mathbf{g}$ satisfy. It only remains to verify continuity on $[a,b]$ because then the final claim follows from Lemma 5.15.3. If ϕ is not continuous on $[a,b]$, then there exists a sequence, $\{t_n\} \subseteq [a,b]$ such that $t_n \to t$ but $\phi(t_n)$ fails to converge to $\phi(t)$. Therefore, for some $\varepsilon > 0$ there exists a subsequence, still denoted by n such that $|\phi(t_n) - \phi(t)| \geq \varepsilon$. Using the sequential compactness of $[c,d]$, (See Theorem 5.13.3 on Page 125.) there is a further subsequence, still denoted by n such that $\{\phi(t_n)\}$ converges to a point, s, of $[c,d]$ which is not equal to $\phi(t)$. Thus $\mathbf{g}^{-1} \circ \mathbf{f}(t_n) \to s$ and still $t_n \to t$. Therefore, the continuity of $\mathbf{f}$ and $\mathbf{g}$ imply $\mathbf{f}(t_n) \to \mathbf{g}(s)$ and $\mathbf{f}(t_n) \to \mathbf{f}(t)$. Therefore, $\mathbf{g}(s) = \mathbf{f}(t)$ and so $s = \mathbf{g}^{-1} \circ \mathbf{f}(t) = \phi(t)$, a contradiction. Therefore, ϕ is continuous as claimed.

Theorem 25.4.14 *The length of a smooth curve is not dependent on parameterization.*

Proof: Let C be the curve and suppose $\mathbf{f}:[a,b] \to C$ and $\mathbf{g}:[c,d] \to C$ both satisfy conditions 1 - 5. Is it true that $\int_a^b |\mathbf{f}'(t)|\, dt = \int_c^d |\mathbf{g}'(s)|\, ds$?

Let $\phi(t) \equiv \mathbf{g}^{-1} \circ \mathbf{f}(t)$ for $t \in [a,b]$. Then by the above lemma ϕ is either strictly increasing or strictly decreasing on $[a,b]$. Suppose for the sake of simplicity that it is strictly increasing. The decreasing case is handled similarly.

Let $s_0 \in \phi([a+\delta, b-\delta]) \subset (c,d)$. Then by assumption 4, $g_i'(s_0) \neq 0$ for some i. By continuity of g_i', it follows $g_i'(s) \neq 0$ for all $s \in I$ where I is an open interval contained in $[c,d]$ which contains s_0. It follows that on this interval, g_i is either strictly increasing or strictly decreasing. Therefore, $J \equiv g_i(I)$ is also an open interval and you can define a differentiable function, $h_i : J \to I$ by

$$h_i(g_i(s)) = s.$$

This implies that for $s \in I$,

$$h_i'(g_i(s)) = \frac{1}{g_i'(s)}. \tag{25.3}$$

Now letting $s = \phi(t)$ for $s \in I$, it follows $t \in J_1$, an open interval. Also, for s and t related this way, $\mathbf{f}(t) = \mathbf{g}(s)$ and so in particular, for $s \in I$,

$$g_i(s) = f_i(t).$$

Consequently,

$$s = h_i(f_i(t)) = \phi(t)$$

and so, for $t \in J_1$,

$$\phi'(t) = h_i'(f_i(t)) f_i'(t) = h_i'(g_i(s)) f_i'(t) = \frac{f_i'(t)}{g_i'(\phi(t))} \tag{25.4}$$

which shows that ϕ' exists and is continuous on J_1, an open interval containing $\phi^{-1}(s_0)$. Since s_0 is arbitrary, this shows ϕ' exists on $[a+\delta, b-\delta]$ and is continuous there.

Now $\mathbf{f}(t) = \mathbf{g} \circ \left(\mathbf{g}^{-1} \circ \mathbf{f}\right)(t) = \mathbf{g}(\phi(t))$ and it was just shown that ϕ' is a continuous function on $[a-\delta, b+\delta]$. It follows

$$\mathbf{f}'(t) = \mathbf{g}'(\phi(t)) \phi'(t)$$

and so, by Theorem 25.4.12,

$$\begin{aligned} \int_{\phi(a+\delta)}^{\phi(b-\delta)} |\mathbf{g}'(s)| \, ds &= \int_{a+\delta}^{b-\delta} |\mathbf{g}'(\phi(t))| \left|\phi'(t)\right| dt \\ &= \int_{a+\delta}^{b-\delta} |\mathbf{f}'(t)| \, dt. \end{aligned}$$

Now using the continuity of $\phi, \mathbf{g}'$, and $\mathbf{f}'$ on $[a,b]$ and letting $\delta \to 0+$ in the above, yields

$$\int_c^d |\mathbf{g}'(s)| \, ds = \int_a^b |\mathbf{f}'(t)| \, dt$$

and this proves the theorem.

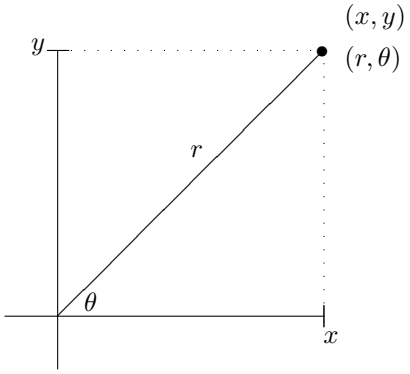
Some Curvilinear Coordinate Systems

26.0.3 Outcomes

1. Recall and use polar coordinates.
2. Graph relations involving polar coordinates.
3. Find the area of regions defined in terms of polar coordinates.
4. Recall and understand the derivation of Kepler's laws.
5. Recall and apply the concept of acceleration in polar coordinates.
6. Recall and use cylindrical and spherical coordinates.

26.1 Polar Coordinates

So far points have been identified in terms of Cartesian coordinates but there are other ways of specifying points in two and three dimensional space. These other ways involve using a list of two or three numbers which have a totally different meaning than Cartesian coordinates to specify a point in two or three dimensional space. In general these lists of numbers which have a different meaning than Cartesian coordinates are called Curvilinear coordinates. Probably the simplest curvilinear coordinate system is that of polar coordinates. The idea is suggested in the following picture.



You see in this picture, the number r identifies the distance of the point from the origin,

$(0,0)$ while θ is the angle shown between the positive x axis and the line from the origin to the point. This angle will always be given in radians and is in the interval $[0, 2\pi)$. Thus the given point, indicated by a small dot in the picture, can be described in terms of the Cartesian coordinates, (x, y) or the polar coordinates, (r, θ). How are the two coordinates systems related? From the picture,

$$x = r\cos(\theta), \ y = r\sin(\theta). \tag{26.1}$$

Example 26.1.1 *The polar coordinates of a point in the plane are* $\left(5, \frac{\pi}{6}\right)$. *Find the Cartesian or rectangular coordinates of this point.*

From (26.1), $x = 5\cos\left(\frac{\pi}{6}\right) = \frac{5}{2}\sqrt{3}$ and $y = 5\sin\left(\frac{\pi}{6}\right) = \frac{5}{2}$. Thus the Cartesian coordinates are $\left(\frac{5}{2}\sqrt{3}, \frac{5}{2}\right)$.

Example 26.1.2 *Suppose the Cartesian coordinates of a point are* $(3, 4)$. *Find the polar coordinates.*

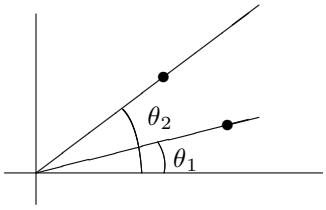
Recall that r is the distance form $(0,0)$ and so $r = 5 = \sqrt{3^2 + 4^2}$. It remains to identify the angle. Note the point is in the first quadrant, (Both the x and y values are positive.) Therefore, the angle is something between 0 and $\pi/2$ and also $3 = 5\cos(\theta)$, and $4 = 5\sin(\theta)$. Therefore, dividing yields $\tan(\theta) = 4/3$. At this point, use a calculator or a table of trigonometric functions to find that at least approximately, $\theta = .927\,295$ radians.

26.1.1 Graphs In Polar Coordinates

Just as in the case of rectangular coordinates, it is possible to use relations between the polar coordinates to specify points in the plane. The process of sketching their graphs is very similar to that used to sketch graphs of functions in rectangular coordinates. I will only consider the case where the relation between the polar coordinates is of the form, $r = f(\theta)$. To graph such a relation, you can make a table of the form

θ	r
θ_1	$f(\theta_1)$
θ_2	$f(\theta_2)$
$\vdots$	$\vdots$

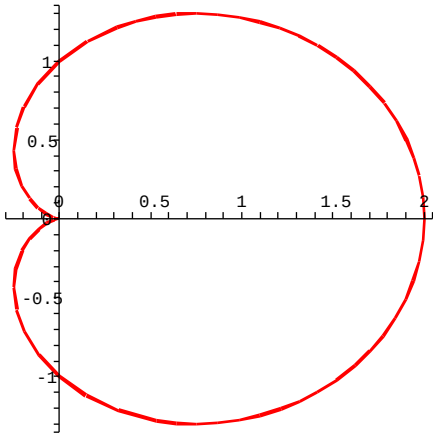
and then graph the resulting points and connect them up with a curve. The following picture illustrates how to begin this process.



To obtain the point in the plane which goes with the pair $(\theta, f(\theta))$, you draw the ray through the origin which makes an angle of θ with the positive x axis. Then you move along this ray a distance of $f(\theta)$ to obtain the point. As in the case with rectangular coordinates, this process is tedious and is best done by a computer algebra system.

Example 26.1.3 *Graph the polar equation,* $r = 1 + \cos\theta$.

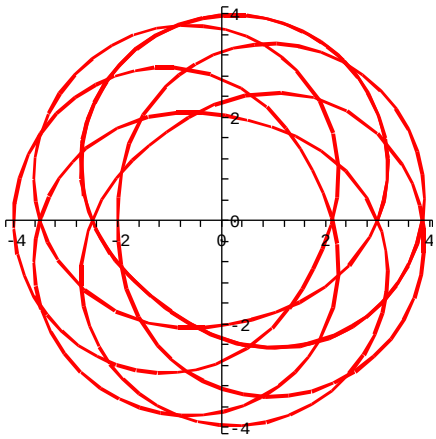
To do this, I will use Maple. The command which produces the polar graph of this is: > plot([1+cos(t),t,t=0..2*Pi],coords=polar); It tells Maple that r is given by $1+\cos(t)$ and that $t \in [0, 2\pi]$. The variable t is playing the role of θ. It is easier to type t than θ in Maple.



You can also see just from your knowledge of the trig. functions that the graph should look something like this. When $\theta = 0, r = 2$ and then as θ increases to $\pi/2$, you see that $\cos\theta$ decreases to 0. Thus the line from the origin to the point on the curve should get shorter as θ goes from 0 to $\pi/2$. Then from $\pi/2$ to π, $\cos\theta$ gets negative eventually equaling -1 at $\theta = \pi$. Thus $r = 0$ at this point. Viewing the graph, you see this is exactly what happens. The above function is called a **cardioid**.

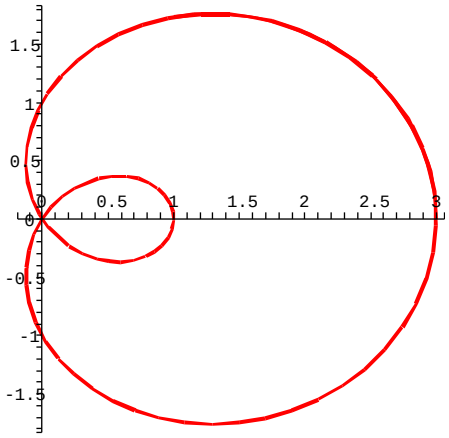
Here is another example. This is the graph obtained from $r = 3 + \sin\left(\frac{7\theta}{6}\right)$.

Example 26.1.4 *Graph* $r = 3 + \sin\left(\frac{7\theta}{6}\right)$ *for* $\theta \in [0, 14\pi]$.



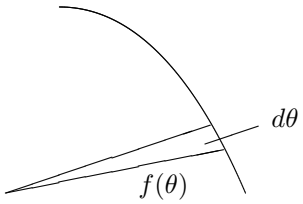
In polar coordinates people sometimes allow r to be negative. When this happens, it means that to obtain the point in the plane, you go in the opposite direction along the ray which starts at the origin and makes an angle of θ with the positive x axis. I do not believe the fussiness occasioned by this extra generality is justified by any sufficiently interesting application so no more will be said about this. It is mainly a fun way to obtain pretty pictures. Here is such an example.

Example 26.1.5 *Graph* $r = 1 + 2\cos\theta$ *for* $\theta \in [0, 2\pi]$.



26.2 The Area In Polar Coordinates

How can you find the area of the region determined by $0 \leq r \leq f(\theta)$ for $\theta \in [a,b]$, assuming this is a well defined set of points in the plane? See Example 26.1.5 with $\theta \in [0,2\pi]$ to see something which it would be better to avoid. I have in mind the situation where every ray through the origin having angle θ for $\theta \in [a,b]$ intersects the graph of $r = f(\theta)$ in exactly one point. To see how to find the area of such a region, consider the following picture.



This is a representation of a small triangle obtained from two rays whose angles differ by only $d\theta$. What is the area of this triangle, dA? It would be

$$\frac{1}{2}\sin(d\theta)\, f(\theta)^2 \approx \frac{1}{2} f(\theta)^2\, d\theta = dA$$

with the approximation getting better as the angle gets smaller. Thus the area should solve the initial value problem,

$$\frac{dA}{d\theta} = \frac{1}{2} f(\theta)^2,\ A(a) = 0.$$

Therefore, the total area would be given by the integral,

$$\frac{1}{2}\int_a^b f(\theta)^2\, d\theta. \tag{26.2}$$

Example 26.2.1 *Find the area of the cardioid, $r = 1 + \cos\theta$ for $\theta \in [0,2\pi]$.*

From the graph of the cardioid presented earlier, you can see the region of interest satisfies the conditions above that every ray intersects the graph in only one point. Therefore, from (26.2) this area is

$$\frac{1}{2}\int_0^{2\pi} (1+\cos(\theta))^2\, d\theta = \frac{3}{2}\pi.$$

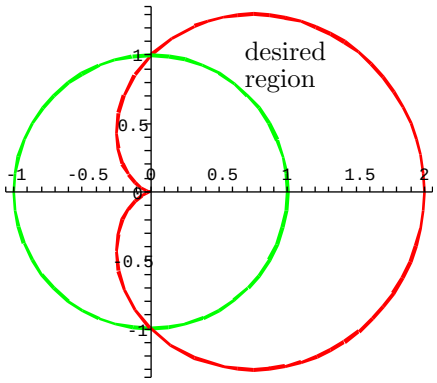
Example 26.2.2 *Verify the area of a circle of radius a is πa^2.*

The polar equation is just $r = a$ for $\theta \in [0, 2\pi]$. Therefore, the area should be

$$\frac{1}{2}\int_0^{2\pi} a^2 d\theta = \pi a^2.$$

Example 26.2.3 *Find the area of the region inside the cardioid, $r = 1 + \cos\theta$ and outside the circle, $r = 1$ for $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.*

As is usual in such cases, it is a good idea to graph the curves involved to get an idea what is wanted.



The area of this region would be the area of the part of the cardioid corresponding to $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ minus the area of the part of the circle in the first quadrant. Thus the area is

$$\frac{1}{2}\int_{-\pi/2}^{\pi/2} (1 + \cos(\theta))^2 d\theta - \frac{1}{2}\int_{-\pi/2}^{\pi/2} 1 d\theta = \frac{1}{4}\pi + 2.$$

This example illustrates the following procedure for finding the area between the graphs of two curves given in polar coordinates.

Procedure 26.2.4 *Suppose that for all $\theta \in [a, b], 0 < g(\theta) < f(\theta)$. To find the area of the region defined in terms of polar coordinates by $g(\theta) < r < f(\theta)$, $\theta \in [a, b]$, you do the following.*

$$\frac{1}{2}\int_a^b \left(f(\theta)^2 - g(\theta)^2\right) d\theta.$$

26.3 Exercises

1. The following are the polar coordinates of points. Find the rectangular coordinates.

   (a) $\left(5, \frac{\pi}{6}\right)$
   (b) $\left(3, \frac{\pi}{3}\right)$
   (c) $\left(4, \frac{2\pi}{3}\right)$
   (d) $\left(2, \frac{3\pi}{4}\right)$
   (e) $\left(3, \frac{7\pi}{6}\right)$

(f) $\left(8, \frac{11\pi}{6}\right)$

2. The following are the rectangular coordinates of points. Find the polar coordinates of these points.

   (a) $\left(\frac{5}{2}\sqrt{2}, \frac{5}{2}\sqrt{2}\right)$
   (b) $\left(\frac{3}{2}, \frac{3}{2}\sqrt{3}\right)$
   (c) $\left(-\frac{5}{2}\sqrt{2}, \frac{5}{2}\sqrt{2}\right)$
   (d) $\left(-\frac{5}{2}, \frac{5}{2}\sqrt{3}\right)$
   (e) $\left(-\sqrt{3}, -1\right)$
   (f) $\left(\frac{3}{2}, -\frac{3}{2}\sqrt{3}\right)$

3. In general it is a stupid idea to try to use algebra to invert and solve for a set of curvilinear coordinates such as polar or cylindrical coordinates in term of Cartesian coordinates. Not only is it often very difficult or even impossible to do it[1], but also it takes you in entirely the wrong direction because the whole point of introducing the new coordinates is to write everything in terms of these new coordinates and not in terms of Cartesian coordinates. However, sometimes this inversion can be done. Describe how to solve for r and θ in terms of x and y in polar coordinates.

4. Suppose $r = \frac{a}{1+e\sin\theta}$ where $e \in [0,1]$. By changing to rectangular coordinates, show this is either a parabola, an ellipse or a hyperbola. Determine the values of e which correspond to the various cases.

5. In Example 26.1.4 suppose you graphed it for $\theta \in [0, k\pi]$ where k is a positive integer. What is the smallest value of k such that the graph will look exactly like the one presented in the example?

6. Suppose you were to graph $r = 3 + \sin\left(\frac{m}{n}\theta\right)$ where m, n are integers. Can you give some description of what the graph will look like for $\theta \in [0, k\pi]$ for k a very large positive integer? How would things change if you did $r = 3 + \sin(r\theta)$ where r is an irrational number?

7. Graph $r = 1 + \sin\theta$ for $\theta \in [0, 2\pi]$.

8. Graph $r = 2 + \sin\theta$ for $\theta \in [0, 2\pi]$.

9. Graph $r = 1 + 2\sin\theta$ for $\theta \in [0, 2\pi]$.

10. Graph $r = 2 + \sin(2\theta)$ for $\theta \in [0, 2\pi]$.

11. Graph $r = 1 + \sin(2\theta)$ for $\theta \in [0, 2\pi]$.

12. Graph $r = 1 + \sin(3\theta)$ for $\theta \in [0, 2\pi]$.

13. Find the area of the bounded region determined by $r = 1 + \sin(3\theta)$ for $\theta \in [0, 2\pi]$.

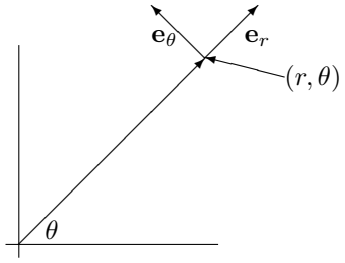
14. Find the area inside $r = 1 + \sin\theta$ and outside the circle $r = 1/2$.

15. Find the area inside the circle $r = 1/2$ and outside the region defined by $r = 1 + \sin\theta$.

[1] It is no problem for these simple cases of curvilinear coordinates. However, it is a major difficulty in general. Algebra is simply not adequate to solve systems of nonlinear equations.

26.4 The Acceleration In Polar Coordinates

Sometimes you have information about forces which act not in the direction of the coordinate axes but in some other direction. When this is the case, it is often useful to express things in terms of different coordinates which are consistent with these directions. A good example of this is the force exerted by the sun on a planet. This force is always directed toward the sun and so the force vector changes as the planet moves. To discuss this, consider the following simple diagram in which two unit vectors, $\mathbf{e}_r$ and $\mathbf{e}_\theta$ are shown.



The vector, $\mathbf{e}_r = (\cos\theta, \sin\theta)$ and the vector, $\mathbf{e}_\theta = (-\sin\theta, \cos\theta)$. You should convince yourself that the picture above corresponds to this definition of the two vectors. Note that $\mathbf{e}_r$ is a unit vector pointing away from $\mathbf{0}$ and

$$\mathbf{e}_\theta = \frac{d\mathbf{e}_r}{d\theta},\ \mathbf{e}_r = -\frac{d\mathbf{e}_\theta}{d\theta}. \tag{26.3}$$

Now consider the position vector from $\mathbf{0}$ of a point in the plane, $\mathbf{r}(t)$. Then

$$\mathbf{r}(t) = r(t)\,\mathbf{e}_r(\theta(t))$$

where $r(t) = |\mathbf{r}(t)|$. Thus $r(t)$ is just the distance from the origin, $\mathbf{0}$ to the point. What is the velocity and acceleration? Using the chain rule,

$$\frac{d\mathbf{e}_r}{dt} = \frac{d\mathbf{e}_r}{d\theta}\theta'(t),\ \frac{d\mathbf{e}_\theta}{dt} = \frac{d\mathbf{e}_\theta}{d\theta}\theta'(t)$$

and so from (26.3),

$$\frac{d\mathbf{e}_r}{dt} = \theta'(t)\,\mathbf{e}_\theta,\ \frac{d\mathbf{e}_\theta}{dt} = -\theta'(t)\,\mathbf{e}_r \tag{26.4}$$

Using (26.4) as needed along with the product rule and the chain rule,

$$\begin{aligned}\mathbf{r}'(t) &= r'(t)\,\mathbf{e}_r + r(t)\,\frac{d}{dt}(\mathbf{e}_r(\theta(t)))\\ &= r'(t)\,\mathbf{e}_r + r(t)\,\theta'(t)\,\mathbf{e}_\theta.\end{aligned}$$

Next consider the acceleration.

$$\begin{aligned}\mathbf{r}''(t) &= r''(t)\,\mathbf{e}_r + r'(t)\,\frac{d\mathbf{e}_r}{dt} + r'(t)\,\theta'(t)\,\mathbf{e}_\theta + r(t)\,\theta''(t)\,\mathbf{e}_\theta + r(t)\,\theta'(t)\,\frac{d}{dt}(\mathbf{e}_\theta)\\ &= r''(t)\,\mathbf{e}_r + 2r'(t)\,\theta'(t)\,\mathbf{e}_\theta + r(t)\,\theta''(t)\,\mathbf{e}_\theta + r(t)\,\theta'(t)\,(-\mathbf{e}_r)\,\theta'(t)\\ &= \left(r''(t) - r(t)\,\theta'(t)^2\right)\mathbf{e}_r + \left(2r'(t)\,\theta'(t) + r(t)\,\theta''(t)\right)\mathbf{e}_\theta.\end{aligned} \tag{26.5}$$

This is a very profound formula. Consider the following examples.

Example 26.4.1 *Suppose an object of mass m moves at a uniform speed, s, around a circle of radius R. Find the force acting on the object.*

By Newton's second law, the force acting on the object is $m\mathbf{r}''$. In this case, $r(t) = R$, a constant and since the speed is constant, $\theta'' = 0$. Therefore, the term in (26.5) corresponding to $\mathbf{e}_\theta$ equals zero and $m\mathbf{r}'' = -R\theta'(t)^2 \mathbf{e}_r$. The speed of the object is s and so it moves s/R radians in unit time. Thus $\theta'(t) = s/R$ and so

$$m\mathbf{r}'' = -mR\left(\frac{s}{R}\right)^2 \mathbf{e}_r = -m\frac{s^2}{R}\mathbf{e}_r.$$

This is the familiar formula for centripetal force from elementary physics, obtained as a very special case of (26.5).

Example 26.4.2 *A platform rotates at a constant speed in the counter clockwise direction and an object of mass m moves from the center of the platform toward the edge at constant speed. What forces act on this object?*

Let v denote the constant speed of the object moving toward the edge of the platform. Then

$$r'(t) = v, r''(t) = 0, \theta''(t) = 0,$$

while $\theta'(t) = \omega$, a positive constant. From (26.5)

$$m\mathbf{r}''(t) = -mr(t)\omega^2 \mathbf{e}_r + m2v\omega\mathbf{e}_\theta.$$

Thus the object experiences centripetal force from the first term and also a funny force from the second term which is in the direction of rotation of the platform. You can observe this by experiment if you like. Go to a playground and have someone spin one of those merry go rounds while you ride it and move from the center toward the edge. The term $2r'\theta'$ is called the Coriolis force.

Suppose at each point of space, $\mathbf{r}$ is associated a force, $\mathbf{F}(\mathbf{r})$ which a given object of mass m will experience if its position vector is $\mathbf{r}$. This is called a force field. a force field is a central force field if $\mathbf{F}(\mathbf{r}) = g(\mathbf{r})\mathbf{e}_r$. Thus in a central force field, the force an object experiences will always be directed toward or away from the origin, $\mathbf{0}$. The following simple lemma is very interesting because it says that in a central force field, objects must move in a plane.

Lemma 26.4.3 *Suppose an object moves in three dimensions in such a way that the only force acting on the object is a central force. Then the motion of the object is in a plane.*

Proof: Let $\mathbf{r}(t)$ denote the position vector of the object. Then from the definition of a central force and Newton's second law,

$$m\mathbf{r}'' = g(\mathbf{r})\mathbf{r}.$$

Therefore, $m\mathbf{r}'' \times \mathbf{r} = m(\mathbf{r}' \times \mathbf{r})' = g(\mathbf{r})\mathbf{r} \times \mathbf{r} = \mathbf{0}$. Therefore, $(\mathbf{r}' \times \mathbf{r}) = \mathbf{n}$, a constant vector and so$\mathbf{r} \cdot \mathbf{n} = \mathbf{r} \cdot (\mathbf{r}' \times \mathbf{r}) = 0$ showing that $\mathbf{n}$ is a normal vector to a plane which contains $\mathbf{r}(t)$ for all t. This proves the lemma.

The next example has as a special case one of Kepler's laws, Kepler's second law , the equal area law.

Example 26.4.4 *An object moves in three dimensions in such a way that the only force acting on the object is a central force. Then the object moves in a plane and the radius vector from the origin to the object sweeps out area at a constant rate.*

The above lemma says the object moves in a plane. From the assumption that the force field is a central force field, it follows from (26.5) that

$$2r'(t)\,\theta'(t) + r(t)\,\theta''(t) = 0$$

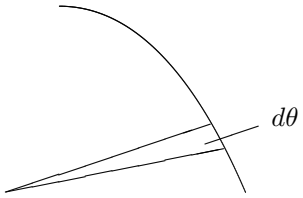
Multiply both sides of this equation by r. This yields

$$2rr'\theta' + r^2\theta'' = \left(r^2\theta'\right)' = 0. \tag{26.6}$$

Consequently,

$$r^2\theta' = c \tag{26.7}$$

for some constant, C. Now consider the following picture.



In this picture, $d\theta$ is the indicated angle and the two lines determining this angle are position vectors for the object at point t and point $t + dt$. The area of the circular sector, dA, is essentially $r^2 d\theta$ and so $dA = \frac{1}{2}r^2 d\theta$. Therefore,

$$\frac{dA}{dt} = \frac{1}{2}r^2\frac{d\theta}{dt} = \frac{c}{2}. \tag{26.8}$$

26.5 Planetary Motion

Kepler's laws of planetary motion state that planets move around the sun along an ellipse, the equal area law described above holds, and there is a formula for the time it takes for the planet to move around the sun. These laws, discovered by Kepler, were shown by Newton to be consequences of his law of gravitation which states that the force acting on a mass, m by a mass, M is given by

$$\mathbf{F} = -GMm\left(\frac{1}{r^3}\right)\mathbf{r} = -\,GMm\left(\frac{1}{r^2}\right)\mathbf{e}_r$$

where r is the distance between centers of mass and $\mathbf{r}$ is the position vector from M to m. Here G is the gravitation constant. This is called an inverse square law. Gravity acts according to this law and so does electrostatic force. The constant, G, is very small when usual units are used and it has been computed using a very delicate experiment. It is now accepted to be

$$6.67 \times 10^{-11} \text{ Newton meter}^2/\text{kilogram}^2.$$

The experiment involved a light source shining on a mirror attached to a quartz fiber from which was suspended a long rod with two equal masses at the ends which were attracted by two larger masses. The gravitation force between the suspended masses and the two large masses caused the fibre to twist ever so slightly and this twisting was measured by observing the deflection of the light reflected from the mirror on a scale placed some distance from the fibre. The constant was first measured successfully by Lord Cavendish in 1798

and the present accepted value was obtained in 1942. Experiments like these are major accomplishments.

In the following argument, M is the mass of the sun and m is the mass of the planet. (It could also be a comet or an asteroid.)

Consider the first of Kepler's laws, the one which states that planets move along ellipses. From Lemma 26.4.3, the motion is in a plane. Now from (26.5) and Newton's second law,

$$\left(r''(t) - r(t)\,\theta'(t)^2\right)\mathbf{e}_r + \left(2r'(t)\,\theta'(t) + r(t)\,\theta''(t)\right)\mathbf{e}_\theta = -\frac{GMm}{m}\left(\frac{1}{r^2}\right)\mathbf{e}_r = -k\left(\frac{1}{r^2}\right)\mathbf{e}_r$$

Thus $k = GM$ and

$$r''(t) - r(t)\,\theta'(t)^2 = -k\left(\frac{1}{r^2}\right),\ 2r'(t)\,\theta'(t) + r(t)\,\theta''(t) = 0. \tag{26.9}$$

As in (26.6), $\left(r^2\theta'\right)' = 0$ and so there exists a constant, c, such that

$$r^2\theta' = c. \tag{26.10}$$

Therefore, also,

$$2rr'\theta' + r^2\theta'' = 0$$

and so

$$\theta'' = \frac{-2rr'\theta'}{r^2} = \frac{-2r'\theta'}{r} = \frac{-2}{r}\frac{dr}{dt}\frac{c}{r^2} \tag{26.11}$$

$$= \frac{-2c}{r^3}\frac{dr}{dt} \tag{26.12}$$

Now consider the first of the above equations. The question of interest is to know how r depends on θ. By the chain rule, regarding r as a function of θ and θ as a function of t,

$$\frac{dr}{d\theta}\frac{d\theta}{dt} = \frac{dr}{dt} \tag{26.13}$$

and by (26.10),

$$\frac{dr}{d\theta} = \frac{c}{r^2}\frac{dr}{dt}. \tag{26.14}$$

Also, by (26.11) and (26.10),

$$\begin{aligned}\frac{d^2\theta}{dt^2} &= \frac{-2c}{r^3}\frac{dr}{dt} = \frac{-2c}{r^3}\left(\frac{dr}{d\theta}\frac{d\theta}{dt}\right)\\ &= \frac{-2c}{r^3}\frac{dr}{d\theta}\left(\frac{c}{r^2}\right) = \frac{-2c^2}{r^5}\frac{dr}{d\theta}\end{aligned}$$

Differentiating (26.13) again with respect to t,

$$\begin{aligned}\frac{d^2r}{dt^2} &= \frac{d^2r}{d\theta^2}\left(\frac{d\theta}{dt}\right)^2 + \frac{dr}{d\theta}\frac{d^2\theta}{dt^2}\\ &= \frac{d^2r}{d\theta^2}\left(\frac{c}{r^2}\right)^2 + \frac{dr}{d\theta}\left(\frac{-2c^2}{r^5}\frac{dr}{d\theta}\right)\\ &= \frac{d^2r}{d\theta^2}\left(\frac{c}{r^2}\right)^2 - \left(\frac{dr}{d\theta}\right)^2\left(\frac{2c^2}{r^5}\right).\end{aligned}$$

It follows that the first equation of (26.9) yields $r''(t) - r(t)\theta'(t)^2 = -k\left(\frac{1}{r^2}\right)$

$$\frac{d^2r}{d\theta^2}\left(\frac{c}{r^2}\right)^2 - \left(\frac{dr}{d\theta}\right)^2\left(\frac{2c^2}{r^5}\right) - r\left(\frac{c}{r^2}\right)^2 = -k\left(\frac{1}{r^2}\right)$$

which is a fairly messy looking differential equation. However, it can be simplified by multiplying both sides by $\frac{r^4}{c^2}$ to get

$$\frac{d^2r}{d\theta^2} - \left(\frac{dr}{d\theta}\right)^2\left(\frac{2}{r}\right) - r = -\frac{k}{c^2}r^2 \tag{26.15}$$

Next consider the above equation in terms of $\rho = r^{-1}$. Thus, from the chain rule,

$$r = \rho^{-1}, \frac{dr}{d\theta} = (-1)\rho^{-2}\frac{d\rho}{d\theta},$$

$$\frac{d^2r}{d\theta^2} = 2\rho^{-3}\left(\frac{d\rho}{d\theta}\right)^2 - \rho^{-2}\frac{d^2\rho}{d\theta^2}.$$

substituting this in to (26.15),

$$\overbrace{2\rho^{-3}\left(\frac{d\rho}{d\theta}\right)^2 - \rho^{-2}\frac{d^2\rho}{d\theta^2}}^{=\frac{d^2r}{d\theta^2}} - \left(\overbrace{(-1)\rho^{-2}\frac{d\rho}{d\theta}}^{=\frac{dr}{d\theta}}\right)^2(2\rho) - \rho^{-1} = -\frac{k}{\rho^2c^2}.$$

Now note that the first and third terms add to zero and so

$$-\rho^{-2}\frac{d^2\rho}{d\theta^2} - \rho^{-1} = -\frac{k}{\rho^2c^2}$$

Multiplying both sides by $-\rho^{-2}$ yields the equation,

$$\frac{d^2\rho}{d\theta^2} + \rho = \frac{k}{c^2},$$

a much more manageable equation. Multiply both sides by $\frac{d\rho}{d\theta}$.

$$\frac{d^2\rho}{d\theta^2}\frac{d\rho}{d\theta} + \rho\frac{d\rho}{d\theta} = \frac{k}{c^2}\frac{d\rho}{d\theta}$$

Then from the product rule,

$$\frac{1}{2}\frac{d}{d\theta}\left(\left(\frac{d\rho}{d\theta}\right)^2 + \rho^2\right) = \frac{k}{c^2}\frac{d\rho}{d\theta}.$$

Therefore, there exists a constant, c_1 such that

$$\frac{1}{2}\left(\left(\frac{d\rho}{d\theta}\right)^2 + \rho^2\right) - \frac{k}{c^2}\rho = c_1$$

and so

$$\begin{aligned}\left(\frac{d\rho}{d\theta}\right)^2 &= 2c_1+\frac{k}{c^2}\rho-\rho^2\\ &= \left(\frac{k^2}{4c^4}+2c_1\right)-\left(\rho-\frac{k}{2c^2}\right)^2\\ &\equiv \delta^2-\left(\rho-\frac{k}{2c^2}\right)^2\end{aligned}$$

Now letting $\rho_1=\rho-\frac{k}{2c^2}$,

$$\frac{1}{\delta^2}\left(\frac{d\rho_1}{d\theta}\right)^2+\left(\frac{\rho_1}{\delta}\right)^2=1$$

which shows that $\left(\frac{1}{\delta}\frac{d\rho_1}{d\theta},\frac{\rho_1}{\delta}\right)$ is a point on the unit circle. Therefore, there exists an angle, $\alpha(\theta)$ such that

$$\frac{d\rho_1}{d\theta}=\delta\cos(\alpha(\theta)),\rho_1=\delta\sin(\alpha(\theta)).$$

Differentiating the second equation with respect to θ,

$$\frac{d\rho_1}{d\theta}=\alpha'(\theta)\,\delta\cos(\alpha(\theta))$$

and so $\alpha'(\theta)=1$. Therefore, $\alpha(\theta)=\theta+\phi$ for some constant, ϕ. Redefining, θ if necessary, (Let $\widetilde{\theta}=\theta+\phi$) it can be assumed that $\phi=0$ so

$$\rho-\frac{k}{2c^2}=\rho_1=\delta\sin\theta.$$

Thus

$$\rho=\frac{k}{c^2}+\delta\sin\theta$$

and so

$$\begin{aligned}r &= \frac{1}{\frac{k}{c^2}+\delta\sin\theta}=\frac{c^2/k}{1+(c^2/k)\,\delta\sin\theta}\\ &= \frac{p\varepsilon}{1+\varepsilon\sin\theta}\end{aligned}\tag{26.16}$$

where

$$\varepsilon=\left(c^2/k\right)\delta\text{ and }p=c^2/k\varepsilon.\tag{26.17}$$

Here all these constants are nonnegative.

Thus

$$r+\varepsilon r\sin\theta=\varepsilon p$$

and so $r=(\varepsilon p-\varepsilon y)$. Then squaring both sides,

$$x^2+y^2=(\varepsilon p-\varepsilon y)^2=\varepsilon^2p^2-2p\varepsilon^2y+\varepsilon^2y^2$$

And so

$$x^2+\left(1-\varepsilon^2\right)y^2=\varepsilon^2p^2-2p\varepsilon^2y.\tag{26.18}$$

In case $\varepsilon=1$, this reduces to the equation of a parabola. If $\varepsilon<1$, this reduces to the equation of an ellipse and if $\varepsilon>1$, this is called a hyperbola. This proves that objects which

are acted on only by a force of the form given in the above example move along hyperbolas, ellipses or circles. The case where $\varepsilon = 0$ corresponds to a circle. The constant, ε is called the eccentricity. This is called Kepler's first law in the case of a planet.

Kepler's third law involves the time it takes for the planet to orbit the sun. From (26.18) you can complete the square and obtain

$$x^2 + \left(1-\varepsilon^2\right)\left(y + \frac{p\varepsilon^2}{1-\varepsilon^2}\right)^2 = \varepsilon^2 p^2 + \frac{p^2\varepsilon^4}{\left(1-\varepsilon^2\right)} = \frac{\varepsilon^2 p^2}{\left(1-\varepsilon^2\right)},$$

and this yields

$$x^2 / \left(\frac{\varepsilon^2 p^2}{1-\varepsilon^2}\right) + \left(y + \frac{p\varepsilon^2}{1-\varepsilon^2}\right)^2 / \left(\frac{\varepsilon^2 p^2}{\left(1-\varepsilon^2\right)^2}\right) = 1. \tag{26.19}$$

Now note this is the equation of an ellipse and that the diameter of this ellipse is

$$\frac{2\varepsilon p}{\left(1-\varepsilon^2\right)} \equiv 2a.$$

This follows because

$$\frac{\varepsilon^2 p^2}{\left(1-\varepsilon^2\right)^2} \geq \frac{\varepsilon^2 p^2}{1-\varepsilon^2}.$$

Now let T denote the time it takes for the planet to make one revolution about the sun. Using this formula, and (26.8) the following equation must hold.

$$\overbrace{\pi \frac{\varepsilon p}{\sqrt{1-\varepsilon^2}} \frac{\varepsilon p}{\left(1-\varepsilon^2\right)}}^{\text{area of ellipse}} = T\frac{c}{2}$$

Therefore,

$$T = \frac{2}{c}\frac{\pi\varepsilon^2 p^2}{\left(1-\varepsilon^2\right)^{3/2}}$$

and so

$$T^2 = \frac{4\pi^2\varepsilon^4 p^4}{c^2\left(1-\varepsilon^2\right)^3}$$

Now using (26.17),

$$\begin{aligned} T^2 &= \frac{4\pi^2\varepsilon^4 p^4}{k\varepsilon p\left(1-\varepsilon^2\right)^3} = \frac{4\pi^2\left(\varepsilon p\right)^3}{k\left(1-\varepsilon^2\right)^3} \\ &= \frac{4\pi^2 a^3}{k} = \frac{4\pi^2 a^3}{GM}. \end{aligned}$$

Written more memorably, this has shown

$$T^2 = \frac{4\pi^2}{GM}\left(\frac{\text{diameter of ellipse}}{2}\right)^3. \tag{26.20}$$

This relationship is known as Kepler's third law. Kepler's second law, the equal area formula, holds for any central force, not just one which satisfies an inverse square law.

26.6 Exercises

1. Suppose you know how the spherical coordinates of a moving point change as a function of t. Can you figure out the velocity of the point? Specifically, suppose $\phi(t) = t, \theta(t) = 1+t$, and $\rho(t) = t$. Find the speed and the velocity of the object in terms of Cartesian coordinates. **Hint:** You would need to find $x'(t), y'(t)$, and $z'(t)$. Then in terms of Cartesian coordinates, the velocity would be $x'(t)\,\mathbf{i} + y'(t)\,\mathbf{j} + z'(t)\,\mathbf{k}$.

2. Find the length of the cardioid, $r = 1 + \cos\theta, \theta \in [0, 2\pi]$. **Hint:** A parameterization is $x(\theta) = (1+\cos\theta)\cos\theta, y(\theta) = (1+\cos\theta)\sin\theta$.

3. In general, show the length of the curve given in polar coordinates by $r = f(\theta), \theta \in [a, b]$ equals $\int_a^b \sqrt{f'(\theta)^2 + f(\theta)^2}\,d\theta$.

4. Suppose the curve given in polar coordinates by $r = f(\theta)$ for $\theta \in [a, b]$ is rotated about the y axis. Find a formula for the resulting surface of revolution.

5. Suppose an object moves in such a way that $r^2\theta'$ is a constant. Show the only force acting on the object is a central force.

6. Explain why low pressure areas rotate counter clockwise in the Northern hemisphere and clockwise in the Southern hemisphere. **Hint:** Note that from the point of view of an observer fixed in space above the North pole, the low pressure area already has a counter clockwise rotation because of the rotation of the earth and its spherical shape. Now consider (26.7). In the low pressure area stuff will move toward the center so r gets smaller. How are things different in the Southern hemisphere?

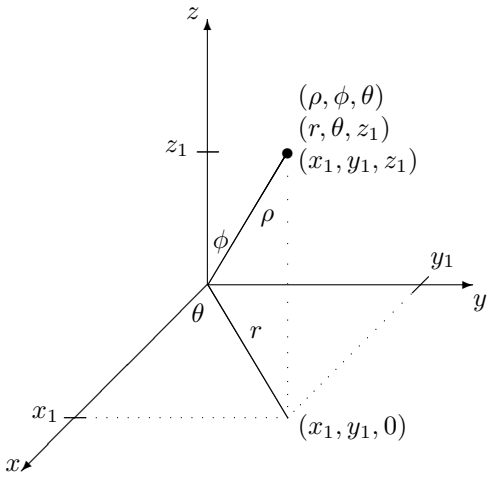
7. What are some physical assumptions which are made in the above derivation of Keplers laws from Newton's laws of motion?

8. The orbit of the earth is pretty nearly circular and the distance from the sun to the earth is about 149×10^6 kilometers. Using (26.20) and the above value of the universal gravitation constant, determine the mass of the sun. The earth goes around it in 365 days. (Actually it is 365.256 days.)

9. It is desired to place a satellite above the equator of the earth which will rotate about the center of mass of the earth every 24 hours. Is it necessary that the orbit be circular? What if you want the satellite to stay above the same point on the earth at all times? If the orbit is to be circular and the satellite is to stay above the same point, at what distance from the center of mass of the earth should the satellite be? You may use that the mass of the earth is 5.98×10^{24} kilograms. Such a satellite is called geosynchronous.

26.7 Spherical And Cylindrical Coordinates

Now consider two three dimensional generalizations of polar coordinates. The following picture serves as motivation for the definition of these two other coordinate systems.



In this picture, ρ is the distance between the origin, the point whose Cartesian coordinates are $(0,0,0)$ and the point indicated by a dot and labeled as (x_1,y_1,z_1), (r,θ,z_1), and (ρ,ϕ,θ). The angle between the positive z axis and the line between the origin and the point indicated by a dot is denoted by ϕ, and θ, is the angle between the positive x axis and the line joining the origin to the point $(x_1,y_1,0)$ as shown, while r is the length of this line. Thus r and θ determine a point in the plane determined by letting $z=0$ and r and θ are the usual polar coordinates. Thus $r\geq 0$ and $\theta\in[0,2\pi)$. Letting z_1 denote the usual z coordinate of a point in three dimensions, like the one shown as a dot, (r,θ,z_1) are the cylindrical coordinates of the dotted point. The spherical coordinates are determined by (ρ,ϕ,θ). When ρ is specified, this indicates that the point of interest is on some sphere of radius ρ which is centered at the origin. Then when ϕ is given, the location of the point is narrowed down to a circle and finally, θ determines which point is on this circle. Let $\phi\in[0,\pi],\theta\in[0,2\pi)$, and $\rho\in[0,\infty)$. The picture shows how to relate these new coordinate systems to Cartesian coordinates. For Cylindrical coordinates,

$$\begin{aligned} x &= r\cos(\theta), \\ y &= r\sin(\theta), \\ z &= z \end{aligned}$$

and for spherical coordinates,

$$\begin{aligned} x &= \rho\sin(\phi)\cos(\theta), \\ y &= \rho\sin(\phi)\sin(\theta), \\ z &= \rho\cos(\phi). \end{aligned}$$

Spherical coordinates should be especially interesting to you because you live on the surface of a sphere. This has been known for several hundred years. You may also know that the standard way to determine position on the earth is to give the longitude and

latitude. The latitude corresponds to ϕ and the longitude corresponds to θ.[2]

Example 26.7.1 *Express the surface, $z = \frac{1}{\sqrt{3}}\sqrt{x^2+y^2}$ in spherical coordinates.*

This is

$$\rho\cos(\phi) = \frac{1}{\sqrt{3}}\sqrt{(\rho\sin(\phi)\cos(\theta))^2 + (\rho\sin(\phi)\sin(\theta))^2} = \frac{1}{3}\sqrt{3}\rho\sin\phi.$$

Therefore, this reduces to

$$\tan\phi = \sqrt{3}$$

and so this is just $\phi = \pi/3$.

Example 26.7.2 *Express the surface, $y = x$ in terms of spherical coordinates.*

This says $\rho\sin(\phi)\sin(\theta) = \rho\sin(\phi)\cos(\theta)$. Thus $\sin\theta = \cos\theta$. You could also write $\tan\theta = 1$.

Example 26.7.3 *Express the surface, $x^2 + y^2 = 4$ in cylindrical coordinates.*

This says $r^2\cos^2\theta + r^2\sin^2\theta = 4$. Thus $r = 2$.

26.8 Exercises

1. The following are the cylindrical coordinates of points. Find the rectangular and spherical coordinates.
   (a) $\left(5, \frac{5\pi}{6}, -3\right)$
   (b) $\left(3, \frac{\pi}{3}, 4\right)$
   (c) $\left(4, \frac{2\pi}{3}, 1\right)$
   (d) $\left(2, \frac{3\pi}{4}, -2\right)$
   (e) $\left(3, \frac{3\pi}{2}, -1\right)$
   (f) $\left(8, \frac{11\pi}{6}, -11\right)$

2. The following are the rectangular coordinates of points. Find the cylindrical and spherical coordinates of these points.
   (a) $\left(\frac{5}{2}\sqrt{2}, \frac{5}{2}\sqrt{2}, -3\right)$
   (b) $\left(\frac{3}{2}, \frac{3}{2}\sqrt{3}, 2\right)$
   (c) $\left(-\frac{5}{2}\sqrt{2}, \frac{5}{2}\sqrt{2}, 11\right)$
   (d) $\left(-\frac{5}{2}, \frac{5}{2}\sqrt{3}, 23\right)$
   (e) $\left(-\sqrt{3}, -1, -5\right)$
   (f) $\left(\frac{3}{2}, -\frac{3}{2}\sqrt{3}, -7\right)$

3. The following are spherical coordinates of points in the form (ρ, ϕ, θ). Find the rectangular and cylindrical coordinates.

[2] Actually latitude is determined on maps and in navigation by measuring the angle from the equator rather than the pole but it is essentially the same idea that we have presented here.

(a) $\left(4, \frac{\pi}{4}, \frac{5\pi}{6}\right)$
(b) $\left(2, \frac{\pi}{3}, \frac{2\pi}{3}\right)$
(c) $\left(3, \frac{5\pi}{6}, \frac{3\pi}{2}\right)$
(d) $\left(4, \frac{\pi}{2}, \frac{7\pi}{4}\right)$
(e) $\left(4, \frac{2\pi}{3}, \frac{\pi}{6}\right)$
(f) $\left(4, \frac{3\pi}{4}, \frac{5\pi}{3}\right)$

4. The following are rectangular coordinates of points. Find the spherical and cylindrical coordinates.

(a) $\left(\sqrt{2}, \sqrt{6}, 2\sqrt{2}\right)$
(b) $\left(-\frac{1}{2}\sqrt{3}, \frac{3}{2}, 1\right)$
(c) $\left(-\frac{3}{4}\sqrt{2}, \frac{3}{4}\sqrt{2}, -\frac{3}{2}\sqrt{3}\right)$
(d) $\left(-\sqrt{3}, 1, 2\sqrt{3}\right)$
(e) $\left(-\frac{1}{4}\sqrt{2}, \frac{1}{4}\sqrt{6}, -\frac{1}{2}\sqrt{2}\right)$
(f) $\left(-\frac{9}{4}\sqrt{3}, \frac{27}{4}, -\frac{9}{2}\right)$

5. Describe how to solve the problem of finding spherical coordinates given rectangular coordinates.

6. A point has Cartesian coordinates, $(1, 2, 3)$. Find its spherical and cylindrical coordinates using a calculator or other electronic gadget.

7. Describe the following surface in rectangular coordinates. $\phi = \pi/4$ where ϕ is the polar angle in spherical coordinates.

8. Describe the following surface in rectangular coordinates. $\theta = \pi/4$ where θ is the angle measured from the postive x axis spherical coordinates.

9. Describe the following surface in rectangular coordinates. $\theta = \pi/4$ where θ is the angle measured from the postive x axis cylindrical coordinates.

10. Describe the following surface in rectangular coordinates. $r = 5$ where r is one of the cylindrical coordinates.

11. Describe the following surface in rectangular coordinates. $\rho = 4$ where ρ is the distance to the origin.

12. Give the cone, $z = \sqrt{x^2 + y^2}$ in cylindrical coordinates and in spherical coordinates.

13. Write the following in spherical coordinates.

(a) $z = x^2 + y^2$.
(b) $x^2 - y^2 = 1$
(c) $z^2 + x^2 + y^2 = 6$
(d) $z = \sqrt{x^2 + y^2}$
(e) $y = x$
(f) $z = x$

14. Write the following in cylindrical coordinates.

    (a) $z = x^2 + y^2$.
    (b) $x^2 - y^2 = 1$
    (c) $z^2 + x^2 + y^2 = 6$
    (d) $z = \sqrt{x^2 + y^2}$
    (e) $y = x$
    (f) $z = x$

Part VI

Vector Calculus In Many Variables

Functions Of Many Variables

27.0.1 Outcomes

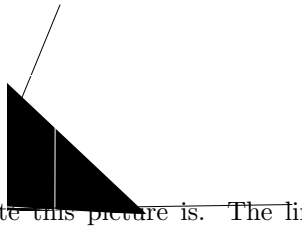
1. Represent a function of two variables by level curves.
2. Identify the characteristics of a function from a graph of its level curves.
3. Recall and use the concept of limit point.
4. Describe the geometrical significance of a directional derivative.
5. Give the relationship between partial derivatives and directional derivatives.
6. Compute partial derivatives and directional derivatives from their definitions.
7. Evaluate higher order partial derivatives.
8. State conditions under which mixed partial derivatives are equal.
9. Verify equations involving partial derivatives.
10. Describe the gradient of a scalar valued function and use to compute the directional derivative.
11. Explain why the directional derivative is maximized in the direction of the gradient and minimized in the direction of minus the gradient.

27.1 The Graph Of A Function Of Two Variables

With vector valued functions of many variables, it doesn't take long before it is impossible to draw meaningful pictures. This is because one needs more than three dimensions to accomplish the task and we can only visualize things in three dimensions. Ultimately, one of the main purposes of calculus is to free us from the tyranny of art. In calculus, we are permitted and even required to think in a meaningful way about things which cannot be drawn. However, it is certainly interesting to consider some things which can be visualized and this will help to formulate and understand more general notions which make sense in contexts which cannot be visualized. One of these is the concept of a scalar valued function of two variables.

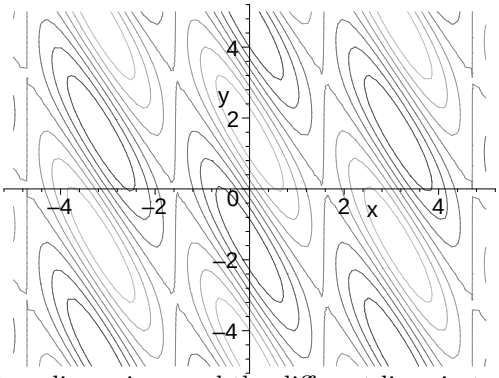
Let $f(x,y)$ denote a scalar valued function of two variables evaluated at the point (x,y). Its graph consists of the set of points, (x,y,z) such that $z = f(x,y)$. How does one go about depicting such a graph? The usual way is to fix one of the variables, say x and consider the function $z = f(x,y)$ where y is allowed to vary and x is fixed. Graphing this would give a curve which lies in the surface to be depicted. Then do the same thing for other

values of x and the result would depict the graph desired graph. Computers do this very well. The following is the graph of the function $z = \cos(x)\sin(2x+y)$ drawn using Maple, a computer algebra system.[1].



Notice how elaborate this picture is. The lines in the drawing correspond to taking one of the variables constant and graphing the curve which results. The computer did this drawing in seconds but you couldn't do it as well if you spent all day on it. I used a grid consisting of 70 choices for x and 70 choices for y.

Sometimes attempts are made to understand three dimensional objects like the above graph by looking at contour graphs in two dimensions. The contour graph of the above three dimensional graph is below and comes from using the computer algebra system again.



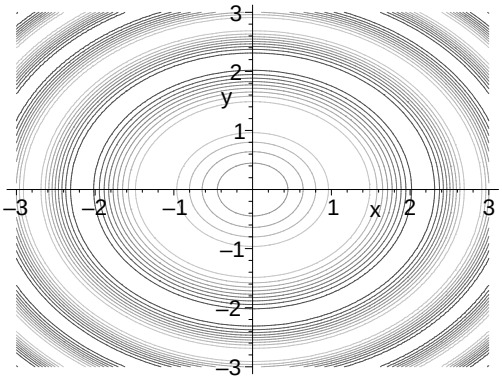
This is in two dimensions and the different lines in two dimensions correspond to points on the three dimensional graph which have the same z value. If you have looked at a weather map, these lines are called isotherms or isobars depending on whether the function involved is temperature or pressure. In a contour geographic map, the contour lines represent constant altitude. If many contour lines are close to each other, this indicates rapid change in the altitude, temperature, pressure, or whatever else may be measured.

A scalar function of three variables, cannot be visualized because four dimensions are required. However, some people like to try and visualize even these examples. This is done by looking at level surfaces in $\mathbb{R}^3$ which are defined as surfaces where the function assumes a constant value. They play the role of contour lines for a function of two variables. As a simple example, consider $f(x,y,z) = x^2+y^2+z^2$. The level surfaces of this function would be concentric spheres centered at $\mathbf{0}$. (Why?) Another way to visualize objects in higher dimensions involves the use of color and animation. However, there really are limits to what you can accomplish in this direction. So much for art.

However, the concept of level curves is quite useful because these can be drawn.

Example 27.1.1 *Determine from a contour map where the function,* $f(x,y) = \sin\left(x^2+y^2\right)$ *is steepest.*

[1] I used Maple and exported the graph as an eps. file which I then imported into this document.



In the picture, the steepest places are where the contour lines are close together because they correspond to various values of the function. You can look at the picture and see where they are close and where they are far. This is the advantage of a contour map.

27.2 Review Of Limits

Recall the concept of limit of a function of many variables. When $\mathbf{f} : D(\mathbf{f}) \to \mathbb{R}^q$ one can only consider in a meaningful way limits at limit points of the set, $D(\mathbf{f})$.

Definition 27.2.1 *Let A denote a nonempty subset of $\mathbb{R}^p$. A point, $\mathbf{x}$ is said to be a limit point of the set, A if for every $r > 0, B(\mathbf{x}, r)$ contains infinitely many points of A.*

Example 27.2.2 *Let S denote the set, $\{(x, y, z) \in \mathbb{R}^3 : x, y, z \text{ are all in } \mathbb{N}\}$. Which points are limit points?*

This set does not have any because any two of these points are at least as far apart as 1. Therefore, if $\mathbf{x}$ is any point of $\mathbb{R}^3, B(\mathbf{x}, 1/4)$ contains at most one point.

Example 27.2.3 *Let U be an open set in $\mathbb{R}^3$. Which points of U are limit points of U?*

They all are. From the definition of U being open, if $\mathbf{x} \in U$, There exists $B(\mathbf{x}, r) \subseteq U$ for some $r > 0$. Now consider the line segment $\mathbf{x} + tr\mathbf{e}_1$ where $t \in [0, 1/2]$. This describes infinitely many points and they are all in $B(\mathbf{x}, r)$ because

$$|\mathbf{x} + tr\mathbf{e}_1 - \mathbf{x}| = tr < r.$$

Therefore, every point of U is a limit point of U.

The case where U is open will be the one of most interest but many other sets have limit points.

Definition 27.2.4 *Let $\mathbf{f} : D(\mathbf{f}) \subseteq \mathbb{R}^p \to \mathbb{R}^q$ where $q, p \geq 1$ be a function and let $\mathbf{x}$ be a limit point of $D(\mathbf{f})$. Then*

$$\lim_{\mathbf{y} \to \mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$$

if and only if the following condition holds. For all $\varepsilon > 0$ there exists $\delta > 0$ such that if

$$0 < |\mathbf{y} - \mathbf{x}| < \delta \text{ and } \mathbf{y} \in D(\mathbf{f})$$

then,

$$|\mathbf{L} - \mathbf{f}(\mathbf{y})| < \varepsilon.$$

The condition that $\mathbf{x}$ must be a limit point of $D(\mathbf{f})$ if you are to take a limit at $\mathbf{x}$ is what makes the limit well defined.

Proposition 27.2.5 *Let $\mathbf{f} : D(\mathbf{f}) \subseteq \mathbb{R}^p \to \mathbb{R}^q$ where $q, p \geq 1$ be a function and let $\mathbf{x}$ be a limit point of $D(\mathbf{f})$. Then if $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y})$ exists, it must be unique.*

Proof: Suppose $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}_1$ and $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}_2$. Then for $\varepsilon > 0$ given, let $\delta_i > 0$ correspond to $\mathbf{L}_i$ in the definition of the limit and let $\delta = \min(\delta_1, \delta_2)$. Since $\mathbf{x}$ is a limit point, there exists $\mathbf{y} \in B(\mathbf{x}, \delta) \cap D(\mathbf{f})$. Therefore,

$$\begin{aligned} |\mathbf{L}_1 - \mathbf{L}_2| &\leq |\mathbf{L}_1 - \mathbf{f}(\mathbf{y})| + |\mathbf{f}(\mathbf{y}) - \mathbf{L}_2| \\ &< \varepsilon + \varepsilon = 2\varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, this shows $\mathbf{L}_1 = \mathbf{L}_2$. The following theorem summarized many important interactions involving continuity. Most of this theorem has been proved in Theorem 23.4.5 on Page 543 and Theorem 23.4.7 on Page 545.

Theorem 27.2.6 *Suppose $\mathbf{x}$ is a limit point of $D(\mathbf{f})$ and $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L}$, $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{g}(\mathbf{y}) = \mathbf{K}$ where $\mathbf{K}$ and $\mathbf{L}$ are vectors in $\mathbb{R}^p$ for $p \geq 1$. Then if $a, b \in \mathbb{R}$,*

$$\lim_{\mathbf{y}\to\mathbf{x}} a\mathbf{f}(\mathbf{y}) + b\mathbf{g}(\mathbf{y}) = a\mathbf{L} + b\mathbf{K}, \tag{27.1}$$

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f} \cdot \mathbf{g}(\mathbf{y}) = \mathbf{L} \cdot \mathbf{K} \tag{27.2}$$

Also, if $\mathbf{h}$ is a continuous function defined near $\mathbf{L}$, then

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{h} \circ \mathbf{f}(\mathbf{y}) = \mathbf{h}(\mathbf{L}). \tag{27.3}$$

For a vector valued function, $\mathbf{f}(\mathbf{y}) = (f_1(\mathbf{y}), \cdots, f_q(\mathbf{y}))$, $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L} = (L_1 \cdots, L_k)^T$ if and only if

$$\lim_{\mathbf{y}\to\mathbf{x}} f_k(\mathbf{y}) = L_k \tag{27.4}$$

for each $k = 1, \cdots, p$.

In the case where $\mathbf{f}$ and $\mathbf{g}$ have values in $\mathbb{R}^3$

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) \times \mathbf{g}(\mathbf{y}) = \mathbf{L} \times \mathbf{K}. \tag{27.5}$$

Also recall Theorem 23.4.6 on Page 545.

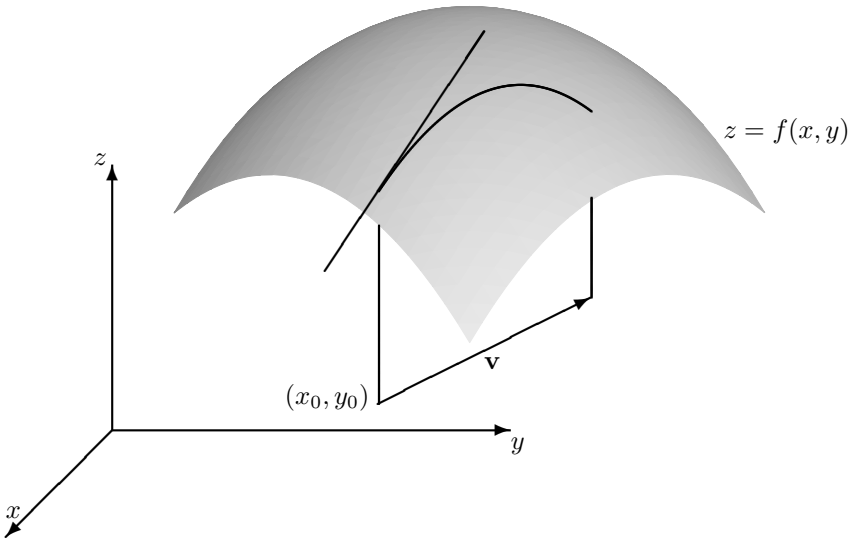
Theorem 27.2.7 *For $\mathbf{f} : D(\mathbf{f}) \to \mathbb{R}^q$ and $\mathbf{x} \in D(\mathbf{f})$ such that $\mathbf{x}$ is a limit point of $D(\mathbf{f})$, it follows $\mathbf{f}$ is continuous at $\mathbf{x}$ if and only if $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x})$.*

27.3 The Directional Derivative And Partial Derivatives

27.3.1 The Directional Derivative

The directional derivative is just what its name suggests. It is the derivative of a function in a particular direction. The following picture illustrates the situation in the case of a

function of two variables.



In this picture, $\mathbf{v} \equiv (v_1, v_2)$ is a unit vector in the xy plane and $\mathbf{x}_0 \equiv (x_0, y_0)$ is a point in the xy plane. When (x, y) moves in the direction of $\mathbf{v}$, this results in a change in $z = f(x, y)$ as shown in the picture. The directional derivative in this direction is defined as

$$\lim_{t \to 0} \frac{f(x_0 + tv_1, y_0 + tv_2) - f(x_0, y_0)}{t}.$$

It tells how fast z is changing in this direction. If you looked at it from the side, you would be getting the slope of the indicated tangent line. A simple example of this is a person climbing a mountain. He could go various directions, some steeper than others. The directional derivative is just a measure of the steepness in a given direction. This motivates the following general definition of the directional derivative.

Definition 27.3.1 *Let $f : U \to \mathbb{R}$ where U is an open set in $\mathbb{R}^n$ and let $\mathbf{v}$ be a unit vector. For $\mathbf{x} \in U$, define the directional derivative of f in the direction, $\mathbf{v}$, at the point $\mathbf{x}$ as*

$$D_{\mathbf{v}} f(\mathbf{x}) \equiv \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t}.$$

Example 27.3.2 *Find the directional derivative of the function, $f(x, y) = x^2 y$ in the direction of $\mathbf{i} + \mathbf{j}$ at the point $(1, 2)$.*

First you need a unit vector which has the same direction as the given vector. This unit vector is $\mathbf{v} \equiv \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$. Then to find the directional derivative from the definition, write the difference quotient described above. Thus $f(\mathbf{x} + t\mathbf{v}) = \left(1 + \frac{t}{\sqrt{2}}\right)^2 \left(2 + \frac{t}{\sqrt{2}}\right)$ and $f(\mathbf{x}) = 2$. Therefore,

$$\frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} = \frac{\left(1 + \frac{t}{\sqrt{2}}\right)^2 \left(2 + \frac{t}{\sqrt{2}}\right) - 2}{t},$$

and to find the directional derivative, you take the limit of this as $t \to 0$. However, this difference quotient equals $\frac{1}{4}\sqrt{2}\left(10 + 4t\sqrt{2} + t^2\right)$ and so, letting $t \to 0$,

$$D_{\mathbf{v}} f(1, 2) = \left(\frac{5}{2}\sqrt{2}\right).$$

There is something you must keep in mind about this. The direction vector must always be a unit vector[2].

27.3.2 Partial Derivatives

There are some special unit vectors which come to mind immediately. These are the vectors, $\mathbf{e}_i$ where

$$\mathbf{e}_i = (0, \cdots, 0, 1, 0, \cdots 0)^T$$

and the 1 is in the i^{th} position.

Definition 27.3.3 *Let U be an open subset of $\mathbb{R}^n$ and let $f : U \to \mathbb{R}$. Then letting $\mathbf{x} = (x_1, \cdots, x_n)^T$ be a typical element of $\mathbb{R}^n$,*

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) \equiv D_{\mathbf{e}_i} f(\mathbf{x}).$$

This is called the partial derivative of f. Thus,

$$\begin{aligned} \frac{\partial f}{\partial x_i}(\mathbf{x}) &\equiv \lim_{t \to 0} \frac{f(\mathbf{x}+t\mathbf{e}_i) - f(\mathbf{x})}{t} \\ &= \lim_{t \to 0} \frac{f(x_1, \cdots, x_i + t, \cdots x_n) - f(x_1, \cdots, x_i, \cdots x_n)}{t}, \end{aligned}$$

and to find the partial derivative, differentiate with respect to the variable of interest and regard all the others as constants. Other notation for this partial derivative is $f_{x_i}, f_{,i}$, or $D_i f$. If $y = f(\mathbf{x})$, the partial derivative of f with respect to x_i may also be denoted by

$$\frac{\partial y}{\partial x_i} \text{ or } y_{x_i}.$$

Example 27.3.4 *Find $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$ if $f(x, y) = y \sin x + x^2 y + z$.*

From the definition above, $\frac{\partial f}{\partial x} = y \cos x + 2xy$, $\frac{\partial f}{\partial y} = \sin x + x^2$, and $\frac{\partial f}{\partial z} = 1$. Having taken one partial derivative, there is no reason to stop doing it. Thus, one could take the partial derivative with respect to y of the partial derivative with respect to x, denoted by $\frac{\partial^2 f}{\partial y \partial x}$ or f_{xy}. In the above example,

$$\frac{\partial^2 f}{\partial y \partial x} = f_{xy} = \cos x + 2x.$$

Also observe that

$$\frac{\partial^2 f}{\partial x \partial y} = f_{yx} = \cos x + 2x.$$

Higher order partial derivatives are defined by analogy to the above. Thus in the above example,

$$f_{yxx} = -\sin x + 2.$$

These partial derivatives, f_{xy} are called mixed partial derivatives.

There is an interesting relationship between the directional derivatives and the partial derivatives, provided the partial derivatives exist and are continuous.

[2]Actually, there is a more general formulation of this known as the Gateaux derivative in which the length of $\mathbf{v}$ is not considered but it will not be considered here.

Definition 27.3.5 *Suppose $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$ where U is an open set and the partial derivatives of f all exist and are continuous on U. Under these conditions, define the gradient of f denoted $\nabla f(\mathbf{x})$ to be the vector*

$$\nabla f(\mathbf{x}) = (f_{x_1}(\mathbf{x}), f_{x_2}(\mathbf{x}), \cdots, f_{x_n}(\mathbf{x}))^T.$$

Proposition 27.3.6 *In the situation of Definition 27.3.5 and for $\mathbf{v}$ a unit vector,*

$$D_{\mathbf{v}} f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v}.$$

This proposition will be proved in a more general setting later. For now, you can use it to compute directional derivatives.

Example 27.3.7 *Find the directional derivative of the function, $f(x,y) = \sin\left(2x^2 + y^3\right)$ at $(1,1)$ in the direction $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T$.*

First find the gradient.

$$\nabla f(x,y) = \left(4x\cos\left(2x^2+y^3\right), 3y^2\cos\left(2x^2+y^3\right)\right)^T.$$

Therefore,

$$\nabla f(1,1) = (4\cos(3), 3\cos(3))^T$$

The directional derivative is therefore,

$$(4\cos(3), 3\cos(3))^T \cdot \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T = \frac{7}{2}(\cos 3)\sqrt{2}.$$

Another important observation is that the gradient gives the direction in which the function changes most rapidly.

Proposition 27.3.8 *In the situation of Definition 27.3.5, suppose $\nabla f(\mathbf{x}) \neq \mathbf{0}$. Then the direction in which f increases most rapidly, that is the direction in which the directional derivative is largest, is the direction of the gradient. Thus $\mathbf{v} = \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$ is the unit vector which maximizes $D_{\mathbf{v}} f(\mathbf{x})$ and this maximum value is $|\nabla f(\mathbf{x})|$. Similarly, $\mathbf{v} = -\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$ is the unit vector which minimizes $D_{\mathbf{v}} f(\mathbf{x})$ and this minimum value is $-|\nabla f(\mathbf{x})|$.*

Proof: Let $\mathbf{v}$ be any unit vector. Then from Proposition 27.3.6,

$$D_{\mathbf{v}} f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v} = |\nabla f(\mathbf{x})|\,|\mathbf{v}| \cos\theta = |\nabla f(\mathbf{x})| \cos\theta$$

where θ is the included angle between these two vectors, $\nabla f(\mathbf{x})$ and $\mathbf{v}$. Therefore, $D_{\mathbf{v}} f(\mathbf{x})$ is maximized when $\cos\theta = 1$ and minimized when $\cos\theta = -1$. The first case corresonds to the angle between the two vectors being 0 which requires they point in the same direction in which case, it must be that $\mathbf{v} = \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$ and $D_{\mathbf{v}} f(\mathbf{x}) = |\nabla f(\mathbf{x})|$. The second case occurs when θ is π and in this case the two vectors point in opposite directions and the directional derivative equals $-|\nabla f(\mathbf{x})|$.

The concept of a directional derivative for a vector valued function is also easy to define although the geometric significance expressed in pictures is not.

Definition 27.3.9 *Let $\mathbf{f} : U \to \mathbb{R}^p$ where U is an open set in $\mathbb{R}^n$ and let $\mathbf{v}$ be a unit vector. For $\mathbf{x} \in U$, define the directional derivative of $\mathbf{f}$ in the direction, $\mathbf{v}$, at the point $\mathbf{x}$ as*

$$D_{\mathbf{v}} \mathbf{f}(\mathbf{x}) \equiv \lim_{t \to 0} \frac{\mathbf{f}(\mathbf{x} + t\mathbf{v}) - \mathbf{f}(\mathbf{x})}{t}.$$

Example 27.3.10 *Let* $\mathbf{f}(x,y) = \left(xy^2, yx\right)^T$. *Find the directional derivative in the direction* $(1,2)^T$ *at the point* (x,y).

First, a unit vector in this direction is $\left(1/\sqrt{5}, 2/\sqrt{5}\right)^T$ and from the definition, the desired limit is

$$\begin{aligned} & \lim_{t\to 0} \frac{\left(\left(x+t\left(1/\sqrt{5}\right)\right)\left(y+t\left(2/\sqrt{5}\right)\right)^2 - xy^2, \left(x+t\left(1/\sqrt{5}\right)\right)\left(y+t\left(2/\sqrt{5}\right)\right) - xy\right)}{t} \\ = & \lim_{t\to 0}\left(\frac{4}{5}xy\sqrt{5} + \frac{4}{5}xt + \frac{1}{5}\sqrt{5}y^2 + \frac{4}{5}ty + \frac{4}{25}t^2\sqrt{5}, \frac{2}{5}x\sqrt{5} + \frac{1}{5}y\sqrt{5} + \frac{2}{5}t\right) \\ = & \left(\frac{4}{5}xy\sqrt{5} + \frac{1}{5}\sqrt{5}y^2, \frac{2}{5}x\sqrt{5} + \frac{1}{5}y\sqrt{5}\right). \end{aligned}$$

You see from this example and the above definition that all you have to do is to form the vector which is obtained by replacing each component of the vector with its directional derivative. In particular, you can take partial derivatives of vector valued functions and use the same notation.

Example 27.3.11 *Find the partial derivative with respect to* x *of the function* $\mathbf{f}(x,y,z,w) = \left(xy^2, z\sin(xy), z^3x\right)^T$.

From the above definition, $\mathbf{f}_x(x,y,z) = D_1\mathbf{f}(x,y,z) = \left(y^2, zy\cos(xy), z^3\right)^T$.

27.4 Mixed Partial Derivatives

Under certain conditions the mixed partial derivatives will always be equal. This astonishing fact is due to Euler in 1734.

Theorem 27.4.1 *Suppose* $f : U \subseteq \mathbb{R}^2 \to \mathbb{R}$ *where* U *is an open set on which* f_x, f_y, f_{xy} *and* f_{yx} *exist. Then if* f_{xy} *and* f_{yx} *are continuous at the point* $(x,y) \in U$, *it follows*

$$f_{xy}(x,y) = f_{yx}(x,y).$$

Proof: Since U is open, there exists $r > 0$ such that $B((x,y),r) \subseteq U$. Now let $|t|, |s| < r/2$ and consider

$$\Delta(s,t) \equiv \frac{1}{st}\{\overbrace{f(x+t,y+s) - f(x+t,y)}^{h(t)} - \overbrace{(f(x,y+s) - f(x,y))}^{h(0)}\}. \tag{27.6}$$

Note that $(x+t, y+s) \in U$ because

$$\begin{aligned} |(x+t,y+s) - (x,y)| &= |(t,s)| = \left(t^2+s^2\right)^{1/2} \\ &\leq \left(\frac{r^2}{4} + \frac{r^2}{4}\right)^{1/2} = \frac{r}{\sqrt{2}} < r. \end{aligned}$$

As implied above, $h(t) \equiv f(x+t,y+s) - f(x+t,y)$. Therefore, by the mean value theorem from calculus and the (one variable) chain rule,

$$\begin{aligned} \Delta(s,t) &= \frac{1}{st}(h(t) - h(0)) = \frac{1}{st}h'(\alpha t)\, t \\ &= \frac{1}{s}\left(f_x(x+\alpha t, y+s) - f_x(x+\alpha t, y)\right) \end{aligned}$$

for some $\alpha \in (0,1)$. Applying the mean value theorem again,

$$\Delta(s,t) = f_{xy}(x+\alpha t, y+\beta s)$$

where $\alpha, \beta \in (0,1)$.

If the terms $f(x+t,y)$ and $f(x,y+s)$ are interchanged in (27.6), $\Delta(s,t)$ is also unchanged and the above argument shows there exist $\gamma, \delta \in (0,1)$ such that

$$\Delta(s,t) = f_{yx}(x+\gamma t, y+\delta s).$$

Letting $(s,t) \to (0,0)$ and using the continuity of f_{xy} and f_{yx} at (x,y),

$$\lim_{(s,t)\to(0,0)} \Delta(s,t) = f_{xy}(x,y) = f_{yx}(x,y).$$

This proves the theorem.

The following is obtained from the above by simply fixing all the variables except for the two of interest.

Corollary 27.4.2 *Suppose U is an open subset of $\mathbb{R}^n$ and $f : U \to \mathbb{R}$ has the property that for two indices, k, l, f_{x_k}, $f_{x_l}, f_{x_l x_k}$, and $f_{x_k x_l}$ exist on U and $f_{x_k x_l}$ and $f_{x_l x_k}$ are both continuous at $\mathbf{x} \in U$. Then $f_{x_k x_l}(\mathbf{x}) = f_{x_l x_k}(\mathbf{x})$.*

It is necessary to assume the mixed partial derivatives are continuous in order to assert they are equal. The following is a well known example [3].

Example 27.4.3 *Let*

$$f(x,y) = \begin{cases} \frac{xy(x^2-y^2)}{x^2+y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$$

From the definition of partial derivatives it follows immediately that $f_x(0,0) = f_y(0,0) = 0$. Using the standard rules of differentiation, for $(x,y) \neq (0,0)$,

$$f_x = y\frac{x^4 - y^4 + 4x^2y^2}{(x^2+y^2)^2},\ f_y = x\frac{x^4 - y^4 - 4x^2y^2}{(x^2+y^2)^2}$$

Now

$$\begin{aligned} f_{xy}(0,0) &\equiv \lim_{y\to 0} \frac{f_x(0,y) - f_x(0,0)}{y} \\ &= \lim_{y\to 0} \frac{-y^4}{(y^2)^2} = -1 \end{aligned}$$

while

$$\begin{aligned} f_{yx}(0,0) &\equiv \lim_{x\to 0} \frac{f_y(x,0) - f_y(0,0)}{x} \\ &= \lim_{x\to 0} \frac{x^4}{(x^2)^2} = 1 \end{aligned}$$

showing that although the mixed partial derivatives do exist at $(0,0)$, they are not equal there.

27.5 Partial Differential Equations

Partial differential equations are equations which involve the partial derivatives of some function. The most famous partial differential equations involve the Laplacian, named after Laplace[3].

Definition 27.5.1 *Let u be a function of n variables. Then $\Delta u \equiv \sum_{k=1}^{n} u_{x_k x_k}$. This is also written as $\nabla^2 u$. The symbol, Δ or ∇^2 is called the Laplacian. When $\Delta u = 0$ the function, u is called harmonic.Laplace's equation is $\Delta u = 0$. The heat equation is $u_t - \Delta u = 0$ and the wave equation is $u_{tt} - \Delta u = 0$.*

Example 27.5.2 *Find the Laplacian of $u(x,y) = x^2 - y^2$.*

$u_{xx} = 2$ while $u_{yy} = -2$. Therefore, $\Delta u = u_{xx} + u_{yy} = 2 - 2 = 0$. Thus this function is harmonic.

Example 27.5.3 *Find $u_t - \Delta u$ where $u(t,x,y) = e^{-t}\cos x$.*

In this case, $u_t = -e^{-t}\cos x$ while $u_{yy} = 0$ and $u_{xx} = -e^{-t}\cos x$ therefore, $u_t - \Delta u = 0$ and so u solves the heat equation.

Example 27.5.4 *Let $u(t,x) = \sin t \cos x$. Find $u_{tt} - \Delta u$.*

In this case, $u_{tt} = -\sin t \cos x$ while $\Delta u = -\sin t \cos x$. Therefore, u is a solution of the wave equation.

27.6 Exercises

1. Find the directional derivative of $f(x,y,z) = x^2 y + z^4$ in the direction of the vector, $(1,3,-1)$ when $(x,y,z) = (1,1,1)$.

2. Find the directional derivative of $f(x,y,z) = \sin\left(x + y^2\right) + z$ in the direction of the vector, $(1,2,-1)$ when $(x,y,z) = (1,1,1)$.

3. Find the directional derivative of $f(x,y,z) = \ln\left(x + y^2\right) + z^2$ in the direction of the vector, $(1,1,-1)$ when $(x,y,z) = (1,1,1)$.

4. Find the largest value of the directional derivative of $f(x,y,z) = \ln\left(x + y^2\right) + z^2$ at the point $(1,1,1)$.

5. Find the smallest value of the directional derivative of $f(x,y,z) = x\sin\left(4xy^2\right) + z^2$ at the point $(1,1,1)$.

6. An ant falls to the top of a stove having temperature $T(x,y) = x^2 \sin(x+y)$ at the point $(2,3)$. In what direction should the ant go to minimize the temperature? In what direction should he go to maximize the temperature?

7. Find the partial derivative with respect to y of the function

$$\mathbf{f}(x,y,z,w) = \left(y^2, z^2\sin(xy), z^3 x\right)^T.$$

[3]Laplace was a great physicist of the 1700's. He made fundamental contributions to mechanics and astronomy.

8. Find the partial derivative with respect to x of the function

$$\mathbf{f}(x,y,z,w) = \left(wx, zx\sin(xy), z^3x\right)^T.$$

9. Find $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$ for $f =$

   (a) $x^2y + \cos(xy) + z^3y$

   (b) $e^{x^2+y^2}z\sin(x+y)$

   (c) $z^2\sin^3\left(e^{x^2+y^3}\right)$

   (d) $x^2\cos\left(\sin\left(\tan\left(z^2+y^2\right)\right)\right)$

   (e) x^{y^2+z}

10. Suppose

$$f(x,y) = \begin{cases} \frac{2xy+6x^3+12xy^2+18yx^2+36y^3+\sin\left(x^3\right)+\tan\left(3y^3\right)}{3x^2+6y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0). \end{cases}$$

    Find $\frac{\partial f}{\partial x}(0,0)$ and $\frac{\partial f}{\partial y}(0,0)$.

11. Why must the vector in the definition of the directional derivative be a unit vector? **Hint:** Suppose not. Would the directional derivative be a correct manifestation of steepness?

12. Find $f_x, f_y, f_z, f_{xy}, f_{yx}, f_{xz}, f_{zx}, f_{zy}, f_{yz}$ for the following and form a conjecture about the mixed partial derivatives.

    (a) $x^2y^3z^4 + \sin(xyz)$

    (b) $\sin(xyz) + x^2yz$

    (c) $z\ln\left|x^2+y^2+1\right|$

    (d) $e^{x^2+y^2+z^2}$

    (e) $\tan(xyz)$

13. Suppose $f : U \to \mathbb{R}$ where U is an open set and suppose that $\mathbf{x} \in U$ has the property that for all $\mathbf{y}$ near $\mathbf{x}$, $f(\mathbf{x}) \leq f(\mathbf{y})$. Prove that if f has all of its partial derivatives at $\mathbf{x}$, then $f_{x_i}(\mathbf{x}) = 0$ for each x_i. **Hint:** This is just a repeat of the usual one variable theorem seen in beginning calculus. You just do this one variable argument for each variable to get the conclusion.

14. As an important application of Problem 13 consider the following. Experiments are done at n times, $t_1, t_2, \cdots, t_n$ and at each time there results a collection of numerical outcomes. Denote by $\{(t_i, x_i)\}_{i=1}^p$ the set of all such pairs and try to find numbers a and b such that the line $x = at + b$ approximates these ordered pairs as well as possible in the sense that out of all choices of a and b, $\sum_{i=1}^p (at_i + b - x_i)^2$ is as small as possible. In other words, you want to minimize the function of two variables, $f(a,b) \equiv \sum_{i=1}^p (at_i + b - x_i)^2$. Find a formula for a and b in terms of the given ordered pairs. You will be finding the formula for the least squares regression line.

15. Show that if $v(x,y) = u(\alpha x, \beta y)$, then $v_x = \alpha u_x$ and $v_y = \beta u_y$. State and prove a generalization to any number of variables.

16. Let f be a function which has continuous derivatives. Show $u(t,x) = f(x-ct)$ solves the wave equation, $u_{tt} - c^2\Delta u = 0$. What about $u(x,t) = f(x+ct)$?

17. D'Alembert found a formula for the solution to the wave equation, $u_{tt} = c^2 u_{xx}$ along with the initial conditions $u(x,0) = f(x), u_t(x,0) = g(x)$. Here is how he did it. He looked for a solution of the form $u(x,t) = h(x+ct) + k(x-ct)$ and then found h and k in terms of the given functions f and g. He ended up with something like

$$u(x,t) = \frac{1}{2c}\int_{x-ct}^{x+ct} g(r)\,dr + \frac{1}{2}\left(f(x+ct) + f(x-ct)\right).$$

Fill in the details.

18. Determine which of the following functions satisfy Laplace's equation.

(a) $x^3 - 3xy^2$
(b) $3x^2y - y^3$
(c) $x^3 - 3xy^2 + 2x^2 - 2y^2$
(d) $3x^2y - y^3 + 4xy$
(e) $3x^2 - y^3 + 4xy$
(f) $3x^2y - y^3 + 4y$
(g) $x^3 - 3x^2y^2 + 2x^2 - 2y^2$

19. Show that if $\Delta u = \lambda u$, then $e^{\lambda t}u$ solves the heat equation, $u_t - \Delta u = 0$.

20. Show that if a, b are scalars and u, v are functions which satisfy Laplace's equation then $au + bv$ also satisfies Laplace's equation. Verify a similar statement for the heat and wave equations.

21. Show that $u(x,t) = \frac{1}{t}e^{-x^2/4c^2t}$ solves the heat equation, $u_t = c^2 u_{xx}$.

The Derivative Of A Function Of Many Variables

28.0.1 Outcomes

1. Define differentiability and explain what the derivative is for a function of n variables.
2. Describe the relation between existence of partial derivatives, continuity, and differentiability.
3. Give examples of functions which have partial derivatives but are not continuous, examples of functions which are differentiable but not C^1, and examples of functions which are continuous without having partial derivatives.
4. Evaluate derivatives of composite functions using the chain rule.
5. Solve related rates problems using the chain rule.

28.1 The Derivative Of Functions Of One Variable

First we review the notion of the derivative of a function of one variable.

Observation 28.1.1 *Suppose a function, f of one variable has a derivative at x. Then*

$$\lim_{h\to 0}\frac{|f(x+h)-f(x)-f'(x)h|}{|h|}=0.$$

This observation follows from the definition of the derivative of a function of one variable, namely

$$f'(x)\equiv\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}.$$

Definition 28.1.2 *A vector valued function of a vector, $\mathbf{v}$ is called $\mathbf{o}(\mathbf{v})$ if*

$$\lim_{|\mathbf{v}|\to 0}\frac{\mathbf{o}(\mathbf{v})}{|\mathbf{v}|}=\mathbf{0}. \tag{28.1}$$

Thus the function $f(x+h)-f(x)-f'(x)h$ is $o(h)$. When we say a function is $o(h)$, it is used like an adjective. It is like saying the function is white or black or green or fat or thin. The term is used very imprecisely. Thus

$$\mathbf{o}(\mathbf{v})=\mathbf{o}(\mathbf{v})+\mathbf{o}(\mathbf{v}),\mathbf{o}(\mathbf{v})=45\mathbf{o}(\mathbf{v}),\mathbf{o}(\mathbf{v})=\mathbf{o}(\mathbf{v})-\mathbf{o}(\mathbf{v}),\text{etc.}$$

When you add two functions with the property of the above definition, you get another one having that same property. When you multiply by 45 the property is also retained as it is when you subtract two such functions. How could something so sloppy be useful? The notation is useful precisely because it prevents you from obsessing over things which are not relevant and should be ignored.

Theorem 28.1.3 *Let $f : (a,b) \to \mathbb{R}$ be a function of one variable. Then $f'(x)$ exists if and only if*

$$f(x+h) - f(x) = ph + o(h) \tag{28.2}$$

In this case, $p = f'(x)$.

Proof: From the above observation it follows that if $f'(x)$ does exist, then (28.2) holds. Suppose then that (28.2) is true. Then

$$\frac{f(x+h) - f(x)}{h} - p = \frac{o(h)}{h}.$$

Taking a limit, you see that

$$p = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

and that in fact this limit exists which shows that $p = f'(x)$. This proves the theorem.

This theorem shows that one way to define $f'(x)$ is as the number, p, if there is one which has the property that

$$f(x+h) = f(x) + ph + o(h).$$

You should think of p as the linear transformation resulting from multiplication by the 1×1 matrix, (p).

Example 28.1.4 *Let $f(x) = x^3$. Find $f'(x)$.*

$f(x+h) = (x+h)^3 = x^3 + 3x^2h + 3xh^2 + h^3 = f(x) + 3x^2h + \left(3xh + h^2\right)h$. Since $\left(3xh + h^2\right)h = o(h)$, it follows $f'(x) = 3x^2$.

Example 28.1.5 *Let $f(x) = \sin(x)$. Find $f'(x)$.*

$$\begin{aligned} f(x+h) - f(x) &= \sin(x+h) - \sin(x) = \sin(x)\cos(h) + \cos(x)\sin(h) - \sin(x) \\ &= \cos(x)\sin(h) + \sin(x)\frac{(\cos(h)-1)}{h}h \\ &= \cos(x)h + \cos(x)\frac{(\sin(h)-h)}{h}h + \sin x\frac{(\cos(h)-1)}{h}h. \end{aligned}$$

Now

$$\cos(x)\frac{(\sin(h)-h)}{h}h + \sin x\frac{(\cos(h)-1)}{h}h = o(h). \tag{28.3}$$

Remember the fundamental limits which allowed you to find the derivative of $\sin(x)$ were

$$\lim_{h \to 0} \frac{\sin(h)}{h} = 1, \ \lim_{h \to 0} \frac{\cos(h) - 1}{h} = 0. \tag{28.4}$$

These same limits are what is needed to verify (28.3).

28.2 The Derivative Of Functions Of Many Variables

This way of thinking about the derivative is exactly what is needed to define the derivative of a function of n variables.

Definition 28.2.1 *Let $\mathbf{f} : U \to \mathbb{R}^p$ where U is an open set in $\mathbb{R}^n$ for $n, p \geq 1$ and let $\mathbf{x} \in U$ be given. Then $\mathbf{f}$ is defined to be differentiable at $\mathbf{x} \in U$ if and only if there exist column vectors, $\mathbf{v}_i$ such that for $\mathbf{h} = (h_1 \cdots, h_n)^T$,*

$$\mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{f}(\mathbf{x}) + \sum_{i=1}^{n} \mathbf{v}_i h_i + \mathbf{o}(\mathbf{h}). \tag{28.5}$$

The derivative of the function, $\mathbf{f}$, denoted by $D\mathbf{f}(\mathbf{x})$, is the linear transformation defined by multiplying by the matrix whose columns are the $p \times 1$ vectors, $\mathbf{v}_i$. Thus if $\mathbf{w}$ is a vector in $\mathbb{R}^n$,

$$D\mathbf{f}(\mathbf{x})\,\mathbf{w} \equiv \begin{pmatrix} | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_n \\ | & & | \end{pmatrix} \mathbf{w}.$$

It is common to think of this matrix as the derivative but strictly speaking, this is incorrect. The derivative is a linear transformation determined by multiplication by this matrix, called the standard matrix because it is based on the standard basis vectors for $\mathbb{R}^n$. The subtle issues involved in a thorough exploration of this issue will be avoided for now. It will be fine to think of the above matrix as the derivative. Other notations which are often used for this matrix or the linear transformation are $\mathbf{f}'(\mathbf{x}), J(\mathbf{x})$, and even $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}$ or $\frac{d\mathbf{f}}{d\mathbf{x}}$.

Theorem 28.2.2 *Suppose $\mathbf{f}$ is as given above in (28.5). Then*

$$\mathbf{v}_k = \lim_{h \to 0} \frac{\mathbf{f}(\mathbf{x}+h\mathbf{e}_k) - \mathbf{f}(\mathbf{x})}{h} \equiv \frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x}),$$

the k^{th} partial derivative.

Proof: Let $\mathbf{h} = (0, \cdots, h, 0, \cdots, 0)^T = h\mathbf{e}_k$ where the h is in the k^{th} slot. Then (28.5) reduces to

$$\mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{f}(\mathbf{x}) + \mathbf{v}_k h + \mathbf{o}(h).$$

Therefore, dividing by h

$$\frac{\mathbf{f}(\mathbf{x}+h\mathbf{e}_k) - \mathbf{f}(\mathbf{x})}{h} = \mathbf{v}_k + \frac{\mathbf{o}(h)}{h}$$

and taking the limit,

$$\lim_{h \to 0} \frac{\mathbf{f}(\mathbf{x}+h\mathbf{e}_k) - \mathbf{f}(\mathbf{x})}{h} = \lim_{h \to 0} \left(\mathbf{v}_k + \frac{\mathbf{o}(h)}{h} \right) = \mathbf{v}_k$$

and so, the above limit exists. This proves the theorem.

Let $\mathbf{f} : U \to \mathbb{R}^q$ where U is an open subset of $\mathbb{R}^p$ and $\mathbf{f}$ is differentiable. It was just shown

$$\mathbf{f}(\mathbf{x}+\mathbf{v}) = \mathbf{f}(\mathbf{x}) + \sum_{j=1}^{p} \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_j} v_j + \mathbf{o}(\mathbf{v}).$$

Taking the i^{th} coordinate of the above equation yields

$$f_i(\mathbf{x}+\mathbf{v}) = f_i(\mathbf{x}) + \sum_{j=1}^{p} \frac{\partial f_i(\mathbf{x})}{\partial x_j} v_j + o(\mathbf{v})$$

and it follows that the term with a sum is nothing more than the i^{th} component of $J(\mathbf{x})\mathbf{v}$ where $J(\mathbf{x})$ is the $q \times p$ matrix,

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_p} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_q}{\partial x_1} & \frac{\partial f_q}{\partial x_2} & \cdots & \frac{\partial f_q}{\partial x_p} \end{pmatrix}.$$

This gives the form of the matrix which defines the linear transformation, $D\mathbf{f}(\mathbf{x})$. Thus

$$\mathbf{f}(\mathbf{x}+\mathbf{v}) = \mathbf{f}(\mathbf{x}) + J(\mathbf{x})\mathbf{v} + \mathbf{o}(\mathbf{v}) \tag{28.6}$$

and to reiterate, the linear transformation which results by multiplication by this $q \times p$ matrix is known as the derivative.

Sometimes we write x, y, z instead of x_1, x_2, and x_3. This is to save on notation and is easier to write and to look at although it lacks generality. When this is done it is understood that $x = x_1, y = x_2$, and $z = x_3$. Thus the derivative is the linear transformation determined by

$$\begin{pmatrix} f_{1x} & f_{1y} & f_{1z} \\ f_{2x} & f_{2y} & f_{2z} \\ f_{3x} & f_{3y} & f_{3z} \end{pmatrix}.$$

Example 28.2.3 *Let A be a constant $m \times n$ matrix and consider $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$. Find $D\mathbf{f}(\mathbf{x})$ if it exists.*

$$\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) = A(\mathbf{x}+\mathbf{h}) - A(\mathbf{x}) = A\mathbf{h} = A\mathbf{h} + \mathbf{o}(\mathbf{h}).$$

In fact in this case, $\mathbf{o}(\mathbf{h}) = \mathbf{0}$. Therefore, $D\mathbf{f}(\mathbf{x}) = A$. Note that this looks the same as the case in one variable, $f(x) = ax$.

28.3 C^1 Functions

Given a function of many variables, how can you tell if it is differentiable? Sometimes you have to go directly to the definition and verify it is differrentiable from the definition. For example, you may have seen the following important example in one variable calculus.

Example 28.3.1 *Let* $f(x) = \begin{cases} x^2 \sin\left(\frac{1}{x}\right) & \text{if } x \neq 0 \\ 0 \text{ if } x = 0 \end{cases}$. *Find* $Df(0)$.

$f(h) - f(0) = 0h + h^2 \sin\left(\frac{1}{h}\right) = o(h)$ and so $Df(0) = 0$. If you find the derivative for $x \neq 0$, it is totally useless information if what you want is $Df(0)$. This is because the derivative, turns out to be discontinuous. Try it. Find the derivative for $x \neq 0$ and try to obtain $Df(0)$ from it. You see, in this example you had to revert to the definition to find the derivative.

It isn't really too hard to use the definition even for more ordinary examples.

Example 28.3.2 *Let* $\mathbf{f}(x,y) = \begin{pmatrix} x^2y + y^2 \\ y^3x \end{pmatrix}$. *Find* $D\mathbf{f}(1,2)$.

First of all note that the thing you are after is a 2×2 matrix.

$$\mathbf{f}(1,2) = \begin{pmatrix} 6 \\ 8 \end{pmatrix}.$$

Then

$$\begin{aligned} & \mathbf{f}(1+h_1, 2+h_2) - \mathbf{f}(1,2) \\ = & \begin{pmatrix} (1+h_1)^2 (2+h_2) + (2+h_2)^2 \\ (2+h_2)^3 (1+h_1) \end{pmatrix} - \begin{pmatrix} 6 \\ 8 \end{pmatrix} \\ = & \begin{pmatrix} 5h_2 + 4h_1 + 2h_1h_2 + 2h_1^2 + h_1^2h_2 + h_2^2 \\ 8h_1 + 12h_2 + 12h_1h_2 + 6h_2^2 + 6h_2^2h_1 + h_2^3 + h_2^3h_1 \end{pmatrix} \\ = & \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \begin{pmatrix} 2h_1h_2 + 2h_1^2 + h_1^2h_2 + h_2^2 \\ 12h_1h_2 + 6h_2^2 + 6h_2^2h_1 + h_2^3 + h_2^3h_1 \end{pmatrix} \\ = & \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \mathbf{o}(\mathbf{h}). \end{aligned}$$

Therefore, the standard matrix of the derivative is $\begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix}$.

Most of the time, there is an easier way to conclude a derivative exists and to find it. It involves the notion of a C^1 function.

Definition 28.3.3 *When $\mathbf{f} : U \to \mathbb{R}^p$ for U an open subset of $\mathbb{R}^n$ and the vector valued functions, $\frac{\partial \mathbf{f}}{\partial x_i}$ are all continuous, (equivalently each $\frac{\partial f_i}{\partial x_j}$ is continuous), the function is said to be $C^1(U)$. If all the partial derivatives up to order k exist and are continuous, then the function is said to be C^k.*

It turns out that for a C^1 function, all you have to do is write the matrix described in Theorem 28.2.2 and this will be the derivative. There is no question of existence for the derivative for such functions. This is the importance of the next few theorems.

Theorem 28.3.4 *Let U be an open subset of $\mathbb{R}^2$ and suppose $f : U \to \mathbb{R}$ has the property that the partial derivatives f_x and f_y exist for $(x,y) \in U$ and are continuous at the point (x_0, y_0). Then*

$$f((x_0,y_0) + (v_1,v_2)) = f(x_0,y_0) + \frac{\partial f}{\partial x}(x_0,y_0)\, v_1 + \frac{\partial f}{\partial x}(x_0,y_0)\, v_2 + o(\mathbf{v}).$$

That is, f is differentiable.

Proof:

$$f((x_0,y_0) + (v_1,v_2)) - \left(f(x_0,y_0) + \frac{\partial f}{\partial x}(x_0,y_0)\, v_1 + \frac{\partial f}{\partial y}(x_0,y_0)\, v_2 \right) \tag{28.7}$$

$$= (f(x_0+v_1, y_0+v_2) - f(x_0,y_0)) - \left(\frac{\partial f}{\partial x}(x_0,y_0)\, v_1 + \frac{\partial f}{\partial y}(x_0,y_0)\, v_2 \right)$$

$$= \left(\overbrace{f(x_0+v_1, y_0+v_2) - f(x_0, y_0+v_2)}^{\text{changes only in first component}} + \overbrace{f(x_0, y_0+v_2) - f(x_0,y_0)}^{\text{changes only in second component}} \right)$$

$$-\left(\frac{\partial f}{\partial x}(x_0,y_0)\,v_1+\frac{\partial f}{\partial y}(x_0,y_0)\,v_2\right)$$

By the mean value theorem, there exist numbers s and t in $[0,1]$ such that this equals

$$=\left(\frac{\partial f}{\partial x}(x_0+tv_1,y_0+v_2)\,v_1+\frac{\partial f}{\partial y}(x_0,y_0+sv_2)\,v_2\right)$$

$$-\left(\frac{\partial f}{\partial x}(x_0,y_0)\,v_1+\frac{\partial f}{\partial y}(x_0,y_0)\,v_2\right)$$

$$=\left(\frac{\partial f}{\partial x}(x_0+tv_1,y_0+v_2)-\frac{\partial f}{\partial x}(x_0,y_0)\right)v_1+\left(\frac{\partial f}{\partial y}(x_0,y_0+sv_2)-\frac{\partial f}{\partial y}(x_0,y_0)\right)v_2$$

Therefore, letting $o(\mathbf{v})$ denote the expression in (28.7), and noticing that $|v_1|$ and $|v_2|$ are both no larger than $|\mathbf{v}|$,

$$|o(\mathbf{v})|\leq\left(\left|\frac{\partial f}{\partial x}(x_0+tv_1,y_0+v_2)-\frac{\partial f}{\partial x}(x_0,y_0)\right|+\left|\frac{\partial f}{\partial y}(x_0,y_0+sv_2)-\frac{\partial f}{\partial y}(x_0,y_0)\right|\right)|\mathbf{v}|.$$

It follows

$$\frac{|o(\mathbf{v})|}{|\mathbf{v}|}\leq\left|\frac{\partial f}{\partial x}(x_0+tv_1,y_0+v_2)-\frac{\partial f}{\partial x}(x_0,y_0)\right|+\left|\frac{\partial f}{\partial y}(x_0,y_0+sv_2)-\frac{\partial f}{\partial y}(x_0,y_0)\right|$$

Therefore, $\lim_{\mathbf{v}\to\mathbf{0}}\frac{|o(\mathbf{v})|}{|\mathbf{v}|}=0$ because of the assumption that f_x and f_y are continuous at the point (x_0,y_0) and this proves the theorem.

Having proved a theorem for scalar valued functions, one for vector valued functions follows immediately.

Theorem 28.3.5 *Let U be an open subset of $\mathbb{R}^p$ for $p\geq 1$ and suppose $\mathbf{f}:U\to\mathbb{R}^q$ has the property that each component function, f_i is differentiable at $\mathbf{x}_0$. Then $\mathbf{f}$ is differentiable at $\mathbf{x}_0$.*

Proof: Let $\mathbf{f}(\mathbf{x})\equiv(f_1(\mathbf{x}),\cdots,f_q(\mathbf{x}))^T$. From the assumption each component function is differentiable, the following holds for each $k=1,\cdots,q$.

$$f_k(\mathbf{x}_0+\mathbf{v})=f_k(\mathbf{x}_0)+\sum_{i=1}^{p}\frac{\partial f_k}{\partial x_i}(\mathbf{x}_0)\,v_i+o_k(\mathbf{v}).$$

Define $\mathbf{o}(\mathbf{v})\equiv(o_1(\mathbf{v}),\cdots,o_q(\mathbf{v}))^T$. Then (28.1) on Page 643 holds for $\mathbf{o}(\mathbf{v})$ because it holds for each of the components of $\mathbf{o}(\mathbf{v})$. The above equation is then equivalent to

$$\mathbf{f}(\mathbf{x}_0+\mathbf{v})=\mathbf{f}(\mathbf{x}_0)+\sum_{i=1}^{p}\frac{\partial\mathbf{f}}{\partial x_i}(\mathbf{x}_0)\,v_i+\mathbf{o}(\mathbf{v})$$

and so $\mathbf{f}$ is differentiable at $\mathbf{x}_0$.

Here is an example to illustrate.

Example 28.3.6 *Let $\mathbf{f}(x,y)=\begin{pmatrix}x^2y+y^2\\ y^3x\end{pmatrix}$. Find $D\mathbf{f}(x,y)$.*

From Theorem 28.3.4 this function is differentiable because all possible partial derivatives are continuous. Thus

$$D\mathbf{f}(x,y) = \begin{pmatrix} 2xy & x^2+2y \\ y^3 & 3y^2x \end{pmatrix}.$$

In particular,

$$D\mathbf{f}(1,2) = \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix}.$$

Not surprisingly, the above theorem has an extension to more variables. First this is illustrated with an example.

Example 28.3.7 *Let* $\mathbf{f}(x_1,x_2,x_3) = \begin{pmatrix} x_1^2x_2+x_2^2 \\ x_2x_1+x_3 \\ \sin(x_1x_2x_3) \end{pmatrix}$. *Find* $D\mathbf{f}(x_1,x_2,x_3)$.

All possible partial derivatives are continuous so the function is differentiable. The matrix for this derivative is therefore the following 3×3 matrix

$$\begin{pmatrix} 2x_1x_2 & x_1^2+2x_2 & 0 \\ x_2 & x_1 & 1 \\ x_2x_3\cos(x_1x_2x_3) & x_1x_3\cos(x_1x_2x_3) & x_1x_2\cos(x_1x_2x_3) \end{pmatrix}$$

The following theorem is the general result.

Theorem 28.3.8 *Let* U *be an open subset of* $\mathbb{R}^p$ *for* $p \geq 1$ *and suppose* $f: U \to \mathbb{R}$ *has the property that the partial derivatives* f_{x_i} *exist for all* $\mathbf{x} \in U$ *and are continuous at the point* $\mathbf{x}_0 \in U$. *Then*

$$f(\mathbf{x}_0+\mathbf{v}) = f(\mathbf{x}_0) + \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(\mathbf{x}_0)\, v_i + o(\mathbf{v}).$$

That is, f *is differentiable at* $\mathbf{x}_0$ *and the derivative of* f *equals the linear transformation obtained by multiplying by the* $1\times p$ *matrix,*

$$\left(\frac{\partial f}{\partial x_1}(\mathbf{x}_0), \cdots, \frac{\partial f}{\partial x_p}(\mathbf{x}_0)\right).$$

Proof: The proof is similar to the case of two variables. Letting $\mathbf{v} = (v_1 \cdots, v_p)^T$, denote by $\theta_i\mathbf{v}$ the vector

$$(0,\cdots,0,v_i,v_{i+1},\cdots,v_p)^T$$

Thus $\theta_0\mathbf{v} = \mathbf{v}$, $\theta_{p-1}(\mathbf{v}) = (0,\cdots,0,v_p)^T$, and $\theta_p\mathbf{v} = \mathbf{0}$. Now

$$f(\mathbf{x}_0+\mathbf{v}) - \left(f(\mathbf{x}_0) + \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(\mathbf{x}_0)\, v_i\right) \tag{28.8}$$

$$= \sum_{i=1}^{p} \left(\overbrace{f(\mathbf{x}_0+\theta_{i-1}\mathbf{v}) - f(\mathbf{x}_0+\theta_i\mathbf{v})}^{\text{changes only in the } i^{th} \text{ position}}\right) - \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(\mathbf{x}_0)\, v_i$$

Now by the mean value theorem there exist numbers $s_i \in (0,1)$ such that the above expression equals

$$= \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(\mathbf{x}_0+\theta_i\mathbf{v}+s_iv_i)\, v_i - \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(\mathbf{x}_0)\, v_i$$

and so letting $o(\mathbf{v})$ equal the expression in (28.8),

$$|o(\mathbf{v})| \leq \sum_{i=1}^{p} \left| \frac{\partial f}{\partial x_i}(\mathbf{x}_0 + \theta_i \mathbf{v} + s_i v_i) - \frac{\partial f}{\partial x_i}(\mathbf{x}_0) \right| |v_i|$$

$$\leq \sum_{i=1}^{p} \left| \frac{\partial f}{\partial x_i}(\mathbf{x}_0 + \theta_i \mathbf{v} + s_i v_i) - \frac{\partial f}{\partial x_i}(\mathbf{x}_0) \right| |\mathbf{v}|$$

and so

$$\lim_{\mathbf{v}\to\mathbf{0}} \frac{|o(\mathbf{v})|}{|\mathbf{v}|} \leq \lim_{\mathbf{v}\to\mathbf{0}} \sum_{i=1}^{p} \left| \frac{\partial f}{\partial x_i}(\mathbf{x}_0 + \theta_i \mathbf{v} + s_i v_i) - \frac{\partial f}{\partial x_i}(\mathbf{x}_0) \right| = 0$$

because of continuity of the f_{x_i} at $\mathbf{x}_0$. This proves the theorem.

Letting $\mathbf{x} - \mathbf{x}_0 = \mathbf{v}$,

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{x}_0) + \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(\mathbf{x}_0)(x_i - x_{0i}) + o(\mathbf{v}) \\ &= f(\mathbf{x}_0) + \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(\mathbf{x}_0) v_i + o(\mathbf{v}). \end{aligned}$$

Example 28.3.9 *Suppose* $f(x,y,z) = xy + z^2$. *Find* $Df(1,2,3)$.

Taking the partial derivatives of f, $f_x = y, f_y = x, f_z = 2z$. These are all continuous. Therefore, the function has a derivative and $f_x(1,2,3) = 1, f_y(1,2,3) = 2$, and $f_z(1,2,3) = 6$. Therefore, $Df(1,2,3)$ is given by

$$Df(1,2,3) = (1,2,6).$$

Also, for (x,y,z) close to $(1,2,3)$,

$$\begin{aligned} f(x,y,z) &\approx f(1,2,3) + 1(x-1) + 2(y-2) + 6(z-3) \\ &= 11 + 1(x-1) + 2(y-2) + 6(z-3) = -12 + x + 2y + 6z \end{aligned}$$

In the case where $\mathbf{f}$ has values in $\mathbb{R}^q$ rather than $\mathbb{R}$, is there a similar theorem about differentiability of a C^1 function?

Theorem 28.3.10 *Let* U *be an open subset of* $\mathbb{R}^p$ *for* $p \geq 1$ *and suppose* $\mathbf{f} : U \to \mathbb{R}^q$ *has the property that the partial derivatives* $\mathbf{f}_{x_i}$ *exist for all* $\mathbf{x} \in U$ *and are continuous at the point* $\mathbf{x}_0 \in U$, *then*

$$\mathbf{f}(\mathbf{x}_0 + \mathbf{v}) = \mathbf{f}(\mathbf{x}_0) + \sum_{i=1}^{p} \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}_0) v_i + \mathbf{o}(\mathbf{v}) \tag{28.9}$$

and so $\mathbf{f}$ *is differentiable at* $\mathbf{x}_0$.

Proof: This follows from Theorem 28.3.5.

When a function is differentiable at $\mathbf{x}_0$ it follows the function must be continuous there. This is the content of the following important lemma.

Lemma 28.3.11 *Let* $\mathbf{f} : U \to \mathbb{R}^q$ *where* U *is an open subset of* $\mathbb{R}^p$. *If* $\mathbf{f}$ *is differentiable, then* $\mathbf{f}$ *is continuous at* $\mathbf{x}_0$. *Furthermore, if* $C \geq \max\left\{ \left| \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}_0) \right|, i = 1, \cdots, p \right\}$, *then whenever* $|\mathbf{x} - \mathbf{x}_0|$ *is small enough,*

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)| \leq (Cp + 1)|\mathbf{x} - \mathbf{x}_0| \tag{28.10}$$

Proof: Suppose $\mathbf{f}$ is differentiable. Since $\mathbf{o}(\mathbf{v})$ satisfies (28.1), there exists $\delta_1 > 0$ such that if $|\mathbf{x}-\mathbf{x}_0| < \delta_1$, then $|\mathbf{o}(\mathbf{x}-\mathbf{x}_0)| < |\mathbf{x}-\mathbf{x}_0|$. But also, by the triangle inequality, Corollary 17.1.5 on Page 444,

$$\left|\sum_{i=1}^{p} \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}_0)(x_i - x_{0i})\right| \leq C\sum_{i=1}^{p}|x_i - x_{0i}| \leq Cp\,|\mathbf{x}-\mathbf{x}_0|$$

Therefore, if $|\mathbf{x}-\mathbf{x}_0| < \delta_1$,

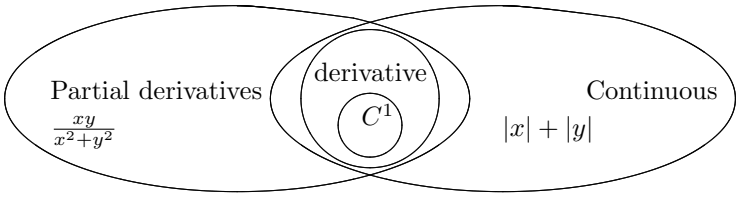
$$\begin{aligned} |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)| &\leq \left|\sum_{i=1}^{p} \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}_0)(x_i - x_{0i})\right| + |\mathbf{x}-\mathbf{x}_0| \\ &< (Cp+1)\,|\mathbf{x}-\mathbf{x}_0| \end{aligned}$$

which verifies (28.10). Now letting $\varepsilon > 0$ be given, let $\delta = \min\left(\delta_1, \frac{\varepsilon}{Cp+1}\right)$. Then for $|\mathbf{x}-\mathbf{x}_0| < \delta$,

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)| < (Cp+1)\,|\mathbf{x}-\mathbf{x}_0| < (Cp+1)\frac{\varepsilon}{Cp+1} = \varepsilon$$

showing $\mathbf{f}$ is continuous at $\mathbf{x}_0$.

There have been quite a few terms defined. First there was the concept of continuity. Next the concept of partial or directional derivative. Next there was the concept of differentiability and the derivative being a linear transformation determined by a certain matrix. Finally, it was shown that if a function is C^1, then it has a derivative. To give a rough idea of the relationships of these topics, here is a picture.



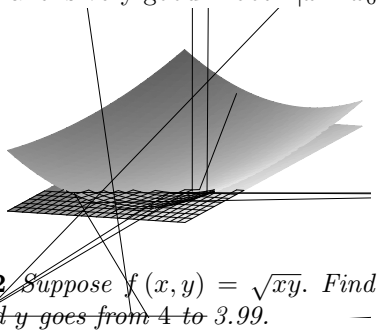
You might ask whether there are examples of functions which are differentiable but not C^1. Of course there are. In fact, Example 28.3.1 is just such an example as explained earlier. Then you should verify that $f'(x)$ exists for all $x \in \mathbb{R}$ but f' fails to be continuous at $x = 0$. Thus the function is differentiable at every point of $\mathbb{R}$ but fails to be C^1 at every point of $\mathbb{R}$.

28.3.1 Approximation With A Tangent Plane

In the case where f is a scalar valued function of two variables, the geometric significance of the derivative can be exhibited in the following picture. Writing $\mathbf{v} \equiv (x-x_0, y-y_0)$, the notion of differentiability at (x_0, y_0) reduces to

$$f(x,y) = f(x_0,y_0) + \frac{\partial f}{\partial x}(x_0,y_0)(x-x_0) + \frac{\partial f}{\partial x}(x_0,y_0)(y-y_0) + o(\mathbf{v})$$

The right side of the above, $f(x_0,y_0)+\frac{\partial f}{\partial x}(x_0,y_0)(x-x_0)+\frac{\partial f}{\partial x}(x_0,y_0)(y-y_0)=z$ is the equation of a plane approximating the graph of $z=f(x,y)$ for (x,y) near (x_0,y_0). Saying that the function is differentiable at (x_0,y_0) amounts to saying that the approximation delivered by this plane is very good if both $|x-x_0|$ and $|y-y_0|$ are small.



Example 28.3.12 *Suppose $f(x,y)=\sqrt{xy}$. Find the approximate change in f if x goes from 1 to 1.01 and y goes from 4 to 3.99.*

We can do this by noting that

$$\begin{aligned} f(1.01,3.99)-f(1,4) &\approx f_x(1,2)(.01)+f_y(1,2)(-.01) \\ &= 1(.01)+\frac{1}{4}(-.01)=7.5\times 10^{-3}. \end{aligned}$$

Of course the exact value would be

$$\sqrt{(1.01)(3.99)}-\sqrt{4}=7.461\,083\,1\times 10^{-3}.$$

28.4 The Chain Rule

As in the case of a function of one variable, it is important to consider the derivative of a composition of two functions. The proof of the chain rule depends on the following fundamental lemma.

Lemma 28.4.1 *Let $\mathbf{g}:U\to\mathbb{R}^p$ where U is an open set in $\mathbb{R}^n$ and suppose $\mathbf{g}$ has a derivative at $\mathbf{x}\in U$. Then $\mathbf{o}(\mathbf{g}(\mathbf{x}+\mathbf{v})-\mathbf{g}(\mathbf{x}))=\mathbf{o}(\mathbf{v})$.*

Proof: It is necessary to show

$$\lim_{\mathbf{v}\to\mathbf{0}}\frac{|\mathbf{o}(\mathbf{g}(\mathbf{x}+\mathbf{v})-\mathbf{g}(\mathbf{x}))|}{|\mathbf{v}|}=0. \tag{28.11}$$

From Lemma 28.3.11, there exists $\delta>0$ such that if $|\mathbf{v}|<\delta$, then

$$|\mathbf{g}(\mathbf{x}+\mathbf{v})-\mathbf{g}(\mathbf{x})|\le(Cn+1)|\mathbf{v}|. \tag{28.12}$$

Now let $\varepsilon>0$ be given. There exists $\eta>0$ such that if $|\mathbf{g}(\mathbf{x}+\mathbf{v})-\mathbf{g}(\mathbf{x})|<\eta$, then

$$|\mathbf{o}(\mathbf{g}(\mathbf{x}+\mathbf{v})-\mathbf{g}(\mathbf{x}))|<\left(\frac{\varepsilon}{Cn+1}\right)|\mathbf{g}(\mathbf{x}+\mathbf{v})-\mathbf{g}(\mathbf{x})| \tag{28.13}$$

Let $|\mathbf{v}|<\min\left(\delta,\frac{\eta}{Cn+1}\right)$. For such $\mathbf{v}$, $|\mathbf{g}(\mathbf{x}+\mathbf{v})-\mathbf{g}(\mathbf{x})|\le\eta$, which implies

$$\begin{aligned} |\mathbf{o}(\mathbf{g}(\mathbf{x}+\mathbf{v})-\mathbf{g}(\mathbf{x}))| &< \left(\frac{\varepsilon}{Cn+1}\right)|\mathbf{g}(\mathbf{x}+\mathbf{v})-\mathbf{g}(\mathbf{x})| \\ &< \left(\frac{\varepsilon}{Cn+1}\right)(Cn+1)|\mathbf{v}| \end{aligned}$$

and so

$$\frac{|\mathbf{o}(\mathbf{g}(\mathbf{x}+\mathbf{v})-\mathbf{g}(\mathbf{x}))|}{|\mathbf{v}|}<\varepsilon$$

which establishes (28.11). This proves the lemma.

Recall the notation $\mathbf{f}\circ\mathbf{g}(\mathbf{x})\equiv\mathbf{f}(\mathbf{g}(\mathbf{x}))$. Thus $\mathbf{f}\circ\mathbf{g}$ is the name of a function and this function is defined by what was just written. The following theorem is known as the chain rule.

Theorem 28.4.2 *(Chain rule) Let U be an open set in $\mathbb{R}^n$, let V be an open set in $\mathbb{R}^p$, let $\mathbf{g}:U\to\mathbb{R}^p$ be such that $\mathbf{g}(U)\subseteq V$, and let $\mathbf{f}:V\to\mathbb{R}^q$. Suppose $D\mathbf{g}(\mathbf{x})$ exists for some $\mathbf{x}\in U$ and that $D\mathbf{f}(\mathbf{g}(\mathbf{x}))$ exists. Then $D(\mathbf{f}\circ\mathbf{g})(\mathbf{x})$ exists and furthermore,*

$$D(\mathbf{f}\circ\mathbf{g})(\mathbf{x})=D\mathbf{f}(\mathbf{g}(\mathbf{x}))\,D\mathbf{g}(\mathbf{x}). \tag{28.14}$$

In particular,

$$\frac{\partial(\mathbf{f}\circ\mathbf{g})(\mathbf{x})}{\partial x_j}=\sum_{i=1}^{p}\frac{\partial\mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial y_i}\frac{\partial g_i(\mathbf{x})}{\partial x_j}. \tag{28.15}$$

Proof: From the assumption that $D\mathbf{f}(\mathbf{g}(\mathbf{x}))$ exists,

$$\mathbf{f}(\mathbf{g}(\mathbf{x}+\mathbf{v}))=\mathbf{f}(\mathbf{g}(\mathbf{x}))+\sum_{i=1}^{p}\frac{\partial\mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial y_i}(g_i(\mathbf{x}+\mathbf{v})-g_i(\mathbf{x}))+\mathbf{o}(\mathbf{g}(\mathbf{x}+\mathbf{v})-\mathbf{g}(\mathbf{x}))$$

which by Lemma 28.4.1 equals

$$(\mathbf{f}\circ\mathbf{g})(\mathbf{x}+\mathbf{v})=\mathbf{f}(\mathbf{g}(\mathbf{x}+\mathbf{v}))=\mathbf{f}(\mathbf{g}(\mathbf{x}))+\sum_{i=1}^{p}\frac{\partial\mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial y_i}(g_i(\mathbf{x}+\mathbf{v})-g_i(\mathbf{x}))+\mathbf{o}(\mathbf{v}).$$

Now since $D\mathbf{g}(\mathbf{x})$ exists, the above becomes

$$\begin{aligned}(\mathbf{f}\circ\mathbf{g})(\mathbf{x}+\mathbf{v}) &= \mathbf{f}(\mathbf{g}(\mathbf{x}))+\sum_{i=1}^{p}\frac{\partial\mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial y_i}\left(\sum_{j=1}^{n}\frac{\partial g_i(\mathbf{x})}{\partial x_j}v_j+\mathbf{o}(\mathbf{v})\right)+\mathbf{o}(\mathbf{v})\\ &= \mathbf{f}(\mathbf{g}(\mathbf{x}))+\sum_{i=1}^{p}\frac{\partial\mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial y_i}\left(\sum_{j=1}^{n}\frac{\partial g_i(\mathbf{x})}{\partial x_j}v_j\right)+\sum_{i=1}^{p}\frac{\partial\mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial y_i}\mathbf{o}(\mathbf{v})+\mathbf{o}(\mathbf{v})\\ &= (\mathbf{f}\circ\mathbf{g})(\mathbf{x})+\sum_{j=1}^{n}\left(\sum_{i=1}^{p}\frac{\partial\mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial y_i}\frac{\partial g_i(\mathbf{x})}{\partial x_j}\right)v_j+\mathbf{o}(\mathbf{v})\end{aligned}$$

because $\sum_{i=1}^{p}\frac{\partial\mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial y_i}\mathbf{o}(\mathbf{v})+\mathbf{o}(\mathbf{v})=\mathbf{o}(\mathbf{v})$. This establishes (28.15) because of Theorem 28.2.2 on Page 645. Thus

$$\begin{aligned}(D(\mathbf{f}\circ\mathbf{g})(\mathbf{x}))_{kj} &= \sum_{i=1}^{p}\frac{\partial f_k(\mathbf{g}(\mathbf{x}))}{\partial y_i}\frac{\partial g_i(\mathbf{x})}{\partial x_j}\\ &= \sum_{i=1}^{p}D\mathbf{f}(\mathbf{g}(\mathbf{x}))_{ki}(D\mathbf{g}(\mathbf{x}))_{ij}.\end{aligned}$$

Then (28.14) follows from the definition of matrix multiplication.

There is an easy way to remember this in terms of the repeated index summation convention presented earlier. Let $\mathbf{y} = \mathbf{g}(\mathbf{x})$ and $\mathbf{z} = \mathbf{f}(\mathbf{y})$. Then the above says

$$\frac{\partial \mathbf{z}}{\partial y_i}\frac{\partial y_i}{\partial x_k} = \frac{\partial \mathbf{z}}{\partial x_k}.$$

Remember there is a sum on the repeated index.

Example 28.4.3 *Let $f(u,v) = \sin(uv)$ and let $u(x,y,t) = t\sin x + \cos y$ and $v(x,y,t,s) = s\tan x + y^2 + ts$. Letting $z = f(u,v)$ where u, v are as just described, find $\frac{\partial z}{\partial t}$ and $\frac{\partial z}{\partial x}$.*

From the above,

$$\frac{\partial z}{\partial t} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial t} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial t} = v\cos(uv)\sin(x) + us\cos(uv).$$

Also,

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial x} = v\cos(uv)\,t\cos(x) + us\sec^2(x)\cos(uv).$$

Clearly you can continue in this way taking partial derivatives with respect to any of the other variables.

Example 28.4.4 *Let $\mathbf{f}(u_1,u_2) = \begin{pmatrix} u_1^2 + u_2 \\ \sin(u_2) + u_1 \end{pmatrix}$ and $\mathbf{g}(x_1,x_2,x_3) = \begin{pmatrix} u_1(x_1,x_2,x_3) \\ u_2(x_1,x_2,x_3) \end{pmatrix} = \begin{pmatrix} x_1x_2 + x_3 \\ x_2^2 + x_1 \end{pmatrix}$. Find $D(\mathbf{f}\circ\mathbf{g})(x_1,x_2,x_3)$.*

To do this,

$$D\mathbf{f}(u_1,u_2) = \begin{pmatrix} 2u_1 & 1 \\ 1 & \cos u_2 \end{pmatrix}, D\mathbf{g}(x_1,x_2,x_3) = \begin{pmatrix} x_2 & x_1 & 1 \\ 1 & 2x_2 & 0 \end{pmatrix}.$$

Then

$$D\mathbf{f}(\mathbf{g}(x_1,x_2,x_3)) = \begin{pmatrix} 2(x_1x_2+x_3) & 1 \\ 1 & \cos\left(x_2^2+x_1\right) \end{pmatrix}$$

and so by the chain rule,

$$\begin{aligned} D(\mathbf{f}\circ\mathbf{g})(x_1,x_2,x_3) &= \overbrace{\begin{pmatrix} 2(x_1x_2+x_3) & 1 \\ 1 & \cos\left(x_2^2+x_1\right) \end{pmatrix}}^{D\mathbf{f}(\mathbf{g}(\mathbf{x}))}\overbrace{\begin{pmatrix} x_2 & x_1 & 1 \\ 1 & 2x_2 & 0 \end{pmatrix}}^{D\mathbf{g}(\mathbf{x})} \\ &= \begin{pmatrix} (2x_1x_2+2x_3)x_2+1 & (2x_1x_2+2x_3)x_1+2x_2 & 2x_1x_2+2x_3 \\ x_2+\cos\left(x_2^2+x_1\right) & x_1+2x_2\left(\cos\left(x_2^2+x_1\right)\right) & 1 \end{pmatrix} \end{aligned}$$

Therefore, in particular,

$$\frac{\partial f_1\circ\mathbf{g}}{\partial x_1}(x_1,x_2,x_3) = (2x_1x_2+2x_3)x_2+1,$$

$$\frac{\partial f_2\circ\mathbf{g}}{\partial x_3}(x_1,x_2,x_3) = 1, \frac{\partial f_2\circ\mathbf{g}}{\partial x_2}(x_1,x_2,x_3) = x_1 + 2x_2\left(\cos\left(x_2^2+x_1\right)\right).$$

etc.

Example 28.4.5 *Let $\mathbf{f}: U \to V$ where U and V are open sets in $\mathbb{R}^n$ and $\mathbf{f}$ is one to one and onto. Suppose also that $\mathbf{f}$ and $\mathbf{f}^{-1}$ are both differentiable. How are $D\mathbf{f}^{-1}$ and $D\mathbf{f}$ related?*

This can be done as follows. From the assumptions, $\mathbf{x} = \mathbf{f}^{-1}(\mathbf{f}(\mathbf{x}))$. Let $I\mathbf{x} = \mathbf{x}$. Then by Example 28.2.3 on Page 646 $DI = I$. By the chain rule,

$$I = DI = D\mathbf{f}^{-1}(\mathbf{f}(\mathbf{x}))(D\mathbf{f}(\mathbf{x})).$$

Therefore,

$$D\mathbf{f}(\mathbf{x})^{-1} = D\mathbf{f}^{-1}(\mathbf{f}(\mathbf{x})).$$

This is equivalent to

$$D\mathbf{f}\left(\mathbf{f}^{-1}(\mathbf{y})\right)^{-1} = D\mathbf{f}^{-1}(\mathbf{y})$$

or

$$D\mathbf{f}(\mathbf{x})^{-1} = D\mathbf{f}^{-1}(\mathbf{y}), \mathbf{y} = \mathbf{f}(\mathbf{x}).$$

This is just like a similar situation for functions of one variable. Remember $\left(f^{-1}\right)'(f(x)) = 1/f'(x)$. In terms of the repeated index summation convention, suppose $\mathbf{y} = \mathbf{f}(\mathbf{x})$ so that $\mathbf{x} = \mathbf{f}^{-1}(\mathbf{y})$. Then the above can be written as

$$\delta_{ij} = \frac{\partial x_i}{\partial y_k}(\mathbf{f}(\mathbf{x}))\frac{\partial y_k}{\partial x_j}(\mathbf{x}).$$

Example 28.4.6 *Recall spherical coordinates are given by*

$$x = \rho\sin\phi\cos\theta,\ y = \rho\sin\phi\sin\theta,\ z = \rho\cos\phi.$$

If an object moves in three dimensions, describe its acceleration in terms of spherical coordinates and the vectors,

$$\mathbf{e}_\rho = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)^T,$$

$$\mathbf{e}_\theta = (-\rho\sin\phi\sin\theta, \rho\sin\phi\cos\theta, 0)^T,$$

and

$$\mathbf{e}_\phi = (\rho\cos\phi\cos\theta, \rho\cos\phi\sin\theta, -\rho\sin\phi)^T.$$

Why these vectors? Note how they were obtained. Let

$$\mathbf{r}(\rho,\theta,\phi) = (\rho\sin\phi\cos\theta,\ \rho\sin\phi\sin\theta,\ \rho\cos\phi)^T$$

and fix ϕ and θ, letting only ρ change, this gives a curve in the direction of increasing ρ. Thus it is a vector which points away from the origin. Letting only ϕ change and fixing θ and ρ, this gives a vector which is tangent to the sphere of radius ρ and points South. Similarly, letting θ change and fixing the other two gives a vector which points East and is tangent to the sphere of radius ρ. It is thought by most people that we live on a large sphere. The model of a flat earth is not believed by anyone except perhaps for beginning physics students. Given we live on a sphere, what directions would be most meaningful? Wouldn't it be the directions of the vectors just described?

Let $\mathbf{r}(t)$ denote the position vector of the object from the origin. Thus

$$\mathbf{r}(t) = \rho(t)\mathbf{e}_\rho(t) = \left((x(t), y(t), z(t))^T\right)$$

Now this implies the velocity is

$$\mathbf{r}'(t) = \rho'(t)\mathbf{e}_\rho(t) + \rho(t)(\mathbf{e}_\rho(t))'. \tag{28.16}$$

You see, $\mathbf{e}_\rho = \mathbf{e}_\rho(\rho,\theta,\phi)$ where each of these variables is a function of t.

$$\frac{\partial \mathbf{e}_\rho}{\partial \phi} = (\cos\phi\cos\theta, \cos\phi\sin\theta, -\sin\phi)^T = \frac{1}{\rho}\mathbf{e}_\phi,$$

$$\frac{\partial \mathbf{e}_\rho}{\partial \theta} = (-\sin\phi\sin\theta, \sin\phi\cos\theta, 0)^T = \frac{1}{\rho}\mathbf{e}_\theta,$$

and

$$\frac{\partial \mathbf{e}_\rho}{\partial \rho} = 0.$$

Therefore, by the chain rule,

$$\begin{aligned}\frac{d\mathbf{e}_\rho}{dt} &= \frac{\partial \mathbf{e}_\rho}{\partial \phi}\frac{d\phi}{dt} + \frac{\partial \mathbf{e}_\rho}{\partial \theta}\frac{d\theta}{dt} \\ &= \frac{1}{\rho}\frac{d\phi}{dt}\mathbf{e}_\phi + \frac{1}{\rho}\frac{d\theta}{dt}\mathbf{e}_\theta.\end{aligned}$$

By (28.16),

$$\mathbf{r}' = \rho'\mathbf{e}_\rho + \frac{d\phi}{dt}\mathbf{e}_\phi + \frac{d\theta}{dt}\mathbf{e}_\theta. \tag{28.17}$$

Now things get interesting. This must be differentiated with respect to t. To do so,

$$\frac{\partial \mathbf{e}_\theta}{\partial \theta} = (-\rho\sin\phi\cos\theta, -\rho\sin\phi\sin\theta, 0)^T = ?$$

where it is desired to find a, b, c such that $? = a\mathbf{e}_\theta + b\mathbf{e}_\phi + c\mathbf{e}_\rho$. Thus

$$\begin{pmatrix} -\rho\sin\phi\sin\theta & \rho\cos\phi\cos\theta & \sin\phi\cos\theta \\ \rho\sin\phi\cos\theta & \rho\cos\phi\sin\theta & \sin\phi\sin\theta \\ 0 & -\rho\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} -\rho\sin\phi\cos\theta \\ -\rho\sin\phi\sin\theta \\ 0 \end{pmatrix}$$

Using Cramer's rule, the solution is $a = 0, b = -\cos\phi\sin\phi$, and $c = -\rho\sin^2\phi$. Thus

$$\begin{aligned}\frac{\partial \mathbf{e}_\theta}{\partial \theta} &= (-\rho\sin\phi\cos\theta, -\rho\sin\phi\sin\theta, 0)^T \\ &= (-\cos\phi\sin\phi)\,\mathbf{e}_\phi + \left(-\rho\sin^2\phi\right)\mathbf{e}_\rho.\end{aligned}$$

Also,

$$\frac{\partial \mathbf{e}_\theta}{\partial \phi} = (-\rho\cos\phi\sin\theta, \rho\cos\phi\cos\theta, 0)^T = (\cot\phi)\,\mathbf{e}_\theta$$

and

$$\frac{\partial \mathbf{e}_\theta}{\partial \rho} = (-\sin\phi\sin\theta, \sin\phi\cos\theta, 0)^T = \frac{1}{\rho}\mathbf{e}_\theta.$$

Now in (28.17) it is also necessary to consider $\mathbf{e}_\phi$.

$$\frac{\partial \mathbf{e}_\phi}{\partial \phi} = (-\rho\sin\phi\cos\theta, -\rho\sin\phi\sin\theta, -\rho\cos\phi)^T = -\rho\mathbf{e}_\rho$$

$$\begin{aligned}\frac{\partial \mathbf{e}_\phi}{\partial \theta} &= (-\rho\cos\phi\sin\theta, \rho\cos\phi\cos\theta, 0)^T \\ &= (\cot\phi)\,\mathbf{e}_\theta\end{aligned}$$

and finally,

$$\frac{\partial \mathbf{e}_\phi}{\partial \rho} = (\cos\phi\cos\theta, \cos\phi\sin\theta, -\sin\phi)^T = \frac{1}{\rho}\mathbf{e}_\phi.$$

With these formulas for various partial derivatives, the chain rule is used to obtain $\mathbf{r}''$ which will yield a formula for the acceleration in terms of the spherical coordinates and these special vectors. By the chain rule,

$$\begin{aligned}\frac{d}{dt}(\mathbf{e}_\rho) &= \frac{\partial \mathbf{e}_\rho}{\partial\theta}\theta' + \frac{\partial \mathbf{e}_\rho}{\partial\phi}\phi' + \frac{\partial \mathbf{e}_\rho}{\partial\rho}\rho' \\ &= \frac{\theta'}{\rho}\mathbf{e}_\theta + \frac{\phi'}{\rho}\mathbf{e}_\phi\end{aligned}$$

$$\begin{aligned}\frac{d}{dt}(\mathbf{e}_\theta) &= \frac{\partial \mathbf{e}_\theta}{\partial\theta}\theta' + \frac{\partial \mathbf{e}_\theta}{\partial\phi}\phi' + \frac{\partial \mathbf{e}_\theta}{\partial\rho}\rho' \\ &= \theta'\left((-\cos\phi\sin\phi)\,\mathbf{e}_\phi + \left(-\rho\sin^2\phi\right)\mathbf{e}_\rho\right) + \phi'(\cot\phi)\,\mathbf{e}_\theta + \frac{\rho'}{\rho}\mathbf{e}_\theta\end{aligned}$$

$$\begin{aligned}\frac{d}{dt}(\mathbf{e}_\phi) &= \frac{\partial \mathbf{e}_\phi}{\partial\theta}\theta' + \frac{\partial \mathbf{e}_\phi}{\partial\phi}\phi' + \frac{\partial \mathbf{e}_\phi}{\partial\rho}\rho' \\ &= \left(\theta'\cot\phi\right)\mathbf{e}_\theta + \phi'(-\rho\mathbf{e}_\rho) + \left(\frac{\rho'}{\rho}\mathbf{e}_\phi\right)\end{aligned}$$

By (28.17),

$$\mathbf{r}'' = \rho''\mathbf{e}_\rho + \phi''\mathbf{e}_\phi + \theta''\mathbf{e}_\theta + \rho'(\mathbf{e}_\rho)' + \phi'(\mathbf{e}_\phi)' + \theta'(\mathbf{e}_\theta)'$$

and from the above, this equals

$$\begin{aligned}&\rho''\mathbf{e}_\rho + \phi''\mathbf{e}_\phi + \theta''\mathbf{e}_\theta + \rho'\left(\frac{\theta'}{\rho}\mathbf{e}_\theta + \frac{\phi'}{\rho}\mathbf{e}_\phi\right) + \\ &\phi'\left(\left(\theta'\cot\phi\right)\mathbf{e}_\theta + \phi'(-\rho\mathbf{e}_\rho) + \left(\frac{\rho'}{\rho}\mathbf{e}_\phi\right)\right) + \\ &\theta'\left(\theta'\left((-\cos\phi\sin\phi)\,\mathbf{e}_\phi + \left(-\rho\sin^2\phi\right)\mathbf{e}_\rho\right) + \phi'(\cot\phi)\,\mathbf{e}_\theta + \frac{\rho'}{\rho}\mathbf{e}_\theta\right)\end{aligned}$$

and now all that remains is to collect the terms. Thus $\mathbf{r}''$ equals

$$\begin{aligned}\mathbf{r}'' &= \left(\rho'' - \rho\left(\phi'\right)^2 - \rho\left(\theta'\right)^2\sin^2(\phi)\right)\mathbf{e}_\rho + \left(\phi'' + \frac{2\rho'\phi'}{\rho} - \left(\theta'\right)^2\cos\phi\sin\phi\right)\mathbf{e}_\phi + \\ &\quad + \left(\theta'' + \frac{2\theta'\rho'}{\rho} + 2\phi'\theta'\cot(\phi)\right)\mathbf{e}_\theta.\end{aligned}$$

and this gives the acceleration in spherical coordinates. Note the prominent role played by the chain rule. All of the above is done in books on mechanics for general curvilinear coordinate systems and in the more general context, special theorems are developed which make things go much faster but these theorems are all exercises in the chain rule.

As an example of how this could be used, consider a rocket. Suppose for simplicity that it experiences a force only in the direction of $\mathbf{e}_\rho$, directly away from the earth. Of course this force produces a corresponding acceleration which can be computed as a function of time. As the fuel is burned, the rocket becomes less massive and so the acceleration will be an increasing function of t. However, this would be a known function, say $a(t)$. Suppose

you wanted to know the latitude and longitude of the rocket as a function of time. (There is no reason to think these will stay the same.) Then all that would be required would be to solve the system of differential equations[1],

$$\begin{aligned}\rho'' - \rho\left(\phi'\right)^2 - \rho\left(\theta'\right)^2 \sin^2(\phi) &= a(t), \\ \phi'' + \frac{2\rho'\phi'}{\rho} - \left(\theta'\right)^2 \cos\phi \sin\phi &= 0, \\ \theta'' + \frac{2\theta'\rho'}{\rho} + 2\phi'\theta' \cot(\phi) &= 0\end{aligned}$$

along with initial conditions, $\rho(0) = \rho_0$ (the distance from the launch site to the center of the earth.), $\rho'(0) = \rho_1$(the initial vertical component of velocity of the rocket, probably 0.) and then initial conditions for $\phi, \phi', \theta, \theta'$. The initial value problems could then be solved numerically and you would know the distance from the center of the earth as a function of t along with θ and ϕ. Thus you could predict where the booster shells would fall to earth so you would know where to look for them. Of course there are many variations of this. You might want to specify forces in the $\mathbf{e}_\theta$ and $\mathbf{e}_\phi$ direction as well and attempt to control the position of the rocket or rather its payload. The point is that if you are interested in doing all this in terms of ϕ, θ, and ρ, the above shows how to do it systematically and you see it is all an exercise in using the chain rule. More could be said here involving moving coordinate systems and the Coriolis force. You really might want to do everything with respect to a coordinate system which is fixed with respect to the moving earth.

28.5 Lagrangian Mechanics*

A difficult and important problem is to come up with differential equations which model mechanical systems. Lagrange gave a way to do this. It will be presented here as a very interesting and important application of the chain rule. Lagrange developed this technique back in the 1700's. The presentation here follows [12]. Assume N point masses, located at the points $\mathbf{x}_1, \cdots, \mathbf{x}_N$ in $\mathbb{R}^3$ and let the mass of the α^{th} mass be m_α. Then according to Newton's second law,

$$m_\alpha \mathbf{x}_\alpha'' = \mathbf{F}_\alpha(\mathbf{x}_\alpha, t). \tag{28.18}$$

The dependence of $\mathbf{F}_\alpha$ on the two indicated quantities is indicative of the situation where the force may change in time and position. Now define

$$\mathbf{x} \equiv (\mathbf{x}_1, \cdots, \mathbf{x}_N) \in \mathbb{R}^{3N}$$

and assume $\mathbf{x} \in M$ which is defined locally in the form $\mathbf{x} = \mathbf{G}(\mathbf{q},t)$. Here $\mathbf{q} \in \mathbb{R}^m$ where typically $m < 3N$ and $\mathbf{G}(\cdot, t)$ is a smooth one to one mapping from V, an open subset of $\mathbb{R}^m$ onto a set of points near $\mathbf{x}$ which are on M. Also assume t is in an open subset of $\mathbb{R}$. In what follows a dot over a variable will indicate a derivative taken with respect to time. Two dots will indicate the second derivative with respect to time, etc. Then define $\mathbf{G}_\alpha$ by

$$\mathbf{x}_\alpha = \mathbf{G}_\alpha(\mathbf{q}, t).$$

Using the summation convention and the chain rule,

$$\frac{d\mathbf{x}_\alpha}{dt} = \frac{\partial \mathbf{G}_\alpha}{\partial q^j}\frac{dq^j}{dt} + \frac{\partial \mathbf{G}_\alpha}{\partial t}.$$

[1] You won't be able to find the solution to equations like these in terms of simple functions. The existence of such functions is being assumed. The reason they exist often depends on the implicit function theorem, a big theorem in advanced calculus.

Therefore, the kinetic energy is of the form

$$\begin{aligned} T &\equiv \sum_{\alpha=1}^{N} \frac{1}{2} m_\alpha \left(\frac{d\mathbf{x}_\alpha}{dt} \cdot \frac{d\mathbf{x}_\alpha}{dt} \right) \\ &= \sum_{\alpha=1}^{N} \frac{1}{2} m_\alpha \left(\sum_j \frac{\partial \mathbf{G}_\alpha}{\partial q^j} \frac{dq^j}{dt} + \frac{\partial \mathbf{G}_\alpha}{\partial t} \cdot \sum_r \frac{\partial \mathbf{G}_\alpha}{\partial q^r} \frac{dq^r}{dt} + \frac{\partial \mathbf{G}_\alpha}{\partial t} \right) \end{aligned}$$

$$= \sum_{j,r} \frac{1}{2} \left[\sum_\alpha m_\alpha \left(\frac{\partial \mathbf{G}_\alpha}{\partial q^j} \cdot \frac{\partial \mathbf{G}_\alpha}{\partial q^r} \right) \right] \dot{q}^r \dot{q}^j + \sum_\alpha \sum_j m_\alpha \left(\frac{\partial \mathbf{G}_\alpha}{\partial q^j} \cdot \frac{\partial \mathbf{G}_\alpha}{\partial t} \right) \dot{q}^j$$

$$+ \sum_\alpha \frac{1}{2} m_\alpha \left(\frac{\partial \mathbf{G}_\alpha}{\partial t} \cdot \frac{\partial \mathbf{G}_\alpha}{\partial t} \right) \tag{28.19}$$

where in the last equation $\dot{q}^k$ indicates $\frac{dq^k}{dt}$. Therefore,

$$\begin{aligned} \frac{\partial T}{\partial \dot{q}^k} &= \sum_{j=1}^{m} \left[\sum_{\alpha=1}^{N} m_\alpha \left(\frac{\partial \mathbf{G}_\alpha}{\partial q^j} \cdot \frac{\partial \mathbf{G}_\alpha}{\partial q^k} \right) \right] \dot{q}^j + \sum_\alpha m_\alpha \left(\frac{\partial \mathbf{G}_\alpha}{\partial q^k} \cdot \frac{\partial \mathbf{G}_\alpha}{\partial t} \right) \\ &= \sum_{\alpha=1}^{N} \left(m_\alpha \frac{\partial \mathbf{G}_\alpha}{\partial q^k} \cdot \sum_{j=1}^{m} \frac{\partial \mathbf{G}_\alpha}{\partial q^j} \dot{q}^j \right) + \sum_\alpha m_\alpha \left(\frac{\partial \mathbf{G}_\alpha}{\partial q^k} \cdot \frac{\partial \mathbf{G}_\alpha}{\partial t} \right) \\ &= \left(\sum_{\alpha=1}^{N} m_\alpha \frac{\partial \mathbf{G}_\alpha}{\partial q^k} \cdot \left(\mathbf{x}'_\alpha - \frac{\partial \mathbf{G}_\alpha}{\partial t} \right) \right) + \sum_\alpha m_\alpha \left(\frac{\partial \mathbf{G}_\alpha}{\partial q^k} \cdot \frac{\partial \mathbf{G}_\alpha}{\partial t} \right) \\ &= \sum_{\alpha=1}^{N} \frac{\partial \mathbf{G}_\alpha}{\partial q^k} \cdot m_\alpha \mathbf{x}'_\alpha \end{aligned}$$

Now using the chain rule and product rule again, along with Newton's second law,

$$\begin{aligned} \frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}^k} \right) &= \left(\sum_{\alpha=1}^{N} \left[\left(\sum_j \frac{\partial^2 \mathbf{G}_\alpha}{\partial q^k \partial q^j} \dot{q}^j \right) + \frac{\partial^2 \mathbf{G}_\alpha}{\partial t \partial q^k} \right] \cdot m_\alpha \mathbf{x}'_\alpha \right) \\ &\quad + \left(\sum_{\alpha=1}^{N} \frac{\partial \mathbf{G}_\alpha}{\partial q^k} \cdot m_\alpha \mathbf{x}''_\alpha \right) \\ &= \left(\sum_{\alpha=1}^{N} \left[\left(\sum_j \frac{\partial^2 \mathbf{G}_\alpha}{\partial q^k \partial q^j} \dot{q}^j \right) + \frac{\partial^2 \mathbf{G}_\alpha}{\partial t \partial q^k} \right] \cdot m_\alpha \mathbf{x}'_\alpha \right) + \\ &\quad + \left(\sum_{\alpha=1}^{N} \frac{\partial \mathbf{G}_\alpha}{\partial q^k} \cdot \mathbf{F}_\alpha \right) \\ &= \left(\sum_{\alpha=1}^{N} \left[\left(\sum_j \frac{\partial^2 \mathbf{G}_\alpha}{\partial q^k \partial q^j} \dot{q}^j \right) + \frac{\partial^2 \mathbf{G}_\alpha}{\partial t \partial q^k} \right] \cdot \right. \\ &\quad \left. m_\alpha \left(\sum_r \frac{\partial \mathbf{G}_\alpha}{\partial q^r} \dot{q}^r + \frac{\partial \mathbf{G}_\alpha}{\partial t} \right) \right) + \left(\sum_{\alpha=1}^{N} \frac{\partial \mathbf{G}_\alpha}{\partial q^k} \cdot \mathbf{F}_\alpha \right) \end{aligned} \tag{28.20}$$

$$= \sum_{rj}\left[\sum_{\alpha=1}^{N} m_\alpha\left(\frac{\partial^2 \mathbf{G}_\alpha}{\partial q^j \partial q^k}\cdot\frac{\partial \mathbf{G}_\alpha}{\partial q^r}\right)\right]\dot{q}^r\dot{q}^j + \sum_{\alpha=1}^{N}\sum_j \frac{\partial^2 \mathbf{G}_\alpha}{\partial q^k \partial q^j}\dot{q}^j \cdot m_\alpha \frac{\partial \mathbf{G}_\alpha}{\partial t}$$
$$+\left(\sum_\alpha\sum_r \frac{\partial^2 \mathbf{G}_\alpha}{\partial t \partial q^k}\cdot m_\alpha \frac{\partial \mathbf{G}_\alpha}{\partial q^r}\dot{q}^r\right) + \sum_\alpha \frac{\partial^2 \mathbf{G}_\alpha}{\partial t \partial q^k}\cdot m_\alpha \frac{\partial \mathbf{G}_\alpha}{\partial t} + \left(\sum_{\alpha=1}^{N}\frac{\partial \mathbf{G}_\alpha}{\partial q^k}\cdot \mathbf{F}_\alpha\right) \quad (28.21)$$

Next consider $\frac{\partial T}{\partial q^k}$. Recall (28.19),

$$T = \sum_{j,r}\frac{1}{2}\left[\sum_\alpha m_\alpha\left(\frac{\partial \mathbf{G}_\alpha}{\partial q^j}\cdot\frac{\partial \mathbf{G}_\alpha}{\partial q^r}\right)\right]\dot{q}^r\dot{q}^j + \sum_\alpha\sum_j m_\alpha\left(\frac{\partial \mathbf{G}_\alpha}{\partial q^j}\cdot\frac{\partial \mathbf{G}_\alpha}{\partial t}\right)\dot{q}^j$$
$$+\sum_\alpha \frac{1}{2}m_\alpha\left(\frac{\partial \mathbf{G}_\alpha}{\partial t}\cdot\frac{\partial \mathbf{G}_\alpha}{\partial t}\right) \quad (28.22)$$

From this formula,

$$\frac{\partial T}{\partial q^k} = \sum_{rj}\left[\sum_{\alpha=1}^{N} m_\alpha\left(\frac{\partial^2 \mathbf{G}_\alpha}{\partial q^j \partial q^k}\cdot\frac{\partial \mathbf{G}_\alpha}{\partial q^r}\right)\right]\dot{q}^r\dot{q}^j +$$
$$\sum_\alpha\sum_j m_\alpha\left(\frac{\partial^2 \mathbf{G}_\alpha}{\partial q^k \partial q^j}\cdot\frac{\partial \mathbf{G}_\alpha}{\partial t}\right)\dot{q}^j + \sum_\alpha\sum_j m_\alpha\left(\frac{\partial \mathbf{G}_\alpha}{\partial q^j}\cdot\frac{\partial^2 \mathbf{G}_\alpha}{\partial q^k \partial t}\right)\dot{q}^j$$
$$+\sum_\alpha m_\alpha\left(\frac{\partial^2 \mathbf{G}_\alpha}{\partial q^k \partial t}\cdot\frac{\partial \mathbf{G}_\alpha}{\partial t}\right). \quad (28.23)$$

Now upon comparing (28.23) and (28.21)

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}^k}\right) - \frac{\partial T}{\partial q^k} = \sum_{\alpha=1}^{N}\frac{\partial \mathbf{G}_\alpha}{\partial q^k}\cdot\mathbf{F}_\alpha.$$

Resolve the force, $\mathbf{F}_\alpha$ into the sum of two forces, $\mathbf{F}_\alpha = \mathbf{F}_\alpha^a + \mathbf{F}_\alpha^c$ where $\mathbf{F}_\alpha^c$ is a force of constraint which is perpendcular to $\frac{\partial \mathbf{G}_\alpha}{\partial q^k}$ and the other force, $\mathbf{F}_\alpha^a$ which is left over is called the applied force. The applied force is allowed to have a component which is perpendicular to $\frac{\partial \mathbf{G}_\alpha}{\partial q^k}$. The only requirement of this sort is placed on $\mathbf{F}_\alpha^c$. Therefore,

$$\frac{\partial \mathbf{G}_\alpha}{\partial q^k}\cdot\mathbf{F}_\alpha = \frac{\partial \mathbf{G}_\alpha}{\partial q^k}\cdot\mathbf{F}_\alpha^a$$

and so in the end, you obtain the following interesting equation which is equivalent to Newton's second law.

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}^k}\right) - \frac{\partial T}{\partial q^k} = \sum_{\alpha=1}^{N}\frac{\partial \mathbf{G}_\alpha}{\partial q^k}\cdot\mathbf{F}_\alpha^a \quad (28.24)$$
$$= \frac{\partial \mathbf{G}}{\partial q^k}\cdot\mathbf{F}^a, \quad (28.25)$$

where $\mathbf{F}^a \equiv (\mathbf{F}_1^\alpha, \cdots, \mathbf{F}_N^\alpha)$ is referred to as the total applied force.

It is particularly agreeable when the total applied force comes as the gradient of a potential function. This means there exists a scalar function of $\mathbf{x}$, ϕ defined near $\mathbf{G}(V)$ such that

$$\mathbf{F}_\alpha^a(\mathbf{x},t) = -\nabla_\alpha\phi(\mathbf{x},t)$$

where the symbol ∇_α denotes the gradient with respect to $\mathbf{x}_\alpha$. More generally,

$$\mathbf{F}_\alpha^a(\mathbf{x},t) = -\nabla_\alpha \phi(\mathbf{x},t) + \mathbf{F}_\alpha^d$$

where $\mathbf{F}_\alpha^d$ is a force which is not a force of constraint or the gradient of a given function. For example, it could be a force of friction. Then

$$\mathbf{F}^a(\mathbf{x},t) = -\nabla \phi(\mathbf{x},t) + \mathbf{F}^d$$

where

$$\mathbf{F}^d = \left(\mathbf{F}_1^d, \cdots, \mathbf{F}_N^d\right)$$

Now let $T(\mathbf{q},\dot{\mathbf{q}}) - \phi(\mathbf{G}(\mathbf{q},t)) = L(\mathbf{q},\dot{\mathbf{q}})$. Then letting x^j denote the usual Cartesian coordinates of $\mathbf{x}$,

$$\begin{aligned}\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^k}\right) - \frac{\partial L}{\partial q^k} &= \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}^k}\right) - \frac{\partial T}{\partial q^k} + \sum_j \frac{\partial \phi(\mathbf{x})}{\partial x^j}\frac{\partial x^j}{\partial q^k} \\ &= \frac{\partial \mathbf{G}}{\partial q^k}\cdot\left(-\nabla\phi(\mathbf{x}) + \mathbf{F}^d\right) + \frac{\partial \mathbf{G}}{\partial q^k}\cdot\nabla\phi = \frac{\partial \mathbf{G}}{\partial q^k}\cdot\mathbf{F}^d. \qquad (28.26)\end{aligned}$$

These are called Lagrange's equations of motion and they are enormously significant because it is often possible to find the kinetic and potential energy in terms of variables q^k which are meaningful for a particular problem. The expression, $L(\mathbf{q},\dot{\mathbf{q}})$ is called the Lagrangian. This has proved part of the following theorem.

Theorem 28.5.1 *In the above context Newton's second law implies*

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^k}\right) - \frac{\partial L}{\partial q^k} = \frac{\partial \mathbf{G}}{\partial q^k}\cdot\mathbf{F}^d. \qquad (28.27)$$

In particular, if the applied force is the gradient of $-\phi$, the right side reduces to 0(28.27). If, in addition to this, the potential function is time independent then the total energy is conserved. That is,

$$T(\mathbf{q},\dot{\mathbf{q}}) + \phi(\mathbf{G}(\mathbf{q},t)) = C \qquad (28.28)$$

for some constant, C.

Proof: It remains to verify the assertion about the energy. In terms of the Cartesian coordinates,

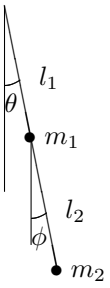
$$E = \sum_\alpha \frac{1}{2} m_\alpha \dot{\mathbf{x}}_\alpha \cdot \dot{\mathbf{x}}_\alpha + \phi(\mathbf{x},t).$$

Recall the applied force is given by $\mathbf{F}_\alpha^a = -\nabla_\alpha\phi(\mathbf{x},t) + \mathbf{F}_\alpha^d$. Differentiating with respect to time,

$$\begin{aligned}\frac{dE}{dt} &= \sum_\alpha m_\alpha \ddot{\mathbf{x}}_\alpha \cdot \dot{\mathbf{x}}_\alpha + \sum_j \frac{\partial \phi}{\partial x^j}\dot{x}^j + \frac{\partial \phi}{\partial t} \\ &= \sum_\alpha \mathbf{F}_\alpha \cdot \dot{\mathbf{x}}_\alpha + \sum_\alpha \nabla_\alpha \phi(\mathbf{x},t)\cdot\dot{\mathbf{x}}_\alpha + \frac{\partial \phi}{\partial t} \\ &= \sum_\alpha \mathbf{F}_\alpha^a \cdot \dot{\mathbf{x}}_\alpha + \sum_\alpha \nabla_\alpha \phi(\mathbf{x},t)\cdot\dot{\mathbf{x}}_\alpha + \frac{\partial \phi}{\partial t} \\ &= \sum_\alpha \left(-\nabla_\alpha\phi(\mathbf{x},t) + \mathbf{F}_\alpha^d\right)\cdot\dot{\mathbf{x}}_\alpha + \sum_\alpha \nabla_\alpha\phi(\mathbf{x},t)\cdot\dot{\mathbf{x}}_\alpha + \frac{\partial\phi}{\partial t} \\ &= \sum_\alpha \mathbf{F}_\alpha^d \cdot \dot{\mathbf{x}}_\alpha + \frac{\partial \phi}{\partial t}.\end{aligned}$$

Therefore, this shows (28.28) because in the case described, $\mathbf{F}_\alpha^d = 0$ and $\frac{\partial \phi}{\partial t} = 0$. In the case of friction, $\mathbf{F}_\alpha^d \cdot \dot{\mathbf{x}}_\alpha \leq 0$ and so in this case, if ϕ is time independent, the total energy is decreasing.

Example 28.5.2 *Consider the double pendulum.*



1.

It is fairly easy to find the equations of motion in terms of the variables, ϕ and θ. These variables are the q^k mentioned above. Because the two rods joining the masses have fixed length, a constraint is introduced on the motion of the two masses. It is clear the position of these masses is specified from the two variables, θ and ϕ. In fact, letting the origin be located at the point at the top where the pendulum is suspended and assuming the vibration is in a plane,

$$\mathbf{x}_1 = (l_1 \sin\theta, -l_1 \cos\theta)$$

and

$$\mathbf{x}_2 = (l_1 \sin\theta + l_2 \sin\phi, -l_1 \cos\theta - l_2 \cos\phi)\,.$$

Therefore,

$$\begin{aligned} \dot{\mathbf{x}}_1 &= \left(l_1\dot{\theta}\cos\theta, l_1\dot{\theta}\sin\theta\right) \\ \dot{\mathbf{x}}_2 &= \left(l_1\dot{\theta}\cos\theta + l_2\dot{\phi}\cos\phi, l_1\dot{\theta}\sin\theta + l_2\dot{\phi}\sin\phi\right). \end{aligned}$$

It follows the kinetic energy is given by

$$T = \frac{1}{2}m_2\left(2l_1\dot{\theta}\left(\cos\theta\right)l_2\dot{\phi}\cos\phi + l_1^2(\dot{\theta})^2 + 2l_1\dot{\theta}\left(\sin\theta\right)l_2\dot{\phi}\sin\phi + l_2^2(\dot{\phi})^2\right) + \frac{1}{2}m_1\left(l_1^2(\dot{\theta})^2\right).$$

There are forces of constraint acting on these masses and there is the force of gravity acting on them. The force from gravity on m_1 is $-m_1 g$ and the force from gravity on m_2 is $-m_2 g$. Our function, ϕ is just the total potential energy. Thus $\phi\left(\mathbf{x}_1, \mathbf{x}_2\right) = m_1 g y_1 + m_2 g y_2$. It follows that $\phi\left(\mathbf{G}\left(\mathbf{q}\right)\right) = m_1 g\left(-l_1\cos\theta\right) + m_2 g\left(-l_1\cos\theta - l_2\cos\phi\right)$. Therefore, the Lagrangian, L, is

$$\begin{aligned} &\frac{1}{2}m_2\left(2l_1l_2\dot{\theta}\dot{\phi}\left(\cos\left(\phi-\theta\right)\right) + l_1^2(\dot{\theta})^2 + l_2^2(\dot{\phi})^2\right) + \frac{1}{2}m_1\left(l_1^2(\dot{\theta})^2\right) \\ &\qquad - \left[m_1 g\left(-l_1\cos\theta\right) + m_2 g\left(-l_1\cos\theta - l_2\cos\phi\right)\right]. \end{aligned}$$

It now becomes an easy task to find the equations of motion in terms of the two angles, θ and ϕ.

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}}\right) - \frac{\partial L}{\partial \theta} =$$

$$\begin{aligned}&\theta''\left(m_1+m_2\right)l_1^2+m_2l_2l_1\cos\left(\phi-\theta\right)\phi''-m_2l_1l_2\sin\left(\phi-\theta\right)\left(\phi'-\theta'\right)\phi'\\&+\left(m_1+m_2\right)gl_1\sin\theta-m_2l_1l_2\theta'\phi'\sin\left(\phi-\theta\right)\\&=\theta''\left(m_1+m_2\right)l_1^2+m_2l_2l_1\cos\left(\phi-\theta\right)\phi''-m_2l_1l_2\sin\left(\phi-\theta\right)\phi'^2\\&+\left(m_1+m_2\right)gl_1\sin\theta=0.\end{aligned}\tag{28.29}$$

To get the other equation,

$$\frac{d}{dt}\left(\frac{\partial L}{\partial\dot{\phi}}\right)-\frac{\partial L}{\partial\phi}=$$

$$\begin{aligned}&\frac{d}{dt}\left[m_2l_1l_2\dot{\theta}\left(\cos\left(\phi-\theta\right)\right)+m_2l_2^2\dot{\phi}\right]+m_2gl_2\sin\phi-\left(-m_2l_1l_2\dot{\theta}\dot{\phi}\sin\left(\phi-\theta\right)\right)\\&=m_2l_1l_2\theta''\cos\left(\phi-\theta\right)-m_2l_1l_2\theta'\sin\left(\phi-\theta\right)\left(\phi'-\theta'\right)+m_2l_2^2\phi''+m_2gl_2\sin\phi+m_2l_2l_1\phi'\theta'\sin\left(\phi-\theta\right)\\&=m_2l_1l_2\theta''\cos\left(\phi-\theta\right)+m_2l_1l_2\left(\theta'\right)^2\sin\left(\phi-\theta\right)+m_2l_2^2\phi''+m_2gl_2\sin\phi=0\end{aligned}\tag{28.30}$$

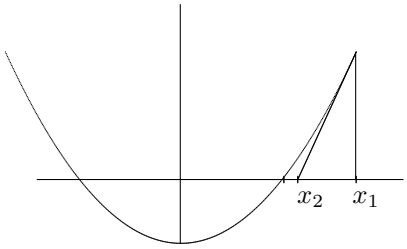
Admittedly, (28.29) and (28.30) are horrific equations, but what would you expect from something as complicated as the double pendulum? They can at least be solved numerically. The conservation of energy gives some idea what is going on. Thus

$$\begin{aligned}&\frac{1}{2}m_2\left(2l_1l_2\dot{\theta}\dot{\phi}\left(\cos\left(\phi-\theta\right)\right)+l_1^2(\dot{\theta})^2+l_2^2(\dot{\phi})^2\right)+\frac{1}{2}m_1\left(l_1^2(\dot{\theta})^2\right)\\&+\left[m_1g\left(-l_1\cos\theta\right)+m_2g\left(-l_1\cos\theta-l_2\cos\phi\right)\right]=C.\end{aligned}$$

28.6 Newton's Method

28.6.1 The Newton Raphson Method In One Dimension

The Newton Raphson method is a way to get approximations of solutions to various equations. For example, suppose you want to find $\sqrt{2}$. The existence of $\sqrt{2}$ is not difficult to establish by considering the continuous function, $f\left(x\right)=x^2-2$ which is negative at $x=0$ and positive at $x=2$. Therefore, by the intermediate value theorem, there exists $x\in(0,2)$ such that $f\left(x\right)=0$ and this x must equal $\sqrt{2}$. The problem consists of how to find this number, not just to prove it exists. The following picture illustrates the procedure of the Newton Raphson method.



In this picture, a first approximation, denoted in the picture as x_1 is chosen and then the tangent line to the curve $y=f\left(x\right)$ at the point $\left(x_1,f\left(x_1\right)\right)$ is obtained. The equation of this tangent line is

$$y-f\left(x_1\right)=f'\left(x_1\right)\left(x-x_1\right).$$

Then extend this tangent line to find where it intersects the x axis. In other words, set $y=0$ and solve for x. This value of x is denoted by x_2. Thus

$$x_2=x_1-\frac{f\left(x_1\right)}{f'\left(x_1\right)}.$$

This second point, x_2 is the second approximation and the same process is done for x_2 that was done for x_1 in order to get the third approximation, x_3. Thus

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}.$$

Continuing this way, yields a sequence of points, $\{x_n\}$ given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{28.31}$$

which hopefully has the property that $\lim_{n\to\infty} x_n = x$ where $f(x) = 0$. You can see from the above picture that this must work out in the case of $f(x) = x^2 - 2$.

Now carry out the computations in the above case for $x_1 = 2$ and $f(x) = x^2 - 2$. From (28.31),

$$x_2 = 2 - \frac{2}{4} = 1.5.$$

Then

$$x_3 = 1.5 - \frac{(1.5)^2 - 2}{2(1.5)} \leq 1.417,$$

$$x_4 = 1.417 - \frac{(1.417)^2 - 2}{2(1.417)} = 1.414\,216\,302\,046\,577,$$

What is the true value of $\sqrt{2}$? To several decimal places this is $\sqrt{2} = 1.414\,213\,562\,373\,095$, showing that the Newton Raphson method has yielded a very good approximation after only a few iterations, even starting with an initial approximation, 2, which was not very good.

This method does not always work. For example, suppose you wanted to find the solution to $f(x) = 0$ where $f(x) = x^{1/3}$. You should check that the sequence of iterates which results does not converge. This is because, starting with x_1 the above procedure yields $x_2 = -2x_1$ and so as the iteration continues, the sequence oscillates between positive and negative values as its absolute value gets larger and larger. The problem is that $f'(0)$ does not exist.

However, if $f(x_0) = 0$ and $f''(x) > 0$ for x near x_0, you can draw a picture to show that the method will yield a sequence which converges to x_0 provided the first approximation, x_1 is taken sufficiently close to x_0. Similarly, if $f''(x) < 0$ for x near x_0, then the method produces a sequence which converges to x_0 provided x_1 is close enough to x_0.

28.6.2 Newton's Method For Nonlinear Systems

The same formula yields a procedure for finding solutions to systems of functions of n variables. This is particularly interesting because you can't make any sense of things from drawing pictures. The technique of graphing and zooming which really works well for functions of one variable is no longer available.

Procedure 28.6.1 *Suppose* $\mathbf{f}$ *is a* C^1 *function of* n *variables and* $\mathbf{f}(\mathbf{z}) = \mathbf{0}$. *Then to find* $\mathbf{z}$, *you use the same iteration which you would use in one dimension,*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - D\mathbf{f}(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k)$$

where $\mathbf{x}_0$ *is an initial approximation chosen close to* $\mathbf{z}$.

Example 28.6.2 *Find a solution to the nonlinear system of equations,*

$$\mathbf{f}(x,y) = \begin{pmatrix} x^3 - 3xy^2 - 3x^2 + 3y^2 + 7x - 5 \\ 3x^2y - y^3 - 6xy + 7y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

You can verify that $(x,y) = (1,2)$, $(1,-2)$, and $(1,0)$ all are solutions to the above system. Suppose then that you didn't know this.

$$D\mathbf{f}(x,y) = \begin{pmatrix} 3x^2 - 3y^2 - 6x + 7 & -6xy + 6y \\ 6xy - 6y & 3x^2 - 3y^2 - 6x + 7 \end{pmatrix}$$

Start with an initial guess $(x_0, y_0) = (1,3)$. Then the next iteration is

$$\begin{pmatrix} 1 \\ 3 \end{pmatrix} - \begin{pmatrix} -23 & 0 \\ 0 & -23 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{72}{23} \end{pmatrix}$$

The next iteration is

$$\begin{aligned} & \begin{pmatrix} 1 \\ \frac{72}{23} \end{pmatrix} - \begin{pmatrix} -3.937\,183\,7 \times 10^{-2} & 0 \\ 0 & -3.937\,183\,7 \times 10^{-2} \end{pmatrix} \begin{pmatrix} 0 \\ -18.155\,338 \end{pmatrix} \\ = & \begin{pmatrix} 1.0 \\ 2.415\,625\,8 \end{pmatrix} \end{aligned}$$

I will not bother to use all the decimals in 2.4156258. The next iteration is

$$\begin{aligned} & \begin{pmatrix} 1.0 \\ 2.4 \end{pmatrix} - \begin{pmatrix} -7.530\,120\,5 \times 10^{-2} & 0 \\ 0 & -7.530\,120\,5 \times 10^{-2} \end{pmatrix} \begin{pmatrix} 0 \\ -4.224 \end{pmatrix} \\ = & \begin{pmatrix} 1.0 \\ 2.081\,927\,7 \end{pmatrix}. \end{aligned}$$

Notice how the process is converging to the solution $(x,y) = (1,2)$. If you do one more iteration, you will be really close.

The above was pretty painful because at every step the derivative had to be re evaluated and the inverse taken. It turns out a simpler procedure will work in which you don't have to constantly re evaluate the inverse of the derivative.

Procedure 28.6.3 *Suppose* $\mathbf{f}$ *is a* C^1 *function of* n *variables and* $\mathbf{f}(\mathbf{z}) = \mathbf{0}$. *Then to find* $\mathbf{z}$, *you can use the following iteration procedure*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - D\mathbf{f}(\mathbf{x}_0)^{-1} \mathbf{f}(\mathbf{x}_k)$$

where $\mathbf{x}_0$ *is an initial approximation chosen close to* $\mathbf{z}$.

To illustrate, I will use this new procedure on the same example.

Example 28.6.4 *Find a solution to the nonlinear system of equations,*

$$\mathbf{f}(x,y) = \begin{pmatrix} x^3 - 3xy^2 - 3x^2 + 3y^2 + 7x - 5 \\ 3x^2y - y^3 - 6xy + 7y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

You can verify that $(x,y) = (1,2)$, $(1,-2)$, and $(1,0)$ all are solutions to the above system. Suppose then that you didn't know this. Take $(x_0, y_0) = (1,3)$ as above. Then a little computation will show

$$D\mathbf{f}(1,3)^{-1} = \begin{pmatrix} -\frac{1}{23} & 0 \\ 0 & -\frac{1}{23} \end{pmatrix}$$

The first iteration is then $\begin{array}{c} 1-3\left(2.116\,087\,3\right)^{2}-3+3\left(2.116\,087\,3\right)^{2}+7-5 \\ 3\left(2.116\,087\,3\right)-\left(2.116\,087\,3\right)^{3}-6\left(2.116\,087\,3\right)+7\left(2.116\,087\,3\right) \end{array} =$

$$\begin{array}{c} 0 \\ -1.011\,120\,4 \end{array}$$

$$\begin{aligned} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} &= \begin{pmatrix} 1 \\ 3 \end{pmatrix} - \begin{pmatrix} -\frac{1}{23} & 0 \\ 0 & -\frac{1}{23} \end{pmatrix} \begin{pmatrix} 0 \\ -15.0 \end{pmatrix} \\ &= \begin{pmatrix} 1.0 \\ 2.347\,826\,1 \end{pmatrix} \end{aligned}$$

The next iteration is

$$\begin{aligned} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} &= \begin{pmatrix} 1.0 \\ 2.347\,826\,1 \end{pmatrix} - \begin{pmatrix} -\frac{1}{23} & 0 \\ 0 & -\frac{1}{23} \end{pmatrix} \begin{pmatrix} 0 \\ -3.550\,587\,8 \end{pmatrix} \\ &= \begin{pmatrix} 1.0 \\ 2.193\,452\,7 \end{pmatrix} \end{aligned}$$

The next iteration is

$$\begin{aligned} \begin{pmatrix} x_3 \\ y_3 \end{pmatrix} &= \begin{pmatrix} 1.0 \\ 2.193\,452\,7 \end{pmatrix} - \begin{pmatrix} -\frac{1}{23} & 0 \\ 0 & -\frac{1}{23} \end{pmatrix} \begin{pmatrix} 0 \\ -1.779\,405 \end{pmatrix} \\ &= \begin{pmatrix} 1.0 \\ 2.116\,087\,3 \end{pmatrix} \end{aligned}$$

The next iteration is

$$\begin{aligned} \begin{pmatrix} x_4 \\ y_4 \end{pmatrix} &= \begin{pmatrix} 1.0 \\ 2.116\,087\,3 \end{pmatrix} - \begin{pmatrix} -\frac{1}{23} & 0 \\ 0 & -\frac{1}{23} \end{pmatrix} \begin{pmatrix} 0 \\ -1.011\,120\,4 \end{pmatrix} \\ &= \begin{pmatrix} 1.0 \\ 2.072\,125\,5 \end{pmatrix}. \end{aligned}$$

You see it appears to be converging to a zero of the nonlinear system. It is doing so more slowly than in the case of Newton's method but there is less trouble involved in each step of the iteration.

Of course there is a question about how to choose the initial approximation. There are methods for doing this called homotopy methods which are based on numerical methods for differential equations. The idea for these methods is to consider the problem

$$(1-t)\left(\mathbf{x}-\mathbf{x}_0\right)+t\mathbf{f}\left(\mathbf{x}\right)=\mathbf{0}.$$

When $t=0$ this reduces to $\mathbf{x}=\mathbf{x}_0$. Then when $t=1$, it reduces to $\mathbf{f}\left(\mathbf{x}\right)=\mathbf{0}$. The equation specifies $\mathbf{x}$ as a function of t (hopefully). Differentiating with respect to t, you see that $\mathbf{x}$ must solve the following initial value problem,

$$-\left(\mathbf{x}-\mathbf{x}_0\right)+\left(1-t\right)\mathbf{x}'+\mathbf{f}\left(\mathbf{x}\right)+tD\mathbf{f}\left(\mathbf{x}\right)\mathbf{x}'=\mathbf{0},\ \mathbf{x}\left(0\right)=\mathbf{x}_0.$$

where $\mathbf{x}'$ denotes the time derivative of the vector $\mathbf{x}$. Initial value problems of this sort are routinely solvable using standard numerical methods. The idea is you solve it on $[0,1]$ and your zero is $\mathbf{x}\left(1\right)$. Because of roundoff error, $\mathbf{x}\left(1\right)$ won't be quite right so you use it as an initial guess in Newton's method and find the zero to great accuracy.

28.7 Convergence Questions*

28.7.1 A Fixed Point Theorem

The message of this section is that under reasonable conditions amounting to an assumption that $D\mathbf{f}(\mathbf{z})^{-1}$ exists, Newton's method will converge whenever you take an initial approximation sufficiently close to $\mathbf{z}$. This is just like the situation for the method in one dimension.

The proof of convergence rests on the following lemma which is somewhat more interesting than Newton's method. It is a case of the contraction mapping principle important in differential and integral equations.

Lemma 28.7.1 *Suppose $T : B(\mathbf{x}_0, \delta) \subseteq \mathbb{R}^p \to \mathbb{R}^p$ and it satisfies*

$$|T\mathbf{x} - T\mathbf{y}| \leq \frac{1}{2}|\mathbf{x} - \mathbf{y}| \text{ for all } \mathbf{x}, \mathbf{y} \in B(\mathbf{x}_0, \delta). \tag{28.32}$$

Suppose also that $|T\mathbf{x}_0 - \mathbf{x}_0| < \frac{\delta}{4}$. Then $\{T^n\mathbf{x}_0\}_{n=1}^{\infty}$ converges to a point, $\mathbf{x} \in B(\mathbf{x}_0, \delta)$ such that $T\mathbf{x} = \mathbf{x}$. This point is called a fixed point. Furthermore, there is at most one fixed point on $B(\mathbf{x}_0, \delta)$.

Proof: From the triangle inequality, and the use of (28.32),

$$\begin{aligned} |T^n\mathbf{x}_0 - \mathbf{x}_0| &\leq \sum_{k=1}^{n} \left|T^k\mathbf{x}_0 - T^{k-1}\mathbf{x}_0\right| \\ &\leq \sum_{k=1}^{n} \left(\frac{1}{2}\right)^{k-1} |T\mathbf{x}_0 - \mathbf{x}_0| \\ &\leq 2|T\mathbf{x}_0 - \mathbf{x}_0| < 2\frac{\delta}{4} = \frac{\delta}{2} < \delta. \end{aligned}$$

Thus the sequence remains in the closed ball, $\overline{B(\mathbf{x}_0, \delta/2)} \subseteq B(\mathbf{x}_0, \delta)$. Also, by similar reasoning,

$$\begin{aligned} |T^n\mathbf{x}_0 - T^m\mathbf{x}_0| &\leq \sum_{k=m}^{n} \left|T^{k+1}\mathbf{x}_0 - T^k\mathbf{x}_0\right| \leq \sum_{k=m}^{n} \left(\frac{1}{2}\right)^{k} |T\mathbf{x}_0 - \mathbf{x}_0| \\ &\leq \frac{\delta}{4}\frac{1}{2^{m-1}}. \end{aligned}$$

It follows, that $\{T^n\mathbf{x}_0\}$ is a Cauchy sequence. Therefore, it converges to a point of $\overline{B(\mathbf{x}_0, \delta/2)} \subseteq B(\mathbf{x}_0, \delta)$. Call this point, $\mathbf{x}$. Then since T is continuous, it follows $\mathbf{x} = \lim_{n\to\infty} T^n\mathbf{x}_0 = T\lim_{n\to\infty} T^{n-1}\mathbf{x}_0 = T\mathbf{x}_0$. If $T\mathbf{x} = \mathbf{x}$ and $T\mathbf{y} = \mathbf{y}$ for $\mathbf{x}, \mathbf{y} \in B(\mathbf{x}_0, \delta)$ then $|\mathbf{x} - \mathbf{y}| = |T\mathbf{x} - T\mathbf{y}| \leq \frac{1}{2}|\mathbf{x} - \mathbf{y}|$ and so $\mathbf{x} = \mathbf{y}$.

28.7.2 The Operator Norm

How do you measure the distance between linear transformations defined on $\mathbb{F}^n$? It turns out there are many ways to do this but I will give the most common one here.

Definition 28.7.2 *$\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ denotes the space of linear transformations mapping $\mathbb{F}^n$ to $\mathbb{F}^m$. For $A \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$, the **operator norm** is defined by*

$$||A|| \equiv \max\{|Ax|_{\mathbb{F}^m} : |x|_{\mathbb{F}^n} \leq 1\} < \infty.$$

Theorem 28.7.3 *Denote by $|\cdot|$ the norm on either $\mathbb{F}^n$ or $\mathbb{F}^m$. Then $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ with this operator norm is a **complete normed linear space** of dimension nm with*

$$||A\mathbf{x}|| \leq ||A||\,|\mathbf{x}|.$$

*Here **Completeness** means that every Cauchy sequence converges.*

Proof: It is necessary to show the norm defined on $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ really is a norm. This means it is necessary to verify

$$||A|| \geq 0 \text{ and equals zero if and only if } A = 0.$$

For α a scalar,

$$||\alpha A|| = |\alpha|\,||A||,$$

and for $A, B \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$,

$$||A + B|| \leq ||A|| + ||B||$$

The first two properties are obvious but you should verify them. It remains to verify the norm is well defined and also to verify the triangle inequality above. First if $|\mathbf{x}| \leq 1$, and (A_{ij}) is the matrix of the linear transformation with respect to the usual basis vectors, then

$$\begin{aligned} ||A|| &= \max\left\{\left(\sum_i |(A\mathbf{x})_i|^2\right)^{1/2} : |\mathbf{x}| \leq 1\right\} \\ &= \max\left\{\left(\sum_i \left|\sum_j A_{ij}x_j\right|^2\right)^{1/2} : |\mathbf{x}| \leq 1\right\} \end{aligned}$$

which is a finite number by the extreme value theorem.

It is clear that a basis for $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ consists of linear transformations whose matrices are of the form E_{ij} where E_{ij} consists of the $m \times n$ matrix having all zeros except for a 1 in the ij^{th} position. In effect, this considers $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ as $\mathbb{F}^{nm}$. Think of the $m \times n$ matrix as a long vector folded up.

If $\mathbf{x} \neq \mathbf{0}$,

$$|A\mathbf{x}|\,\frac{1}{|\mathbf{x}|} = \left|A\frac{\mathbf{x}}{|\mathbf{x}|}\right| \leq ||A|| \tag{28.33}$$

It only remains to verify completeness. Suppose then that $\{A_k\}$ is a Cauchy sequence in $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$. Then from (28.33) $\{A_k\mathbf{x}\}$ is a Cauchy sequence for each $\mathbf{x} \in \mathbb{F}^n$. This follows because

$$|A_k\mathbf{x} - A_l\mathbf{x}| \leq ||A_k - A_l||\,|\mathbf{x}|$$

which converges to 0 as $k, l \to \infty$. Therefore, by completeness of $\mathbb{F}^m$, there exists $A\mathbf{x}$, the name of the thing to which the sequence, $\{A_k\mathbf{x}\}$ converges such that

$$\lim_{k\to\infty} A_k\mathbf{x} = A\mathbf{x}.$$

Then A is linear because

$$\begin{aligned} A(a\mathbf{x}+b\mathbf{y}) &\equiv \lim_{k\to\infty} A_k(a\mathbf{x}+b\mathbf{y}) \\ &= \lim_{k\to\infty} (aA_k\mathbf{x}+bA_k\mathbf{y}) \\ &= a\lim_{k\to\infty} A_k\mathbf{x} + b\lim_{k\to\infty} A_k\mathbf{y} \\ &= aA\mathbf{x}+bA\mathbf{y}. \end{aligned}$$

By the first part of this argument, $||A|| < \infty$ and so $A \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$. This proves the theorem.

The following is an interesting exercise which is left for you.

Proposition 28.7.4 *Let $A(\mathbf{x}) \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ for each $\mathbf{x} \in U \subseteq \mathbb{F}^p$. Then letting $(A_{ij}(\mathbf{x}))$ denote the matrix of $A(\mathbf{x})$ with respect to the standard basis, it follows A_{ij} is continuous at $\mathbf{x}$ for each i, j if and only if for all $\varepsilon > 0$, there exists a $\delta > 0$ such that if $|\mathbf{x}-\mathbf{y}| < \delta$, then $||A(\mathbf{x}) - A(\mathbf{y})|| < \varepsilon$. That is, A is a continuous function having values in $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ at $\mathbf{x}$.*

Proof: Suppose first the second condition holds. Then from the material on linear transformations,

$$\begin{aligned} |A_{ij}(\mathbf{x}) - A_{ij}(\mathbf{y})| &= |\mathbf{e}_i \cdot (A(\mathbf{x}) - A(\mathbf{y}))\mathbf{e}_j| \\ &\le |\mathbf{e}_i|\,|(A(\mathbf{x}) - A(\mathbf{y}))\mathbf{e}_j| \\ &\le ||A(\mathbf{x}) - A(\mathbf{y})||. \end{aligned}$$

Therefore, the second condition implies the first.

Now suppose the first condition holds. That is each A_{ij} is continuous at $\mathbf{x}$. Let $|\mathbf{v}| \le 1$.

$$\begin{aligned} |(A(\mathbf{x}) - A(\mathbf{y}))(\mathbf{v})| &= \left(\sum_i \left|\sum_j (A_{ij}(\mathbf{x}) - A_{ij}(\mathbf{y}))\, v_j\right|^2\right)^{1/2} \qquad (28.34) \\ &\le \left(\sum_i \left(\sum_j |A_{ij}(\mathbf{x}) - A_{ij}(\mathbf{y})|\,|v_j|\right)^2\right)^{1/2}. \end{aligned}$$

By continuity of each A_{ij}, there exists a $\delta > 0$ such that for each i, j

$$|A_{ij}(\mathbf{x}) - A_{ij}(\mathbf{y})| < \frac{\varepsilon}{n\sqrt{m}}$$

whenever $|\mathbf{x}-\mathbf{y}| < \delta$. Then from (28.34), if $|\mathbf{x}-\mathbf{y}| < \delta$,

$$\begin{aligned} |(A(\mathbf{x}) - A(\mathbf{y}))(\mathbf{v})| &< \left(\sum_i \left(\sum_j \frac{\varepsilon}{n\sqrt{m}}|\mathbf{v}|\right)^2\right)^{1/2} \\ &\le \left(\sum_i \left(\sum_j \frac{\varepsilon}{n\sqrt{m}}\right)^2\right)^{1/2} = \varepsilon \end{aligned}$$

This proves the proposition.

The proposition implies that a function is C^1 if and only if the derivative, $D\mathbf{f}$ exists and the function, $\mathbf{x} \to D\mathbf{f}(\mathbf{x})$ is continuous in the usual way. That is, for all $\varepsilon > 0$ there exists $\delta > 0$ such that if $|\mathbf{x} - \mathbf{y}| < \delta$, then $||D\mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{y})|| < \varepsilon$.

The following is a version of the mean value theorem valid for functions defined on $\mathbb{R}^n$.

Theorem 28.7.5 *Suppose U is an open subset of $\mathbb{R}^p$ and $\mathbf{f} : U \to \mathbb{R}^q$ has the property that $D\mathbf{f}(\mathbf{x})$ exists for all $\mathbf{x}$ in U and that, $\mathbf{x}+t(\mathbf{y}-\mathbf{x}) \in U$ for all $t \in [0,1]$. (The line segment joining the two points lies in U.) Suppose also that for all points on this line segment,*

$$||D\mathbf{f}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))|| \leq M.$$

Then

$$||\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})|| \leq M\,||\mathbf{y}-\mathbf{x}||.$$

Proof: Let

$$S \equiv \{t \in [0,1] : \text{ for all } s \in [0,t],$$

$$||\mathbf{f}(\mathbf{x}+s(\mathbf{y}-\mathbf{x})) - \mathbf{f}(\mathbf{x})|| \leq (M+\varepsilon)\,s\,||\mathbf{y}-\mathbf{x}||\}.$$

Then $0 \in S$ and by continuity of $\mathbf{f}$, it follows that if $t \equiv \sup S$, then $t \in S$ and if $t < 1$,

$$||\mathbf{f}(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \mathbf{f}(\mathbf{x})|| = (M+\varepsilon)\,t\,||\mathbf{y}-\mathbf{x}||. \tag{28.35}$$

If $t < 1$, then there exists a sequence of positive numbers, $\{h_k\}_{k=1}^{\infty}$ converging to 0 such that

$$||\mathbf{f}(\mathbf{x}+(t+h_k)(\mathbf{y}-\mathbf{x})) - \mathbf{f}(\mathbf{x})|| > (M+\varepsilon)(t+h_k)\,||\mathbf{y}-\mathbf{x}||$$

which implies that

$$||\mathbf{f}(\mathbf{x}+(t+h_k)(\mathbf{y}-\mathbf{x})) - \mathbf{f}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))||$$

$$+\,||\mathbf{f}(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \mathbf{f}(\mathbf{x})|| > (M+\varepsilon)(t+h_k)\,||\mathbf{y}-\mathbf{x}||.$$

By (28.35), this inequality implies

$$||\mathbf{f}(\mathbf{x}+(t+h_k)(\mathbf{y}-\mathbf{x})) - \mathbf{f}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))|| > (M+\varepsilon)\,h_k\,||\mathbf{y}-\mathbf{x}||$$

which yields upon dividing by h_k and taking the limit as $h_k \to 0$,

$$||D\mathbf{f}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))(\mathbf{y}-\mathbf{x})|| \geq (M+\varepsilon)\,||\mathbf{y}-\mathbf{x}||.$$

Now by the definition of the norm of a linear operator,

$$M\,||\mathbf{y}-\mathbf{x}|| \geq ||D\mathbf{f}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))||\,||\mathbf{y}-\mathbf{x}|| \geq ||D\mathbf{f}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))(\mathbf{y}-\mathbf{x})|| \geq (M+\varepsilon)\,||\mathbf{y}-\mathbf{x}||,$$

a contradiction. Therefore, $t = 1$ and so

$$||\mathbf{f}(\mathbf{x}+(\mathbf{y}-\mathbf{x})) - \mathbf{f}(\mathbf{x})|| \leq (M+\varepsilon)\,||\mathbf{y}-\mathbf{x}||.$$

Since $\varepsilon > 0$ is arbitrary, this proves the theorem.

28.7.3 A Method For Finding Zeros

Theorem 28.7.6 *Suppose* $\mathbf{f} : U \subseteq \mathbb{R}^p \to \mathbb{R}^p$ *is a* C^1 *function and suppose* $\mathbf{f}(\mathbf{z}) = \mathbf{0}$. *Suppose also that for all* $\mathbf{x}$ *sufficiently close to* $\mathbf{z}$, *it follows that* $D\mathbf{f}(\mathbf{x})^{-1}$ *exists. Let* $\delta > 0$ *be small enough that for all* $\mathbf{x}, \mathbf{x}_0 \in B(\mathbf{z}, 2\delta)$

$$\left|\left|I - D\mathbf{f}(\mathbf{x}_0)^{-1} D\mathbf{f}(\mathbf{x})\right|\right| < \frac{1}{2}. \tag{28.36}$$

Now pick $\mathbf{x}_0 \in B(\mathbf{z}, \delta)$ *also close enough to* $\mathbf{z}$ *such that*

$$\left|\left|D\mathbf{f}(\mathbf{x}_0)^{-1}\right|\right| |\mathbf{f}(\mathbf{x}_0)| < \frac{\delta}{4}.$$

Define

$$T\mathbf{x} \equiv \mathbf{x} - D\mathbf{f}(\mathbf{x}_0)^{-1} \mathbf{f}(\mathbf{x}).$$

Then the sequence, $\{T^n \mathbf{x}_0\}_{n=1}^{\infty}$ *converges to* $\mathbf{z}$.

Proof: First note that $|T\mathbf{x}_0 - \mathbf{x}_0| = \left|D\mathbf{f}(\mathbf{x}_0)^{-1} \mathbf{f}(\mathbf{x}_0)\right| \leq \left|\left|D\mathbf{f}(\mathbf{x}_0)^{-1}\right|\right| |\mathbf{f}(\mathbf{x}_0)| < \frac{\delta}{4}$. Also on $B(\mathbf{x}_0, \delta) \subseteq B(\mathbf{z}, 2\delta)$ the inequality, (28.36), the chain rule, and Theorem 28.7.5 shows that for $\mathbf{x}, \mathbf{y} \in B(\mathbf{x}_0, \delta)$,

$$|T\mathbf{x} - T\mathbf{y}| \leq \frac{1}{2} |\mathbf{x} - \mathbf{y}|.$$

This follows because $DT\mathbf{x} = I - D\mathbf{f}(\mathbf{x}_0)^{-1} \mathbf{f}(\mathbf{x})$. The conclusion now follows from Lemma 28.7.1. This proves the lemma.

28.7.4 Newton's Method

Theorem 28.7.7 *Suppose* $\mathbf{f} : U \subseteq \mathbb{R}^p \to \mathbb{R}^p$ *is a* C^1 *function and suppose* $\mathbf{f}(\mathbf{z}) = \mathbf{0}$. *Suppose that for all* $\mathbf{x}$ *sufficiently close to* $\mathbf{z}$, *it follows that* $D\mathbf{f}(\mathbf{x})^{-1}$ *exists. Suppose also that*[2]

$$\left|\left|D\mathbf{f}(\mathbf{x}_2)^{-1} - D\mathbf{f}(\mathbf{x}_1)^{-1}\right|\right| \leq K |\mathbf{x}_2 - \mathbf{x}_1|. \tag{28.37}$$

Then there exists $\delta > 0$ *small enough that for all* $\mathbf{x}_1, \mathbf{x}_2 \in B(\mathbf{z}, 2\delta)$

$$\left|\mathbf{x}_1 - \mathbf{x}_2 - D\mathbf{f}(\mathbf{x}_2)^{-1} (\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2))\right| \leq \frac{1}{4} |\mathbf{x}_1 - \mathbf{x}_2|, \tag{28.38}$$

$$|\mathbf{f}(\mathbf{x}_1)| < \frac{1}{4K}. \tag{28.39}$$

Now pick $\mathbf{x}_0 \in B(\mathbf{z}, \delta)$ *also close enough to* $\mathbf{z}$ *such that*

$$\left|\left|D\mathbf{f}(\mathbf{x}_0)^{-1}\right|\right| |\mathbf{f}(\mathbf{x}_0)| < \frac{\delta}{4}.$$

Define

$$T\mathbf{x} \equiv \mathbf{x} - D\mathbf{f}(\mathbf{x})^{-1} \mathbf{f}(\mathbf{x}).$$

Then the sequence, $\{T^n \mathbf{x}_0\}_{n=1}^{\infty}$ *converges to* $\mathbf{z}$.

[2]The following condition as well as the preceeding can be shown to hold if you simply assume $\mathbf{f}$ is a C^2 function and $D\mathbf{f}(\mathbf{z})^{-1}$ exists. This requires the use of the inverse function theorem, one of the major theorems which should be studied in an advanced calculus class.

Proof: The left side of (28.38) equals

$$\begin{aligned} & \left|\mathbf{x}_1-\mathbf{x}_2-D\mathbf{f}\left(\mathbf{x}_2\right)^{-1}\left(D\mathbf{f}\left(\mathbf{x}_2\right)\left(\mathbf{x}_1-\mathbf{x}_2\right)+\mathbf{f}\left(\mathbf{x}_1\right)-\mathbf{f}\left(\mathbf{x}_2\right)-D\mathbf{f}\left(\mathbf{x}_2\right)\left(\mathbf{x}_1-\mathbf{x}_2\right)\right)\right| \\ = & \left|D\mathbf{f}\left(\mathbf{x}_2\right)^{-1}\left(\mathbf{f}\left(\mathbf{x}_1\right)-\mathbf{f}\left(\mathbf{x}_2\right)-D\mathbf{f}\left(\mathbf{x}_2\right)\left(\mathbf{x}_1-\mathbf{x}_2\right)\right)\right| \\ \leq & C\left|\mathbf{f}\left(\mathbf{x}_1\right)-\mathbf{f}\left(\mathbf{x}_2\right)-D\mathbf{f}\left(\mathbf{x}_2\right)\left(\mathbf{x}_1-\mathbf{x}_2\right)\right| \end{aligned}$$

because (28.37) implies $\left|\left|D\mathbf{f}\left(\mathbf{x}\right)^{-1}\right|\right|$ is bounded for $\mathbf{x}\in B\left(\mathbf{z},\delta\right)$. Now use the assumption that $\mathbf{f}$ is C^1 and Proposition 28.7.4 to conclude there exists δ small enough that $||D\mathbf{f}\left(\mathbf{x}\right)-D\mathbf{f}\left(\mathbf{z}\right)||<\frac{1}{8}$ for all $\mathbf{x}\in B\left(\mathbf{z},2\delta\right)$. Then let $\mathbf{x}_1,\mathbf{x}_2\in B\left(\mathbf{z},2\delta\right)$. Define $\mathbf{h}\left(\mathbf{x}\right)\equiv \mathbf{f}\left(\mathbf{x}\right)-\mathbf{f}\left(\mathbf{x}_2\right)-D\mathbf{f}\left(\mathbf{x}_2\right)\left(\mathbf{x}-\mathbf{x}_2\right)$. Then

$$\begin{aligned} ||D\mathbf{h}\left(\mathbf{x}\right)|| &= ||D\mathbf{f}\left(\mathbf{x}\right)-D\mathbf{f}\left(\mathbf{x}_2\right)|| \\ &\leq ||D\mathbf{f}\left(\mathbf{x}\right)-D\mathbf{f}\left(\mathbf{z}\right)||+||D\mathbf{f}\left(\mathbf{z}\right)-D\mathbf{f}\left(\mathbf{x}_2\right)|| \\ &\leq \frac{1}{8}+\frac{1}{8}=\frac{1}{4}. \end{aligned}$$

It follows from Theorem 28.7.5

$$\begin{aligned} \left|\mathbf{h}\left(\mathbf{x}_1\right)-\mathbf{h}\left(\mathbf{x}_2\right)\right| &= \left|\mathbf{f}\left(\mathbf{x}_1\right)-\mathbf{f}\left(\mathbf{x}_2\right)-D\mathbf{f}\left(\mathbf{x}_2\right)\left(\mathbf{x}_1-\mathbf{x}_2\right)\right| \\ &\leq \frac{1}{4}\left|\mathbf{x}_1-\mathbf{x}_2\right|. \end{aligned}$$

This proves (28.38). (28.39) can be satisfied by taking δ still smaller if necessary and using $\mathbf{f}\left(\mathbf{z}\right)=\mathbf{0}$ and the continuity of $\mathbf{f}$.

Now let $\mathbf{x}_0\in B\left(\mathbf{z},\delta\right)$ be as described. Then

$$\left|T\mathbf{x}_0-\mathbf{x}_0\right|=\left|D\mathbf{f}\left(\mathbf{x}_0\right)^{-1}\mathbf{f}\left(\mathbf{x}_0\right)\right|\leq\left|\left|D\mathbf{f}\left(\mathbf{x}_0\right)^{-1}\right|\right|\left|\mathbf{f}\left(\mathbf{x}_0\right)\right|<\frac{\delta}{4}.$$

Letting $\mathbf{x}_1,\mathbf{x}_2\in B\left(\mathbf{x}_0,\delta\right)\subseteq B\left(\mathbf{z},2\delta\right)$,

$$\begin{aligned} \left|T\mathbf{x}_1-T\mathbf{x}_2\right| &= \left|\mathbf{x}_1-D\mathbf{f}\left(\mathbf{x}_1\right)^{-1}\mathbf{f}\left(\mathbf{x}_1\right)-\left(\mathbf{x}_2-D\mathbf{f}\left(\mathbf{x}_2\right)^{-1}\mathbf{f}\left(\mathbf{x}_2\right)\right)\right| \\ &\leq \left|\mathbf{x}_1-\mathbf{x}_2-D\mathbf{f}\left(\mathbf{x}_2\right)^{-1}\left(\mathbf{f}\left(\mathbf{x}_1\right)-\mathbf{f}\left(\mathbf{x}_2\right)\right)\right|+\left|\left(D\mathbf{f}\left(\mathbf{x}_1\right)^{-1}-D\mathbf{f}\left(\mathbf{x}_2\right)^{-1}\right)\mathbf{f}\left(\mathbf{x}_1\right)\right| \\ &\leq \frac{1}{4}\left|\mathbf{x}_1-\mathbf{x}_2\right|+K\left|\mathbf{x}_1-\mathbf{x}_2\right|\left|\mathbf{f}\left(\mathbf{x}_1\right)\right|\leq\frac{1}{2}\left|\mathbf{x}_1-\mathbf{x}_2\right|. \end{aligned}$$

The desired result now follows from Lemma 28.7.1.

28.8 Exercises

1. Suppose $\mathbf{f}:U\to\mathbb{R}^q$ and let $\mathbf{x}\in U$ and $\mathbf{v}$ be a unit vector. Show $D_{\mathbf{v}}\mathbf{f}\left(\mathbf{x}\right)=D\mathbf{f}\left(\mathbf{x}\right)\mathbf{v}$. Recall that
$$D_{\mathbf{v}}\mathbf{f}\left(\mathbf{x}\right)\equiv\lim_{h\to0}\frac{\mathbf{f}\left(\mathbf{x}+t\mathbf{v}\right)-\mathbf{f}\left(\mathbf{x}\right)}{t}.$$

2. Let $f\left(x,y\right)=\begin{cases} xy\sin\left(\frac{1}{x}\right) & \text{if } x\neq0 \\ 0 \text{ if } x=0 \end{cases}$. Find where f is differentiable and compute the derivative at all these points.

3. Let
$$f(x,y) = \begin{cases} x \text{ if } |y| > |x| \\ -x \text{ if } |y| \leq |x| \end{cases}.$$
Show f is continuous at $(0,0)$ and that the partial derivatives exist at $(0,0)$ but the function is not differentiable at $(0,0)$.

4. Let
$$\mathbf{f}(x,y,z) = \begin{pmatrix} x^2 \sin y + z^3 \\ \sin(x+y) + z^3 \cos x \end{pmatrix}.$$
Find $D\mathbf{f}(1,2,3)$.

5. Let
$$\mathbf{f}(x,y,z) = \begin{pmatrix} x \tan y + z^3 \\ \cos(x+y) + z^3 \cos x \end{pmatrix}.$$
Find $D\mathbf{f}(1,2,3)$.

6. Let
$$\mathbf{f}(x,y,z) = \begin{pmatrix} x \sin y + z^3 \\ \sin(x+y) + z^3 \cos x \\ x^5 + y^2 \end{pmatrix}.$$
Find $D\mathbf{f}(x,y,z)$.

7. Let
$$f(x,y) = \begin{cases} \frac{\left(x^2-y^4\right)^2}{(x^2+y^4)^2} \text{ if } (x,y) \neq (0,0) \\ 1 \text{ if } (x,y) = (0,0) \end{cases}.$$
Show that all directional derivatives of f exist at $(0,0)$, and are all equal to zero but the function is not even continuous at $(0,0)$. Therefore, it is not differentiable. Why?

8. In the example of Problem 7 show the partial derivatives exist but are not continuous.

9. A certain building is shaped like the top half of the ellipsoid, $\frac{x^2}{900} + \frac{y^2}{900} + \frac{z^2}{400} = 1$ determined by letting $z \geq 0$. Here dimensions are measured in meters. The building needs to be painted. The paint, when applied is about .005 meters thick. About how many cubic meters of paint will be needed. **Hint:**This is going to replace the numbers, 900 and 400 with slightly larger numbers when the ellipsoid is fattened slightly by the paint. The volume of the top half of the ellipsoid, $x^2/a^2 + y^2/b^2 + z^2/c^2 \leq 1, z \geq 0$ is $(2/3)\,abc$.

10. Show carefully that the usual one variable version of the chain rule is a special case of Theorem 28.4.2.

11. Let $z = f(\mathbf{y}) = \left(y_1^2 + \sin y_2 + \tan y_3\right)$ and $\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_2 \\ x_2^2 - x_1 + x_2 \\ x_2^2 + x_1 + \sin x_2 \end{pmatrix}$. Find $D(f \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2$.

12. Let $z = f(\mathbf{y}) = \left(y_1^2 + \cot y_2 + \sin y_3\right)$ and $\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_4 + x_3 \\ x_2^2 - x_1 + x_2 \\ x_2^2 + x_1 + \sin x_4 \end{pmatrix}$. Find $D(f \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2, 3, 4$.

13. Let $z = f(\mathbf{y}) = \left(y_1^2 + y_2^2 + \sin y_3 + y_4\right)$ and $\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_4 + x_3 \\ x_2^2 - x_1 + x_2 \\ x_2^2 + x_1 + \sin x_4 \\ x_4 + x_2 \end{pmatrix}$. Find $D(f \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z}{\partial x_i}$ for $i = 1, 2, 3, 4$.

14. Let $\mathbf{z} = \mathbf{f}(\mathbf{y}) = \begin{pmatrix} y_1^2 + \sin y_2 + \tan y_3 \\ y_1^2 y_2 + y_3 \end{pmatrix}$ and $\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_2 \\ x_2^2 - x_1 + x_2 \\ x_2^2 + x_1 + \sin x_2 \end{pmatrix}$. Find $D(\mathbf{f} \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z_k}{\partial x_i}$ for $i = 1, 2$ and $k = 1, 2$.

15. Let $\mathbf{z} = \mathbf{f}(\mathbf{y}) = \begin{pmatrix} y_1^2 + \sin y_2 + \tan y_3 \\ y_1^2 y_2 + y_3 \\ \cos\left(y_1^2\right) + y_2^3 y_3 \end{pmatrix}$ and $\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_4 \\ x_2^2 - x_1 + x_3 \\ x_3^2 + x_1 + \sin x_2 \end{pmatrix}$. Find $D(\mathbf{f} \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z_k}{\partial x_i}$ for $i = 1, 2, 3, 4$ and $k = 1, 2, 3$.

16. Let $z = \mathbf{f}(\mathbf{y}) = \begin{pmatrix} y_2^2 + \sin y_1 + \sec y_2 + y_4 \\ y_1^2 y_2 + y_3^3 \\ y_2^3 y_4 + y_1 \\ y_1 + y_2 \end{pmatrix}$ and $\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + 2x_4 \\ x_2^2 - 2x_1 + x_3 \\ x_3^2 + x_1 + \cos x_1 \\ x_2^2 \end{pmatrix}$. Find $D(\mathbf{f} \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z_k}{\partial x_i}$ for $i = 1, 2, 3, 4$ and $k = 1, 2, 3, 4$.

17. Let $\mathbf{f}(\mathbf{y}) = \begin{pmatrix} y_1^2 + \sin y_2 + \tan y_3 \\ y_1^2 y_2 + y_3 \\ \cos\left(y_1^2\right) + y_2^3 y_3 \end{pmatrix}$ and $\mathbf{y} = \mathbf{g}(\mathbf{x}) \equiv \begin{pmatrix} x_1 + x_4 \\ x_2^2 - x_1 + x_3 \\ x_3^2 + x_1 + \sin x_2 \end{pmatrix}$. Find $D(\mathbf{f} \circ \mathbf{g})(\mathbf{x})$. Use to write $\frac{\partial z_k}{\partial x_i}$ for $i = 1, 2, 3, 4$ and $k = 1, 2, 3$.

18. Suppose $\mathbf{r}_1(t) = (\cos t, \sin t, t)$, $\mathbf{r}_2(t) = (t, 2t, 1)$, and $\mathbf{r}_3(t) = (1, t, 1)$. Find the rate of change with respect to t of the volume of the parallelepiped determined by these three vectors when $t = 1$.

19. A trash compacter is compacting a rectangular block of trash. The width is changing at the rate of -1 inches per second, the length is changing at the rate of -2 inches per second and the height is changing at the rate of -3 inches per second. How fast is the volume changing when the length is 20, the height is 10, and the width is 10.

20. A trash compacter is compacting a rectangular block of trash. The width is changing at the rate of -2 inches per second, the length is changing at the rate of -1 inches per second and the height is changing at the rate of -4 inches per second. How fast is the surface area changing when the length is 20, the height is 10, and the width is 10.

21. The ideal gas law is $PV = kT$ where k is a constant which depends on the number of moles and on the gas being considered. If V is changing at the rate of 2 cubic cm. per second and T is changing at the rate of 3 degrees Kelvin per second, how fast is the pressure changing when $T = 300$ and V equals 400 cubic cm.?

22. Let S denote a level surface of the form $f(x_1, x_2, x_3) = C$. Suppose now that $\mathbf{r}(t)$ is a space curve which lies in this level surface. Thus $f(r_1(t), r_2(t), r_3(t))$. Show using the chain rule that $Df(r_1(t), r_2(t), r_3(t))(r_1'(t), r_2'(t), r_3'(t))^T = 0$. Note that $Df(x_1, x_2, x_3) = (f_{x_1}, f_{x_2}, f_{x_3})$. This is denoted by $\nabla f(x_1, x_2, x_3) = (f_{x_1}, f_{x_2}, f_{x_3})^T$. This 3×1 matrix or column vector is called the gradient vector. Argue that

$$\nabla f(r_1(t), r_2(t), r_3(t)) \cdot (r_1'(t), r_2'(t), r_3'(t))^T = 0.$$

What geometric fact have you just established?

23. Suppose $\mathbf{f}$ is a C^1 function which maps U, an open subset of $\mathbb{R}^n$ one to one and onto V, an open set in $\mathbb{R}^m$ such that the inverse map, $\mathbf{f}^{-1}$ is also C^1. What must be true of m and n? Why? **Hint:** Consider Example 28.4.5 on Page 654.

The Gradient

29.0.1 Outcomes

1. Interpret the gradient of a function as a normal to a level curve or a level surface.
2. Find the normal line and tangent plane to a smooth surface at a given point.
3. Find the angles between curves and surfaces.

Here we review the concept of the gradient. This has already been considered in the special case of a C^1 function. However, you do not need so much to talk of the gradient.

29.1 Fundamental Properties

Let $f : U \to \mathbb{R}$ where U is an open subset of $\mathbb{R}^n$ and suppose f is differentiable on U. Thus if $\mathbf{x} \in U$,

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \sum_{j=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_i} v_i + o(\mathbf{v}). \tag{29.1}$$

Recall Proposition 27.3.6, a more general version of which is stated here for convenience. It is more general because we only assume f is differentiable, not C^1.

Proposition 29.1.1 *If f is differentiable at $\mathbf{x}$ and for $\mathbf{v}$ a unit vector,*

$$D_{\mathbf{v}} f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v}.$$

Proof:

$$\begin{aligned}
\frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t} &= \frac{1}{t}\left(f(\mathbf{x}) + \sum_{j=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_i} t v_i + o(t\mathbf{v}) - f(\mathbf{x}) \right) \\
&= \frac{1}{t}\left(\sum_{j=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_i} t v_i + o(t\mathbf{v}) \right) \\
&= \sum_{j=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_i} v_i + \frac{o(t\mathbf{v})}{t}.
\end{aligned}$$

Now $\lim_{t\to 0} \frac{o(t\mathbf{v})}{t} = 0$ and so

$$D_{\mathbf{v}} f(\mathbf{x}) = \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t} = \sum_{j=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_i} v_i = \nabla f(\mathbf{x}) \cdot \mathbf{v}$$

as claimed.

Definition 29.1.2 *When f is differentiable, define $\nabla f(\mathbf{x}) \equiv \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \cdots, \frac{\partial f}{\partial x_n}(\mathbf{x})\right)^T$ just as was done in the special case where f is C^1. As before, this vector is called the gradient vector.*

This defines the gradient for a differentiable scalar valued function. There are ways to define the gradient for vector valued functions but this will not be attempted in this book.

It follows immediately from (29.1) that

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \mathbf{v} + o(\mathbf{v}) \tag{29.2}$$

An important aspect of the gradient is its relation with the directional derivative. From (29.2), for $\mathbf{v}$ a unit vector,

$$\begin{aligned} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t} &= \nabla f(\mathbf{x}) \cdot \mathbf{v} + \frac{o(t\mathbf{v})}{t} \\ &= \nabla f(\mathbf{x}) \cdot \mathbf{v} + \frac{o(t)}{t}. \end{aligned}$$

Therefore, taking $t \to 0$,

$$D_{\mathbf{v}} f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v}. \tag{29.3}$$

Example 29.1.3 *Let $f(x,y,z) = x^2 + \sin(xy) + z$. Find $D_{\mathbf{v}} f(1,0,1)$ where $\mathbf{v} = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$.*

Note this vector which is given is already a unit vector. Therefore, from the above, it is only necessary to find $\nabla f(1,0,1)$ and take the dot product. $\nabla f(x,y,z) = (2x, x\cos(xy), 1)$. Therefore, $\nabla f(1,0,1) = (2,1,1)$. Therefore, the directional derivative is $(2,1,1) \cdot \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right) = \frac{4}{3}\sqrt{3}$.

Because of (29.3) it is easy to find the largest possible directional derivative and the smallest possible directional derivative. That which follows is a more algebraic treatment of an earlier result with the trigonometry removed.

Proposition 29.1.4 *Let $f : U \to \mathbb{R}$ be a differentiable function and let $\mathbf{x} \in U$. Then*

$$\max\{D_{\mathbf{v}} f(\mathbf{x}) : |\mathbf{v}| = 1\} = |\nabla f(x)| \tag{29.4}$$

and

$$\min\{D_{\mathbf{v}} f(\mathbf{x}) : |\mathbf{v}| = 1\} = -|\nabla f(x)|. \tag{29.5}$$

Furthermore, the maximum in (29.4) occurs when $\mathbf{v} = \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$ and the minimum in (29.5) occurs when $\mathbf{v} = -\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$.

Proof: From (29.3) and the Cauchy Schwarz inequality,

$$|D_{\mathbf{v}} f(\mathbf{x})| \le |\nabla f(\mathbf{x})|$$

and so for any choice of $\mathbf{v}$ with $|\mathbf{v}| = 1$,

$$-|\nabla f(\mathbf{x})| \le D_{\mathbf{v}} f(\mathbf{x}) \le |\nabla f(\mathbf{x})|.$$

The proposition is proved by noting that if $\mathbf{v} = -\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$, then

$$\begin{aligned} D_{\mathbf{v}} f(\mathbf{x}) &= \nabla f(\mathbf{x}) \cdot (-\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|) \\ &= -|\nabla f(\mathbf{x})|^2 / |\nabla f(\mathbf{x})| = -|\nabla f(\mathbf{x})| \end{aligned}$$

while if $\mathbf{v} = \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$, then

$$\begin{aligned} D_{\mathbf{v}} f(\mathbf{x}) &= \nabla f(\mathbf{x}) \cdot (\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|) \\ &= |\nabla f(\mathbf{x})|^2 / |\nabla f(\mathbf{x})| = |\nabla f(\mathbf{x})|. \end{aligned}$$

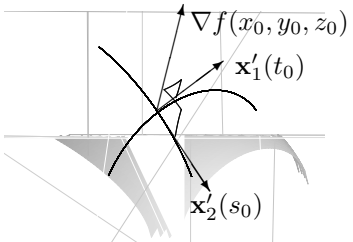
The conclusion of the above proposition is important in many physical models. For example, consider some material which is at various temperatures depending on location. Because it has cool places and hot places, it is expected that the heat will flow from the hot places to the cool places. Consider a small surface having a unit normal, $\mathbf{n}$. Thus $\mathbf{n}$ is a normal to this surface and has unit length. If it is desired to find the rate in calories per second at which heat crosses this little surface in the direction of $\mathbf{n}$ it is defined as $\mathbf{J} \cdot \mathbf{n}A$ where A is the area of the surface and $\mathbf{J}$ is called the heat flux. It is reasonable to suppose the rate at which heat flows across this surface will be largest when $\mathbf{n}$ is in the direction of greatest rate of decrease of the temperature. In other words, heat flows most readily in the direction which involves the maximum rate of decrease in temperature. This expectation will be realized by taking $\mathbf{J} = -K\nabla u$ where K is a positive scalar function which can depend on a variety of things. The above relation between the heat flux and ∇u is usually called the Fourier heat conduction law and the constant, K is known as the coefficient of thermal conductivity. It is a material property, different for iron than for aluminum. In most applications, K is considered to be a constant but this is wrong. Experiments show this scalar should depend on temperature. Nevertheless, things get very difficult if this dependence is allowed. The constant can depend on position in the material or even on time.

An identical relationship is usually postulated for the flow of a diffusing species. In this problem, something like a pollutant diffuses. It may be an insecticide in ground water for example. Like heat, it tries to move from areas of high concentration toward areas of low concentration. In this case $\mathbf{J} = -K\nabla c$ where c is the concentration of the diffusing species. When applied to diffusion, this relationship is known as Fick's law. Mathematically, it is indistinguishable from the problem of heat flow.

Note the importance of the gradient in formulating these models.

29.2 Tangent Planes

The gradient has fundamental geometric significance illustrated by the following picture.



In this picture, the surface is a piece of a level surface of a function of three variables, $f(x, y, z)$. Thus the surface is defined by $f(x, y, z) = c$ or more completely as $\{(x, y, z) : f(x, y, z) = c\}$. For example, if $f(x, y, z) = x^2 + y^2 + z^2$, this would be a piece of a sphere. There are two smooth curves in this picture which lie in the surface having parameterizations, $\mathbf{x}_1(t) = (x_1(t), y_1(t), z_1(t))$ and $\mathbf{x}_2(s) = (x_2(s), y_2(s), z_2(s))$ which intersect at the point, (x_0, y_0, z_0) on this surface[1]. This intersection occurs when $t = t_0$ and $s = s_0$. Since the points, $\mathbf{x}_1(t)$ for t in an interval lie in the level surface, it follows

[1]Do there exist any smooth curves which lie in the level surface of f and pass through the point (x_0, y_0, z_0)? It turns out there do if $\nabla f(x_0, y_0, z_0) \neq \mathbf{0}$ and if the function, f, is C^1. However, this is a

$$f(x_1(t), y_1(t), z_1(t)) = c$$

for all t in some interval. Therefore, taking the derivative of both sides and using the chain rule on the left,

$$\frac{\partial f}{\partial x}(x_1(t), y_1(t), z_1(t)) x_1'(t) +$$

$$\frac{\partial f}{\partial y}(x_1(t), y_1(t), z_1(t)) y_1'(t) + \frac{\partial f}{\partial z}(x_1(t), y_1(t), z_1(t)) z_1'(t) = 0.$$

In terms of the gradient, this merely states

$$\nabla f(x_1(t), y_1(t), z_1(t)) \cdot \mathbf{x}_1'(t) = 0.$$

Similarly,

$$\nabla f(x_2(s), y_2(s), z_2(s)) \cdot \mathbf{x}_2'(s) = 0.$$

Letting $s = s_0$ and $t = t_0$, it follows

$$\nabla f(x_0, y_0, z_0) \cdot \mathbf{x}_1'(t_0) = 0,\ \nabla f(x_0, y_0, z_0) \cdot \mathbf{x}_2'(s_0) = 0.$$

It follows $\nabla f(x_0, y_0, z_0)$ is perpendicular to both the direction vectors of the two indicated curves shown. Surely if things are as they should be, these two direction vectors would determine a plane which deserves to be called the tangent plane to the level surface of f at the point (x_0, y_0, z_0) and that $\nabla f(x_0, y_0, z_0)$ is perpendicular to this tangent plane at the point, (x_0, y_0, z_0).

Example 29.2.1 *Find the equation of the tangent plane to the level surface, $f(x, y, z) = 6$ of the function, $f(x, y, z) = x^2 + 2y^2 + 3z^2$ at the point $(1, 1, 1)$.*

First note that $(1, 1, 1)$ is a point on this level surface. To find the desired plane it suffices to find the normal vector to the proposed plane. But $\nabla f(x, y, z) = (2x, 4y, 6z)$ and so $\nabla f(1, 1, 1) = (2, 4, 6)$. Therefore, from this problem, the equation of the plane is

$$(2, 4, 6) \cdot (x - 1, y - 1, z - 1) = 0$$

or in other words,

$$2x - 12 + 4y + 6z = 0.$$

Example 29.2.2 *The point, $\left(\sqrt{3}, 1, 4\right)$ is on both the surfaces, $z = x^2 + y^2$ and $z = 8 - \left(x^2 + y^2\right)$. Find the cosine of the angle between the two tangent planes at this point.*

Recall this is the same as the angle between two normal vectors. Of course there is some ambiguity here because if $\mathbf{n}$ is a normal vector, then so is $-\mathbf{n}$ and replacing $\mathbf{n}$ with $-\mathbf{n}$ in the formula for the cosine of the angle will change the sign. We agree to look for the acute angle and its cosine rather than the optuse angle. The normals are $\left(2\sqrt{3}, 2, -1\right)$ and $\left(2\sqrt{3}, 2, 1\right)$. Therefore, the cosine of the angle desired is

$$\frac{\left(2\sqrt{3}\right)^2 + 4 - 1}{17} = \frac{15}{17}.$$

consequence of the implicit function theorem, one of the greatest theorems in all mathematics and a topic for an advanced calculus class.

Example 29.2.3 *The point,* $\left(1,\sqrt{3},4\right)$ *is on the surface,* $z=x^2+y^2$*. Find the line perpendicular to the surface at this point.*

All we need is the direction vector of this line. The surface is the level surface, $x^2+y^2-z=0$. The normal to this surface is given by the gradient at this point. Thus the line desired is

$$\left(1,\sqrt{3},4\right)+t\left(2,2\sqrt{3},-1\right).$$

29.3 Exercises

1. Find the gradients of $f=$
   (a) x^2y+z^3 at $(1,1,2)$
   (b) $z\sin\left(x^2y\right)+2^{x+y}$ at $(1,1,0)$
   (c) $u\ln\left(x+y+z^2+w\right)$ at $(x,y,z,w,u)=(1,1,1,1,2)$

2. Find the directional derivatives of f at the indicated point in the direction, $\left(\frac{1}{2},\frac{1}{2},\frac{1}{\sqrt{2}}\right)$.
   (a) x^2y+z^3 at $(1,1,1)$
   (b) $z\sin\left(x^2y\right)+2^{x+y}$ at $(1,1,2)$
   (c) $xy+z^2+w$ at $(1,2,3)$

3. Find the tangent plane to the indicated level surface at the indicated point.
   (a) $x^2y+z^3=2$ at $(1,1,1)$
   (b) $z\sin\left(x^2y\right)+2^{x+y}=2\sin 1+4$ at $(1,1,2)$
   (c) $\cos(x)+z\sin(x+y)=1$ at $\left(-\pi,\frac{3\pi}{2},2\right)$

4. Explain why the displacement vector of an object moving in $\mathbb{R}^3$ is always perpendicular to the velocity vector if the object is always at a fixed distance from a given point.

5. The point $\left(1,1,\sqrt{2}\right)$ is a point on the level surface, $x^2+y^2+z^2=4$. Find the line perpendicular to the surface at this point.

6. The point $\left(1,1,\sqrt{2}\right)$ is a point on the level surface, $x^2+y^2+z^2=4$ and the level surface, $y^2+2z^2=5$. Find the angle between the two tangent planes at this point.

7. The level surfaces $x^2+y^2+z^2=4$ and $z+x^2+y^2=4$ have the point $\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2},1\right)$ in the curve formed by the intersection of these surfaces. Find a direction vector for this curve at this point. **Hint:** Recall the gradients of the two surfaces are perpendicular to the corresponding surfaces at this point. A direction vector for the desired curve should be perpendicular to both of these gradients.

8. In a slightly more general setting, suppose $f_1(x,y,z)=0$ and $f_2(x,y,z)=0$ are two level surfaces which intersect in a curve which has parameterization, $(x(t),y(t),z(t))$. Find a differential equation for this curve.

Optimization

30.0.1 Outcomes

1. Define what is meant by a local extreme point.
2. Find candidates for local extrema using the gradient.
3. Find the local extreme values and saddle points of a C^2 function.
4. Use the second derivative test to identify the nature of a singluar point.
5. Find the extreme values of a function defined on a closed and bounded region.
6. Solve word problems involving maximum and minimum values.
7. Use the method of Lagrange to determine the extreme values of a function subject to a constraint.
8. Solve word problems using the method of Lagrange multipliers.

Suppose $f : D(f) \to \mathbb{R}$ where $D(f) \subseteq \mathbb{R}^n$.

30.1 Local Extrema

Definition 30.1.1 *A point $\mathbf{x} \in D(f)$ is called a local minimum if $f(\mathbf{x}) \leq f(\mathbf{y})$ for all $\mathbf{y} \in D(f)$ sufficiently close to $\mathbf{x}$. A point $\mathbf{x} \in D(f)$ is called a local maximum if $f(\mathbf{x}) \geq f(\mathbf{y})$ for all $\mathbf{y} \in D(f)$ sufficiently close to $\mathbf{x}$. A local extremum is a point of $D(f)$ which is either a local minimum or a local maximum. The plural for extremum is extrema.*

Procedure 30.1.2 *To find candidates for local extrema which are interior points of $D(f)$ where f is a C^1 function, you simply identify those points where ∇f equals the zero vector. To justify this, note that the graph of f is the level surface*

$$F(\mathbf{x},z) \equiv f(\mathbf{x}) - z = 0$$

and the local extrema at such interior points must have horizontal tangent planes. Therefore, a normal vector at such points must be a multiple of $(0,\cdots,0,1)$. Thus ∇F at such points must be a multiple of this vector. That is, if $\mathbf{x}$ is such a point,

$$k(0,\cdots,0,1) = (f_{x_1}(\mathbf{x}),\cdots,f_{x_n}(\mathbf{x}),-1).$$

Thus $\nabla f(\mathbf{x}) = \mathbf{0}$.

Definition 30.1.3 *A singular point for f is a point* $\mathbf{x}$ *where* $\nabla f(\mathbf{x}) = \mathbf{0}$.

Example 30.1.4 *Find the local extrema for the function,* $f(x,y) \equiv xy - x - y$ *for* $x, y > 0$.

Note that here $D(f)$ is an open set and so every point is an interior point. Where is the gradient equal to zero?

$$f_x = y - 1 = 0,\ f_y = x - 1 = 0$$

and so there is exactly one candidate for a local extrema, $(1,1)$.

Example 30.1.5 *Find the volume of the smallest tetrahedron made up of the coordinate planes in the first octant and a plane which is tangent to the sphere* $x^2 + y^2 + z^2 = 4$.

The normal to the sphere at a point, (x_0, y_0, z_0) on a point of the sphere is $\left(x_0, y_0, \sqrt{4 - x_0^2 - y_0^2}\right)$ and so the equation of the tangent plane at this point is

$$x_0(x - x_0) + y_0(y - y_0) + \sqrt{4 - x_0^2 - y_0^2}\left(z - \sqrt{4 - x_0^2 - y_0^2}\right) = 0$$

When $x = y = 0$,

$$z = \frac{4}{\sqrt{(4 - x_0^2 - y_0^2)}}$$

When $z = 0 = y$,

$$x = \frac{4}{x_0},$$

and when $z = x = 0$,

$$y = \frac{4}{y_0}.$$

Therefore, the function to minimize is

$$f(x,y) = \frac{1}{6}\frac{64}{xy\sqrt{(4 - x^2 - y^2)}}$$

This is because in beginning calculus it was shown that the volume of a pyramid is 1/3 the area of the base times the height. Therefore, you simply need to find the gradient of this and set it equal to zero. Thus upon taking the partial derivatives, you need to have

$$\frac{-4 + 2x^2 + y^2}{x^2 y\left(-4 + x^2 + y^2\right)\sqrt{(4 - x^2 - y^2)}} = 0,$$

and

$$\frac{-4 + x^2 + 2y^2}{xy^2\left(-4 + x^2 + y^2\right)\sqrt{(4 - x^2 - y^2)}} = 0.$$

Therefore, $x^2 + 2y^2 = 4$ and $2x^2 + y^2 = 4$. Thus $x = y$ and so $x = y = \frac{2}{\sqrt{3}}$. It follows from the equation for z that $z = \frac{2}{\sqrt{3}}$ also.

Example 30.1.6 *An open box is to contain 32 cubic feet. Find the dimensions which will result in the least surface area.*

Let the height of the box be z and the length and width be x and y respectively. Then $xyz = 32$ and so $z = 32/xy$. The total area is $xy + 2xz + 2yz$ and so in terms of the two variables, x and y, the area is

$$A = xy + \frac{64}{y} + \frac{64}{x}$$

To find best dimensions you note these must result in a local minimum.

$$A_x = \frac{yx^2 - 64}{x^2} = 0,\ A_y = \frac{xy^2 - 64}{y^2}.$$

Therefore, $yx^2 - 64 = 0$ and $xy^2 - 64 = 0$ so $xy^2 = yx^2$. For sure the answer excludes the case where any of the variables equals zero. Therefore, $x = y$ and so $x = 4 = y$. Then $z = 2$ from the requirement that $xyz = 32$.

30.2 The Second Derivative Test

There is a version of the second derivative test in the case that the function and its first and second partial derivatives are all continuous. A discussion of its proof is given in Section 30.3.

Theorem 30.2.1 *Let $f : U \to \mathbb{R}$ for U an open set in $\mathbb{R}^n$ and let f be a C^2 function and suppose that at some $\mathbf{x} \in U$, $\nabla f(\mathbf{x}) = \mathbf{0}$. Also let μ and λ be respectively, the largest and smallest eigenvalues of the matrix, H. If $\lambda > 0$ then f has a local minimum at $\mathbf{x}$. If $\mu < 0$ then f has a local maximum at $\mathbf{x}$. If either λ or μ equals zero, the test fails. If $\lambda < 0$ and $\mu > 0$ there exists a direction in which when f is evaluated on the line through the critical point having this direction, the resulting function of one variable has a local minimum and there exists a direction in which when f is evaluated on the line through the critical point having this direction, the resulting function of one variable has a local maximum. This last case is called a saddle point.*

Example 30.2.2 *Let $f(x,y) = 2x^4 - 4x^3 + 14x^2 + 12yx^2 - 12yx - 12x + 2y^2 + 4y + 2$. Find the critical points and determine whether they are local minimums, local maximums, or saddle points.*

$f_x(x,y) = 8x^3 - 12x^2 + 28x + 24yx - 12y - 12$ and $f_y(x,y) = 12x^2 - 12x + 4y + 4$. The points at which both f_x and f_y equal zero are $\left(\frac{1}{2}, -\frac{1}{4}\right)$, $(0,-1)$, and $(1,-1)$.

The Hessian matrix is

$$\begin{pmatrix} 24x^2 + 28 + 24y - 24x & 24x - 12 \\ 24x - 12 & 4 \end{pmatrix}.$$

and the thing to determine is the sign of its eigenvalues evaluated at the critical points.

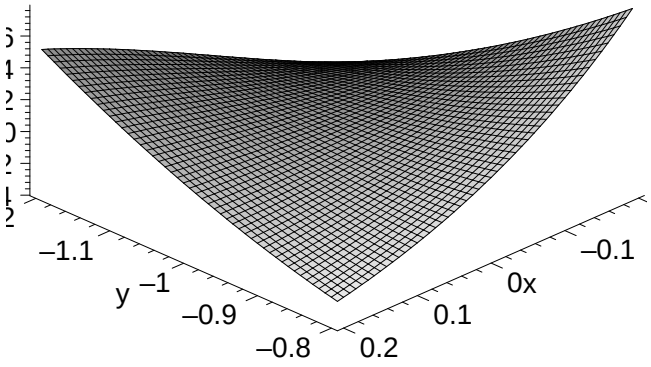
First consider the point $\left(\frac{1}{2}, -\frac{1}{4}\right)$. This matrix is $\begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}$ and its eigenvalues are $16, 4$ showing that this is a local minimum.

Next consider $(0,-1)$ at this point the Hessian matrix is $\begin{pmatrix} 4 & -12 \\ -12 & 4 \end{pmatrix}$ and the eigenvalues are $16, -8$. Therefore, this point is a saddle point.

Finally consider the point $(1,-1)$. At this point the Hessian is $\begin{pmatrix} 4 & 12 \\ 12 & 4 \end{pmatrix}$ and the eigenvalues are $16, -8$ so this point is also a saddle point.

The geometric significance of a saddle point was explained above. In one direction it looks like a local minimum while in another it looks like a local maximum. In fact, they

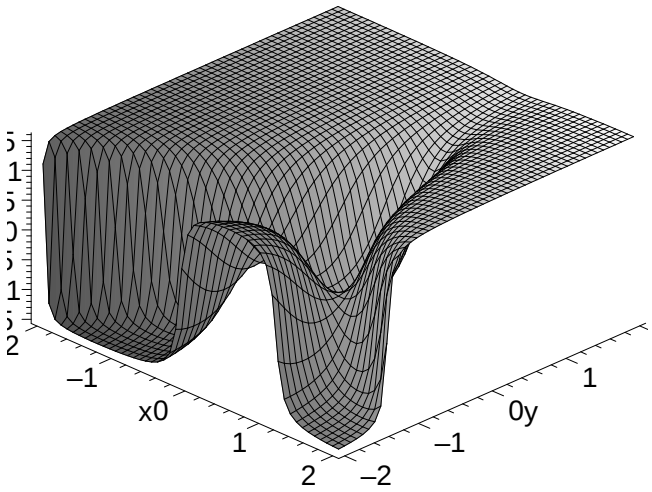
do look like a saddle. Here is a picture of the graph of the above function near the saddle point, $(0,-1)$.



You see it is a lot like the place where you sit on a saddle. If you want to get a better picture, you could graph instead

$$f(x,y) = \arctan\left(2x^4 - 4x^3 + 14x^2 + 12yx^2 - 12yx - 12x + 2y^2 + 4y + 2\right).$$

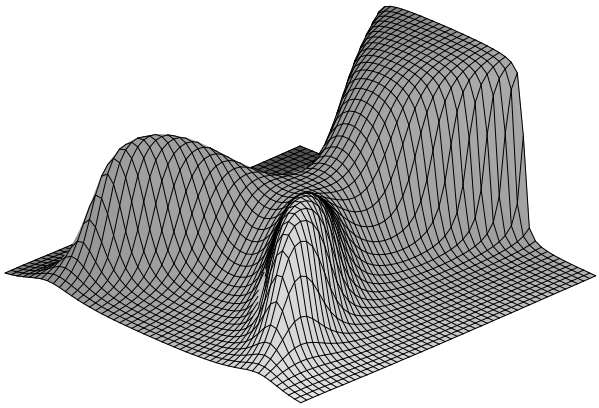
Since arctan is a strictly increasing function, it preserves all the information about whether the given function is increasing or decreasing in certain directions. Below is a graph of this function which illustrates the behavior near the point $(1,-1)$.



Or course sometimes the second derivative test is inadequate to determine what is going on. This should be no surprise since this was the case even for a function of one variable. For a function of two variables, a nice example is the Monkey saddle.

Example 30.2.3 *Suppose* $f(x,y) = \arctan\left(6xy^2 - 2x^3 - 3y^4\right)$. *Show* $(0,0)$ *is a critical point for which the second derivative test gives no information.*

Before doing anything it might be interesting to look at the graph of this function of two variables plotted using Maple.



This picture should indicate why this is called a monkey saddle. It is because the monkey can sit in the saddle and have a place for his tail. Now to see $(0,0)$ is a critical point, note that

$$\frac{\partial\left(\arctan\left(g\left(x,y\right)\right)\right)}{\partial x}=\frac{1}{1+g\left(x,y\right)^{2}}g_{x}\left(x,y\right)$$

and that a similar formula holds for the partial derivative with respect to y. Therefore, it suffices to verify that for

$$g\left(x,y\right)=6xy^{2}-2x^{3}-3y^{4}$$

$g_{x}\left(0,0\right)=g_{y}\left(0,0\right)=0.$

$$g_{x}\left(x,y\right)=6y^{2}-6x^{2},\ \ g_{y}\left(x,y\right)=12xy-12y^{3}$$

and clearly $(0,0)$ is a critical point. So are $(1,1)$ and $(1,-1)$. Now $g_{xx}\left(0,0\right)=0$ and so does $g_{xy}\left(0,0\right)$ and $g_{yy}\left(0,0\right)$. This implies f_{xx},f_{xy},f_{yy} are all equal to zero at $(0,0)$ also. (Why?) Therefore, the Hessian matrix is the zero matrix and clearly has only the zero eigenvalue. Therefore, the second derivative test is totally useless at this point.

However, suppose you took $x=t$ and $y=t$ and evaluated this function on this line. This reduces to $h\left(t\right)=f\left(t,t\right)=\arctan(4t^{3}-t^{4})$, which is strictly increasing near $t=0$. This shows the critical point, $(0,0)$ of f is neither a local max. nor a local min. Next let $x=0$ and $y=t$. Then $p\left(t\right)\equiv f\left(0,t\right)=-3t^{4}$. Therefore, along the line, $(0,t)$, f has a local maximum at $(0,0)$.

30.3 Proof Of The Second Derivative Test

Definition 30.3.1 *The matrix,* $\left(\frac{\partial^{2}f}{\partial x_{i}\partial x_{j}}\left(\mathbf{x}\right)\right)$ *is called the Hessian matrix, denoted by* $H\left(\mathbf{x}\right)$.

Now recall the Taylor formula with the Lagrange form of the remainder. See Theorem 11.1.1 on page 299 for a proof.

Theorem 30.3.2 *Let* $h:(-\delta,1+\delta)\to\mathbb{R}$ *have* $m+1$ *derivatives. Then there exists* $t\in(0,1)$ *such that*

$$h\left(1\right)=h\left(0\right)+\sum_{k=1}^{m}\frac{h^{(k)}\left(0\right)}{k!}+\frac{h^{(m+1)}\left(t\right)}{(m+1)!}.$$

Now let $f : U \to \mathbb{R}$ where U is an open subset of $\mathbb{R}^n$. Suppose $f \in C^2(U)$. Let $\mathbf{x} \in U$ and let $r > 0$ be such that

$$B(\mathbf{x},r) \subseteq U.$$

Then for $||\mathbf{v}|| < r$ consider

$$f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x}) \equiv h(t)$$

for $t \in [0,1]$. Then from Taylor's theorem for the case where $m = 2$ and the chain rule, using the repeated index summation convention and the chain rule,

$$h'(t) = \frac{\partial f}{\partial x_i}(\mathbf{x}+t\mathbf{v})\, v_i,\ h''(t) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}+t\mathbf{v})\, v_i v_j.$$

Thus

$$h''(t) = \mathbf{v}^T H(\mathbf{x}+t\mathbf{v})\,\mathbf{v}.$$

From Theorem 30.3.2 there exists $t \in (0,1)$ such that

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \frac{\partial f}{\partial x_i}(\mathbf{x})\, v_i + \frac{1}{2}\mathbf{v}^T H(\mathbf{x}+t\mathbf{v})\,\mathbf{v}$$

By the continuity of the second partial derivative

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x})\cdot\mathbf{v} + \frac{1}{2}\mathbf{v}^T H(\mathbf{x})\,\mathbf{v} + \frac{1}{2}\left(\mathbf{v}^T\left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right) \tag{30.1}$$

where the last term satisfies

$$\lim_{|\mathbf{v}|\to 0} \frac{1}{2}\frac{\left(\mathbf{v}^T\left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right)}{|\mathbf{v}|^2} = 0 \tag{30.2}$$

because of the continuity of the entries of $H(\mathbf{x})$.

Theorem 30.3.3 *Suppose* $\mathbf{x}$ *is a critical point for* f. *That is, suppose* $\frac{\partial f}{\partial x_i}(\mathbf{x}) = 0$ *for each* i. *Then if* $H(\mathbf{x})$ *has all positive eigenvalues,* $\mathbf{x}$ *is a local minimum. If* $H(\mathbf{x})$ *has all negative eigenvalues, then* $\mathbf{x}$ *is a local maximum. If* $H(\mathbf{x})$ *has a positive eigenvalue, then there exists a direction in which* f *has a local minimum at* $\mathbf{x}$, *while if* $H(\mathbf{x})$ *has a negative eigenvalue, there exists a direction in which* f *has a local maximum at* $\mathbf{x}$.

Proof: Since $\nabla f(\mathbf{x}) = 0$, formula (30.1) implies

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \frac{1}{2}\mathbf{v}^T H(\mathbf{x})\,\mathbf{v} + \frac{1}{2}\left(\mathbf{v}^T\left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right) \tag{30.3}$$

and by continuity of the second derivatives, these mixed second derivatives are equal and so $H(\mathbf{x})$ is a symmetric matrix . Thus, by Corollary 21.4.8 on Page 524 $H(\mathbf{x})$ has all real eigenvalues. Suppose first that $H(\mathbf{x})$ has all positive eigenvalues and that all are larger than $\delta^2 > 0$. Then by this corollary, $H(\mathbf{x})$ has an orthonormal basis of eigenvectors, $\{\mathbf{v}_i\}_{i=1}^n$ and so if $\mathbf{u}$ is an arbitrary vector, there exist scalars, u_i such that $\mathbf{u} = \sum_{j=1}^n u_j \mathbf{v}_j$. Taking the dot product of both sides with $\mathbf{v}_j$ it follows $u_j = \mathbf{u}\cdot\mathbf{v}_j$. Thus

$$\begin{aligned}\mathbf{u}^T H(\mathbf{x})\,\mathbf{u} &= \left(\sum_{k=1}^n u_k \mathbf{v}_k^T\right) H(\mathbf{x}) \left(\sum_{j=1}^n u_j \mathbf{v}_j\right) \\ &= \sum_{k,j} u_k \mathbf{v}_k^T H(\mathbf{x})\,\mathbf{v}_j u_j\end{aligned}$$

$$= \sum_{j=1}^{n} u_j^2 \lambda_j \geq \delta^2 \sum_{j=1}^{n} u_j^2 = \delta^2 |\mathbf{u}|^2 .$$

From (30.3) and (30.2), if $\mathbf{v}$ is small enough,

$$f(\mathbf{x}+\mathbf{v}) \geq f(\mathbf{x}) + \frac{1}{2}\delta^2 |\mathbf{v}|^2 - \frac{1}{4}\delta^2 |\mathbf{v}|^2 = f(\mathbf{x}) + \frac{\delta^2}{4} |\mathbf{v}|^2 .$$

This shows the first claim of the theorem. The second claim follows from similar reasoning. Suppose $H(\mathbf{x})$ has a positive eigenvalue λ^2. Then let $\mathbf{v}$ be an eigenvector for this eigenvalue. Then from (30.3), replacing $\mathbf{v}$ with $s\mathbf{v}$ and letting t depend on s,

$$f(\mathbf{x}+s\mathbf{v}) = f(\mathbf{x}) + \frac{1}{2}s^2 \mathbf{v}^T H(\mathbf{x})\mathbf{v} +$$

$$\frac{1}{2}s^2 \left(\mathbf{v}^T \left(H(\mathbf{x}+ts\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right)$$

which implies

$$\begin{aligned} f(\mathbf{x}+s\mathbf{v}) &= f(\mathbf{x}) + \frac{1}{2}s^2\lambda^2 |\mathbf{v}|^2 + \frac{1}{2}s^2 \left(\mathbf{v}^T \left(H(\mathbf{x}+ts\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right) \\ &\geq f(\mathbf{x}) + \frac{1}{4}s^2\lambda^2 |\mathbf{v}|^2 \end{aligned}$$

whenever s is small enough. Thus in the direction $\mathbf{v}$ the function has a local minimum at $\mathbf{x}$. The assertion about the local maximum in some direction follows similarly. This proves the theorem.

30.4 Exercises

1. Use the second derivative test on the critical points $(1,1)$, and $(1,-1)$ for Example 30.2.3.

2. If $H = H^T$ and $H\mathbf{x} = \lambda\mathbf{x}$ while $H\mathbf{x} = \mu\mathbf{x}$ for $\lambda \neq \mu$, show $\mathbf{x} \cdot \mathbf{y} = 0$.

3. Show the points $\left(\frac{1}{2}, -\frac{21}{4}\right), (0,-4)$, and $(1,-4)$ are critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

   $f(x,y) = -x^4 + 2x^3 + 39x^2 + 10yx^2 - 10yx - 40x - y^2 - 8y - 16.$

   Answer:

   The Hessian matrix is

   $$\begin{pmatrix} -12x^2 + 78 + 20y + 12x & 20x - 10 \\ 20x - 10 & -2 \end{pmatrix}$$

   The eigenvalues must be checked at the critical points. First consider the point $\left(\frac{1}{2}, -\frac{21}{4}\right)$. At this point, the Hessian is

   $$\begin{pmatrix} -24 & 0 \\ 0 & -2 \end{pmatrix}$$

   and its eigenvalues are $-24, -2$, both negative. Therefore, the function has a local maximum at this point.

Next consider $(0,-4)$. At this point the Hessian matrix is

$$\begin{pmatrix} -2 & -10 \\ -10 & -2 \end{pmatrix}$$

and the eigenvalues are $8,-12$ so the function has a saddle point.

Finally consider the point $(1,-4)$. The Hessian equals

$$\begin{pmatrix} -2 & 10 \\ 10 & -2 \end{pmatrix}$$

having eigenvalues: $8,-12$ and so there is a saddle point here.

4. Show the points $\left(\frac{1}{2},-\frac{53}{12}\right),(0,-4)$, and $(1,-4)$ are critical points of the following function of three variables and classify them according to whether they are local minimums, local maximums or saddle points.

$f(x,y)=-3x^4+6x^3+37x^2+10yx^2-10yx-40x-3y^2-24y-48.$

Answer:

The Hessian matrix is

$$\begin{pmatrix} -36x^2+74+20y+36x & 20x-10 \\ 20x-10 & -6 \end{pmatrix}.$$

Check its eigenvalues at the critical points. First consider the point $\left(\frac{1}{2},-\frac{53}{12}\right)$. At this point the Hessian is

$$\begin{pmatrix} -\frac{16}{3} & 0 \\ 0 & -6 \end{pmatrix}$$

and its eigenvalues are $-\frac{16}{3},-6$ so there is a local maximum at this point. The same analysis shows there are saddle points at the other two critical points.

5. Show the points $\left(\frac{1}{2},\frac{37}{20}\right),(0,2)$, and $(1,2)$ are critical points of the following function of three variables and classify them according to whether they are local minimums, local maximums or saddle points.

$f(x,y)=5x^4-10x^3+17x^2-6yx^2+6yx-12x+5y^2-20y+20.$

Answer:

The Hessian matrix is

$$\begin{pmatrix} 60x^2+34-12y-60x & -12x+6 \\ -12x+6 & 10 \end{pmatrix}.$$

Check its eigenvalues at the critical points. First consider the point $\left(\frac{1}{2},\frac{37}{20}\right)$. At this point, the Hessian matrix is

$$\begin{pmatrix} -\frac{16}{5} & 0 \\ 0 & 10 \end{pmatrix}$$

and its eigenvalues are $-\frac{16}{5},10$. Therefore, there is a saddle point.

Next consider $(0,2)$ at this point the Hessian matrix is

$$\begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$$

and the eigenvalues are $16, 4$. Therefore, there is a local minimum at this point. There is also a local minimum at the critical point, $(1, 2)$.

6. Show the points $\left(\frac{1}{2}, -\frac{17}{8}\right), (0, -2)$, and $(1, -2)$ are critical points of the following function of three variables and classify them according to whether they are local minimums, local maximums or saddle points.

   $f(x, y) = 4x^4 - 8x^3 - 4yx^2 + 4yx + 8x - 4x^2 + 4y^2 + 16y + 16.$

   Answer:

   The Hessian matrix is $\begin{pmatrix} 48x^2 - 8 - 8y - 48x & -8x + 4 \\ -8x + 4 & 8 \end{pmatrix}$. Check its eigenvalues at the critical points. First consider the point $\left(\frac{1}{2}, -\frac{17}{8}\right)$. This matrix is

   $\begin{pmatrix} -3 & 0 \\ 0 & 8 \end{pmatrix}$ and its eigenvalues are $-3, 8$.

   Next consider $(0, -2)$ at this point the Hessian matrix is

   $\begin{pmatrix} 8 & 4 \\ 4 & 8 \end{pmatrix}$ and the eigenvalues are $12, 4$. Finally consider the point $(1, -2)$.

   $\begin{pmatrix} 8 & -4 \\ -4 & 8 \end{pmatrix}$, eigenvalues: $12, 4$.

   If the eigenvalues are both negative, then local max. If both positive, then local min. Otherwise the test fails.

7. Find the critical points of the following function of three variables and classify them according to whether they are local minimums, local maximums or saddle points.

   $f(x, y, z) = \frac{1}{3}x^2 + \frac{32}{3}x + \frac{4}{3} - \frac{16}{3}yx - \frac{58}{3}y - \frac{4}{3}zx - \frac{46}{3}z + \frac{1}{3}y^2 - \frac{4}{3}zy - \frac{5}{3}z^2.$

   Answer:

   The critical point is at $(-2, 3, -5)$. The eigenvalues of the Hessian matrix at this point are $-6, -2$, and 6.

8. Find the critical points of the following function of three variables and classify them according to whether they are local minimums, local maximums or saddle points.

   $f(x, y, z) = -\frac{5}{3}x^2 + \frac{2}{3}x - \frac{2}{3} + \frac{8}{3}yx + \frac{2}{3}y + \frac{14}{3}zx - \frac{28}{3}z - \frac{5}{3}y^2 + \frac{14}{3}zy - \frac{8}{3}z^2.$

   Answer:

   The eigenvalues are $4, -10$, and -6 and the only critical point is $(1, 1, 0)$.

9. Find the critical points of the following function of three variables and classify them according to whether they are local minimums, local maximums or saddle points.

   $f(x, y, z) = -\frac{11}{3}x^2 + \frac{40}{3}x - \frac{56}{3} + \frac{8}{3}yx + \frac{10}{3}y - \frac{4}{3}zx + \frac{22}{3}z - \frac{11}{3}y^2 - \frac{4}{3}zy - \frac{5}{3}z^2.$

10. Find the critical points of the following function of three variables and classify them according to whether they are local minimums, local maximums or saddle points.

    $f(x, y, z) = -\frac{2}{3}x^2 + \frac{28}{3}x + \frac{37}{3} + \frac{14}{3}yx + \frac{10}{3}y - \frac{4}{3}zx - \frac{26}{3}z - \frac{2}{3}y^2 - \frac{4}{3}zy + \frac{7}{3}z^2.$

11. Show that if f has a critical point and some eigenvalue of the Hessian matrix is positive, then there exists a direction in which when f is evaluated on the line through the critical point having this direction, the resulting function of one variable has a local minimum. State and prove a similar result in the case where some eigenvalue of the Hessian matrix is negative.

12. Suppose $\mu = 0$ but there are negative eigenvalues of the Hessian at a critical point. Show by giving examples that the second derivative tests fails.

13. Show the points $\left(\frac{1}{2}, -\frac{9}{2}\right), (0,-5)$, and $(1,-5)$ are critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

    $f(x,y) = 2x^4 - 4x^3 + 42x^2 + 8yx^2 - 8yx - 40x + 2y^2 + 20y + 50.$

14. Show the points $\left(1, -\frac{11}{2}\right), (0,-5)$, and $(2,-5)$ are critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

    $f(x,y) = 4x^4 - 16x^3 - 4x^2 - 4yx^2 + 8yx + 40x + 4y^2 + 40y + 100.$

15. Show the points $\left(\frac{3}{2}, \frac{27}{20}\right), (0,0)$, and $(3,0)$ are critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

    $f(x,y) = 5x^4 - 30x^3 + 45x^2 + 6yx^2 - 18yx + 5y^2.$

16. Find the critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

    $f(x,y,z) = \frac{10}{3}x^2 - \frac{44}{3}x + \frac{64}{3} - \frac{10}{3}yx + \frac{16}{3}y + \frac{2}{3}zx - \frac{20}{3}z + \frac{10}{3}y^2 + \frac{2}{3}zy + \frac{4}{3}z^2.$

17. Find the critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

    $f(x,y,z) = -\frac{7}{3}x^2 - \frac{146}{3}x + \frac{83}{3} + \frac{16}{3}yx + \frac{4}{3}y - \frac{14}{3}zx + \frac{94}{3}z - \frac{7}{3}y^2 - \frac{14}{3}zy + \frac{8}{3}z^2.$

18. Find the critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

    $f(x,y,z) = \frac{2}{3}x^2 + 4x + 75 - \frac{14}{3}yx - 38y - \frac{8}{3}zx - 2z + \frac{2}{3}y^2 - \frac{8}{3}zy - \frac{1}{3}z^2.$

19. Find the critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.

    $f(x,y,z) = 4x^2 - 30x + 510 - 2yx + 60y - 2zx - 70z + 4y^2 - 2zy + 4z^2.$

20. Show the critical points of the following function are points of the form, $(x,y,z) = \left(t, 2t^2 - 10t, -t^2 + 5t\right)$ for $t \in \mathbf{R}$ and classify them as local minimums, local maximums or saddle points.

    $f(x,y,z) = -\frac{1}{6}x^4 + \frac{5}{3}x^3 - \frac{25}{6}x^2 + \frac{10}{3}yx^2 - \frac{50}{3}yx + \frac{19}{3}zx^2 - \frac{95}{3}zx - \frac{5}{3}y^2 - \frac{10}{3}zy - \frac{1}{6}z^2.$

    The verification that the critical points are of the indicated form is left for you.

    The Hessian is

$$\begin{pmatrix} -2x^2 + 10x - \frac{25}{3} + \frac{20}{3}y + \frac{38}{3}z & \frac{20}{3}x - \frac{50}{3} & \frac{38}{3}x - \frac{95}{3} \\ \frac{20}{3}x - \frac{50}{3} & -\frac{10}{3} & -\frac{10}{3} \\ \frac{38}{3}x - \frac{95}{3} & -\frac{10}{3} & -\frac{1}{3} \end{pmatrix}$$

    at a critical point it is

$$\begin{pmatrix} -\frac{4}{3}t^2 + \frac{20}{3}t - \frac{25}{3} & \frac{20}{3}(t) - \frac{50}{3} & \frac{38}{3}(t) - \frac{95}{3} \\ \frac{20}{3}(t) - \frac{50}{3} & -\frac{10}{3} & -\frac{10}{3} \\ \frac{38}{3}(t) - \frac{95}{3} & -\frac{10}{3} & -\frac{1}{3} \end{pmatrix}.$$

The eigenvalues are

$$0, -\frac{2}{3}t^2 + \frac{10}{3}t - 6 + \frac{2}{3}\sqrt{(t^4 - 10t^3 + 493t^2 - 2340t + 2916)},$$

and

$$-\frac{2}{3}t^2 + \frac{10}{3}t - 6 - \frac{2}{3}\sqrt{(t^4 - 10t^3 + 493t^2 - 2340t + 2916)}.$$

If you graph these functions of t you find the second is always positive and the third is always negative. Therefore, all these critical points are saddle points.

21. Show the critical points of the following function are $(0,-3,0)$, $(2,-3,0)$, and $\left(1,-3,-\frac{1}{3}\right)$ and classify them as local minimums, local maximums or saddle points.

$f(x,y,z) = -\frac{3}{2}x^4 + 6x^3 - 6x^2 + zx^2 - 2zx - 2y^2 - 12y - 18 - \frac{3}{2}z^2$.

The Hessian is

$$\begin{pmatrix} -12 + 36x + 2z - 18x^2 & 0 & -2 + 2x \\ 0 & -4 & 0 \\ -2 + 2x & 0 & -3 \end{pmatrix}$$

Now consider the critical point, $\left(1,-3,-\frac{1}{3}\right)$. At this point the Hessian matrix equals

$$\begin{pmatrix} \frac{16}{3} & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & -3 \end{pmatrix},$$

The eigenvalues are $\frac{16}{3}, -3, -4$ and so this point is a saddle point.

Next consider the critical point, $(2,-3,0)$. At this point the Hessian matrix is

$$\begin{pmatrix} -12 & 0 & 2 \\ 0 & -4 & 0 \\ 2 & 0 & -3 \end{pmatrix}$$

The eigenvalues are $-4, -\frac{15}{2} + \frac{1}{2}\sqrt{97}, -\frac{15}{2} - \frac{1}{2}\sqrt{97}$, all negative so at this point there is a local max.

Finally consider the critical point, $(0,-3,0)$. At this point the Hessian is

$$\begin{pmatrix} -12 & 0 & -2 \\ 0 & -4 & 0 \\ -2 & 0 & -3 \end{pmatrix}$$

and the eigenvalues are the same as the above, all negative. Therefore, there is a local maximum at this point.

22. Show the critical points of the following function are points of the form, $(x,y,z) = \left(t, 2t^2 + 6t, -t^2 - 3t\right)$ for $t \in \mathbf{R}$ and classify them as local minimums, local maximums or saddle points.

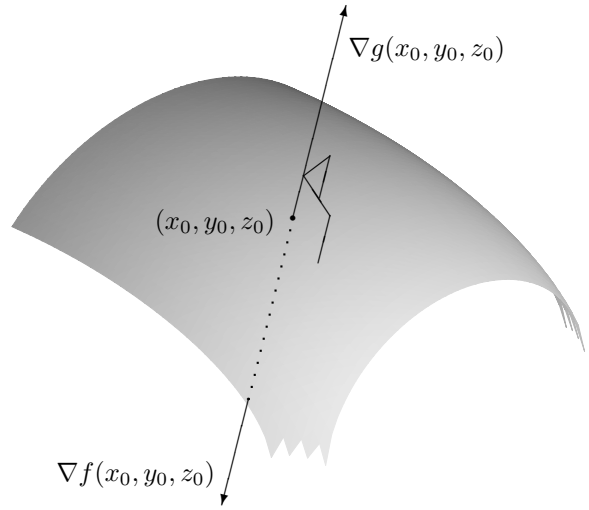
$f(x,y,z) = -2yx^2 - 6yx - 4zx^2 - 12zx + y^2 + 2yz$.

23. Show the critical points of the following function are $(0,-1,0)$, $(4,-1,0)$, and $(2,-1,-12)$ and classify them as local minimums, local maximums or saddle points.

$f(x,y,z) = \frac{1}{2}x^4 - 4x^3 + 8x^2 - 3zx^2 + 12zx + 2y^2 + 4y + 2 + \frac{1}{2}z^2$.

30.5 Lagrange Multipliers

Lagrange multipliers are used to solve extremum problems for a function defined on a level set of another function. For example, suppose you want to maximize xy given that $x+y=4$. This is not too hard to do using methods developed earlier. Solve for one of the variables, say y, in the constraint equation, $x+y=4$ to find $y=4-x$. Then the function to maximize is $f(x)=x(4-x)$ and the answer is clearly $x=2$. Thus the two numbers are $x=y=2$. This was easy because you could easily solve the constraint equation for one of the variables in terms of the other. Now what if you wanted to maximize $f(x,y,z)=xyz$ subject to the constraint that $x^2+y^2+z^2=4$? It is still possible to do this using using similar techniques. Solve for one of the variables in the constraint equation, say z, substitute it into f, and then find where the partial derivatives equal zero to find candidates for the extremum. However, it seems you might encounter many cases and it does look a little fussy. However, sometimes you can't solve the constraint equation for one variable in terms of the others. Also, what if you had many constraints. What if you wanted to maximize $f(x,y,z)$ subject to the constraints $x^2+y^2=4$ and $z=2x+3y^2$. Things are clearly getting more involved and messy. It turns out that at an extremum, there is a simple relationship between the gradient of the function to be maximized and the gradient of the constraint function. This relation can be seen geometrically as in the following picture.



In the picture, the surface represents a piece of the level surface of $g(x,y,z)=0$ and $f(x,y,z)$ is the function of three variables which is being maximized or minimized on the level surface and suppose the extremum of f occurs at the point (x_0,y_0,z_0). As shown above, $\nabla g(x_0,y_0,z_0)$ is perpendicular to the surface or more precisely to the tangent plane. However, if $\mathbf{x}(t)=(x(t),y(t),z(t))$ is a point on a smooth curve which passes through (x_0,y_0,z_0) when $t=t_0$, then the function, $h(t)=f(x(t),y(t),z(t))$ must have either a maximum or a minimum at the point, $t=t_0$. Therefore, $h'(t_0)=0$. But this means

$$\begin{aligned} 0 &= h'(t_0)=\nabla f(x(t_0),y(t_0),z(t_0))\cdot\mathbf{x}'(t_0) \\ &= \nabla f(x_0,y_0,z_0)\cdot\mathbf{x}'(t_0) \end{aligned}$$

and since this holds for any such smooth curve, $\nabla f(x_0,y_0,z_0)$ is also perpendicular to the surface. This picture represents a situation in three dimensions and you can see that it is

intuitively clear that this implies $\nabla f(x_0, y_0, z_0)$ is some scalar multiple of $\nabla g(x_0, y_0, z_0)$. Thus

$$\nabla f(x_0, y_0, z_0) = \lambda \nabla g(x_0, y_0, z_0)$$

This λ is called a Lagrange multiplier after Lagrange who considered such problems in the 1700's.

Of course the above argument is at best only heuristic. It does not deal with the question of existence of smooth curves lying in the constraint surface passing through (x_0, y_0, z_0). Nor does it consider all cases, being essentially confined to three dimensions. In addition to this, it fails to consider the situation in which there are many constraints. However, I think it is likely a geometric notion like that presented above which led Lagrange to formulate the method.

Example 30.5.1 *Maximize xyz subject to $x^2 + y^2 + z^2 = 27$.*

Here $f(x, y, z) = xyz$ while $g(x, y, z) = x^2 + y^2 + z^2 - 27$. Then $\nabla g(x, y, z) = (2x, 2y, 2z)$ and $\nabla f(x, y, z) = (yz, xz, xy)$. Then at the point which maximizes this function[1],

$$(yz, xz, xy) = \lambda (2x, 2y, 2z).$$

Therefore, each of $2\lambda x^2, 2\lambda y^2, 2\lambda z^2$ equals xyz. It follows that at any point which maximizes xyz, $|x| = |y| = |z|$. Therefore, the only candidates for the point where the maximum occurs are $(3, 3, 3), (-3, -3, 3) (-3, 3, 3)$, etc. The maximum occurs at $(3, 3, 3)$ which can be verified by plugging in to the function which is being maximized.

Example 30.5.2 *Maximize$f(x, y) = xy + y$ subject to the constraint, $x^2 + y^2 \leq 1$.*

Here I know there is a maximum because the set is the closed circle, a closed and bounded set. Therefore, it is just a matter of finding it. Look for singular points on the interior of the circle. $\nabla f(x, y) = (y, x + 1) = (0, 0)$. There are no points on the interior of the circle where the gradient equals zero. Therefore, the maximum occurs on the boundary of the circle. That is the problem reduces to maximizing $xy + y$ subject to $x^2 + y^2 = 1$. From the above,

$$(y, x + 1) - \lambda (2x, 2y) = 0.$$

Hence $y^2 - 2\lambda xy = 0$ and $x(x + 1) - 2\lambda xy = 0$ so $y^2 = x(x + 1)$. Therefore from the constraint, $x^2 + x(x + 1) = 1$ and the solution is $x = -1, x = \frac{1}{2}$. Then the candidates for a solution are $(-1, 0), \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right), \left(\frac{1}{2}, \frac{-\sqrt{3}}{2}\right)$. Then

$$f(-1, 0) = 0, f\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) = \frac{3\sqrt{3}}{4}, f\left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right) = -\frac{3\sqrt{3}}{4}.$$

It follows the maximum value of this function is $\frac{3\sqrt{3}}{4}$ and it occurs at $\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$. The minimum value is $-\frac{3\sqrt{3}}{4}$ and it occurs at $\left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right)$.

This illustrates how to use the method of Lagrange multipliers to identify the extrema for a function defined on a closed and bounded set. You try and consider the boundary as a level curve or level surface and then use the method of Lagrange multipliers on it and look for singular points on the interior of the set.

There are no magic bullets here. It was still required to solve a system of nonlinear equations to get the answer. However, it does often help to do it this way.

[1] There exists such a point because the sphere is closed and bounded.

The above generalizes to a general procedure which is described in the following major Theorem. All correct proofs of this theorem will involve some appeal to the implicit function theorem or to fundamental existence theorems from differential equations. A complete proof is very fascinating but it will not come cheap. Good advanced calculus books will usually give a correct proof. First here is a simple definition explaining one of the terms in the statement of this theorem.

Definition 30.5.3 *Let A be an $m \times n$ matrix. A submatrix is any matrix which can be obtained from A by deleting some rows and some columns.*

Theorem 30.5.4 *Let U be an open subset of $\mathbb{R}^n$ and let $f : U \to \mathbb{R}$ be a C^1 function. Then if $\mathbf{x}_0 \in U$ is either a local maximum or local minimum of f subject to the constraints*

$$g_i(\mathbf{x}) = 0, \ i = 1, \cdots, m \tag{30.4}$$

and if some $m \times m$ submatrix of

$$D\mathbf{g}(\mathbf{x}_0) \equiv \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) & g_{1x_2}(\mathbf{x}_0) & \cdots & g_{1x_n}(\mathbf{x}_0) \\ \vdots & \vdots & & \vdots \\ g_{mx_1}(\mathbf{x}_0) & g_{mx_2}(\mathbf{x}_0) & \cdots & g_{mx_n}(\mathbf{x}_0) \end{pmatrix}$$

has nonzero determinant, then there exist scalars, $\lambda_1, \cdots, \lambda_m$ such that

$$\begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \tag{30.5}$$

holds.

To help remember how to use (30.5) it may be helpful to do the following. First write the Lagrangian,

$$L = f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x})$$

and then proceed to take derivatives with respect to each of the components of $\mathbf{x}$ and also derivatives with respect to each λ_i and set all of these equations equal to 0. The formula (30.5) is what results from taking the derivatives of L with respect to the components of $\mathbf{x}$. When you take the derivatives with respect to the Lagrange multipliers, and set what results equal to 0, you just pick up the constraint equations. This yields $n + m$ equations for the $n + m$ unknowns, $x_1, \cdots, x_n, \lambda_1, \cdots, \lambda_m$. Then you proceed to look for solutions to these equations. Of course these might be impossible to find using methods of algebra, but you just do your best and hope it will work out.

Example 30.5.5 *Minimize xyz subject to the constraints $x^2 + y^2 + z^2 = 4$ and $x - 2y = 0$.*

Form the Lagrangian,

$$L = xyz - \lambda\left(x^2 + y^2 + z^2 - 4\right) - \mu(x - 2y)$$

and proceed to take derivatives with respect to every possible variable, leading to the following system of equations.

$$\begin{aligned} yz - 2\lambda x - \mu &= 0 \\ xz - 2\lambda y + 2\mu &= 0 \\ xy - 2\lambda z &= 0 \\ x^2 + y^2 + z^2 &= 4 \\ x - 2y &= 0 \end{aligned}$$

Now you have to find the solutions to this system of equations. In general, this could be very hard or even impossible. If $\lambda = 0$, then from the third equation, either x or y must equal 0. Therefore, from the first two equations, $\mu = 0$ also. If $\mu = 0$ and $\lambda \neq 0$, then from the first two equations, $xyz = 2\lambda x^2$ and $xyz = 2\lambda y^2$ and so either $x = y$ or $x = -y$, which requires that both x and y equal zero thanks to the last equation. But then from the fourth equation, $z = \pm 2$ and now this contradicts the third equation. Thus μ and λ are either both equal to zero or neither one is and the expression, xyz equals zero in this case. However, I know this is not the best value for a minimizer because I can take $x = 2\sqrt{\frac{3}{5}}, y = \sqrt{\frac{3}{5}}$, and $z = -1$. This satisfies the constraints and the product of these numbers equals a negative number. Therefore, both μ and λ must be non zero. Now use the last equation eliminate x and write the following system.

$$\begin{gathered} 5y^2 + z^2 = 4 \\ y^2 - \lambda z = 0 \\ yz - \lambda y + \mu = 0 \\ yz - 4\lambda y - \mu = 0 \end{gathered}$$

From the last equation, $\mu = (yz - 4\lambda y)$. Substitute this into the third and get

$$\begin{gathered} 5y^2 + z^2 = 4 \\ y^2 - \lambda z = 0 \\ yz - \lambda y + yz - 4\lambda y = 0 \end{gathered}$$

$y = 0$ will not yield the minimum value from the above example. Therefore, divide the last equation by y and solve for λ to get $\lambda = (2/5)\, z$. Now put this in the second equation to conclude

$$\begin{gathered} 5y^2 + z^2 = 4 \\ y^2 - (2/5)\, z^2 = 0 \end{gathered} \quad ,$$

a system which is easy to solve. Thus $y^2 = 8/15$ and $z^2 = 4/3$. Therefore, candidates for minima are $\left(2\sqrt{\frac{8}{15}}, \sqrt{\frac{8}{15}}, \pm\sqrt{\frac{4}{3}}\right)$, and $\left(-2\sqrt{\frac{8}{15}}, -\sqrt{\frac{8}{15}}, \pm\sqrt{\frac{4}{3}}\right)$, a choice of 4 points to check. Clearly the one which gives the smallest value is

$$\left(2\sqrt{\frac{8}{15}}, \sqrt{\frac{8}{15}}, -\sqrt{\frac{4}{3}}\right)$$

or $\left(-2\sqrt{\frac{8}{15}}, -\sqrt{\frac{8}{15}}, -\sqrt{\frac{4}{3}}\right)$ and the minimum value of the function subject to the constraints is $-\frac{2}{5}\sqrt{30} - \frac{2}{3}\sqrt{3}$.

You should rework this problem first solving the second easy constraint for x and then producing a simpler problem involving only the variables y and z.

The method of Lagrange multipliers allows you to consider maximization of functions defined on closed and bounded sets. Recall that any continuous function defined on a closed and bounded set has a maximum and a minimum on the set. Candidates for the extremum on the interior of the set can be located by setting the gradient equal to zero. The consideration of the boundary can then sometimes be handled with the method of Lagrange multipliers.

Example 30.5.6 *Find the maximum and minimum values of the function, $f(x,y) = xy - x^2$ on the set, $\left\{(x,y) : x^2 + 2xy + y^2 \leq 4\right\}$.*

First, the only point where ∇f equals zero is $(x,y) = (0,0)$ and this is in the desired set. In fact it is an interior point of this set. This takes care of the interior points. What about

those on the boundary $x^2+2xy+y^2=4$? The problem is to maximize $xy-x^2$ subject to the constraint, $x^2+2xy+y^2=4$. The Lagrangian is $xy-x^2-\lambda\left(x^2+2xy+y^2-4\right)$ and this yields the following system.

$$\begin{gathered} y-2x-\lambda\left(2x+2y\right)=0 \\ x-2\lambda\left(x+y\right)=0 \\ 2x^2+2xy+y^2=4 \end{gathered}$$

From the first two equations,

$$\begin{gathered} \left(2+2\lambda\right)x-\left(1-2\lambda\right)y=0 \\ \left(1-2\lambda\right)x-2\lambda y=0 \end{gathered}$$

Since not both x and y equal zero, it follows

$$\det\begin{pmatrix} 2+2\lambda & 2\lambda-1 \\ 1-2\lambda & -2\lambda \end{pmatrix}=0$$

which yields

$$\lambda=1/8$$

Therefore,

$$y=-\frac{3}{4}x \tag{30.6}$$

From the constraint equation,

$$2x^2+2x\left(-\frac{3}{4}x\right)+\left(-\frac{3}{4}x\right)^2=4$$

and so

$$x=\frac{8}{17}\sqrt{17}\text{ or }-\frac{8}{17}\sqrt{17}$$

Now from (30.6), the points of interest on the boundary of this set are

$$\left(\frac{8}{17}\sqrt{17},-\frac{6}{17}\sqrt{17}\right),\text{ and }\left(-\frac{8}{17}\sqrt{17},\frac{6}{17}\sqrt{17}\right). \tag{30.7}$$

$$\begin{aligned} f\left(\frac{8}{17}\sqrt{17},-\frac{6}{17}\sqrt{17}\right) &= \left(\frac{8}{17}\sqrt{17}\right)\left(-\frac{6}{17}\sqrt{17}\right)-\left(\frac{8}{17}\sqrt{17}\right)^2 \\ &= -\frac{112}{17} \end{aligned}$$

$$\begin{aligned} f\left(-\frac{8}{17}\sqrt{17},\frac{6}{17}\sqrt{17}\right) &= \left(-\frac{8}{17}\sqrt{17}\right)\left(\frac{6}{17}\sqrt{17}\right)-\left(-\frac{8}{17}\sqrt{17}\right)^2 \\ &= -\frac{112}{17} \end{aligned}$$

It follows the maximum value of this function on the given set occurs at $(0,0)$ and is equal to zero and the minimum occurs at either of the two points in (30.7) and has the value $-112/17$.

30.6 Exercises

1. Maximize $2x + 3y - 6z$ subject to the constraint, $x^2 + 2y^2 + 3z^2 = 9$.
2. Find the dimensions of the largest rectangle which can be inscribed in a circle of radius r.
3. Maximize $2x + y$ subject to the condition that $\frac{x^2}{4} + \frac{y^2}{9} \leq 1$.
4. Maximize $x + 2y$ subject to the condition that $x^2 + \frac{y^2}{9} \leq 1$.
5. Maximize $x + y$ subject to the condition that $x^2 + \frac{y^2}{9} + z^2 \leq 1$.
6. Maximize $x + y + z$ subject to the condition that $x^2 + \frac{y^2}{9} + z^2 \leq 1$.
7. Find the points on $y^2 x = 9$ which are closest to $(0, 0)$.
8. Find points on $xy = 4$ farthest from $(0, 0)$ if any exist. If none exist, tell why. What does this say about the method of Lagrange multipliers?
9. A can is supposed to have a volume of 36π cubic centimeters. Find the dimensions of the can which minimizes the surface area.
10. A can is supposed to have a volume of 36π cubic centimeters. The top and bottom of the can are made of tin costing 4 cents per square centimeter and the sides of the can are made of aluminum costing 5 cents per square centimeter. Find the dimensions of the can which minimizes the cost.
11. Minimize $\sum_{j=1}^{n} x_j$ subject to the constraint $\sum_{j=1}^{n} x_j^2 = a^2$. Your answer should be some function of a which you may assume is a positive number.
12. Find the point, (x, y, z) on the level surface, $4x^2 + y^2 - z^2 = 1$which is closest to $(0, 0, 0)$.
13. A curve is formed from the intersection of the plane, $2x + 3y + z = 3$ and the cylinder $x^2 + y^2 = 4$. Find the point on this curve which is closest to $(0, 0, 0)$.
14. A curve is formed from the intersection of the plane, $2x + 3y + z = 3$ and the sphere $x^2 + y^2 + z^2 = 16$. Find the point on this curve which is closest to $(0, 0, 0)$.
15. Find the point on the plane, $2x + 3y + z = 4$ which is closest to the point $(1, 2, 3)$.
16. Let $A = (A_{ij})$ be an $n \times n$ matrix which is symmetric. Thus $A_{ij} = A_{ji}$ and recall $(A\mathbf{x})_i = A_{ij} x_j$ where as usual sum over the repeated index. Show $\frac{\partial}{\partial x_i}(A_{ij} x_j x_i) = 2A_{ij} x_j$. Show that when you use the method of Lagrange multipliers to maximize the function, $A_{ij} x_j x_i$ subject to the constraint, $\sum_{j=1}^{n} x_j^2 = 1$, the value of λ which corresponds to the maximum value of this functions is such that $A_{ij} x_j = \lambda x_i$. Thus $A\mathbf{x} = \lambda \mathbf{x}$. Thus λ is an eigenvalue of the matrix, A.
17. Here are two lines. $\mathbf{x} = (1 + 2t, 2 + t, 3 + t)^T$ and $\mathbf{x} = (2 + s, 1 + 2s, 1 + 3s)^T$. Find points $\mathbf{p}_1$ on the first line and $\mathbf{p}_2$ on the second with the property that $|\mathbf{p}_1 - \mathbf{p}_2|$ is at least as small as the distance between any other pair of points, one chosen on one line and the other on the other line.
18. Find the dimensions of the largest triangle which can be inscribed in a circle of radius r.

19. Find the point on the intersection of $z = x^2 + y^2$ and $x + y + z = 1$ which is closest to $(0, 0, 0)$.

20. Minimize $4x^2 + y^2 + 9z^2$ subject to $x + y - z = 1$ and $x - 2y + z = 0$.

21. Minimize xyz subject to the constraints $x^2 + y^2 + z^2 = r^2$ and $x - y = 0$.

22. Let n be a positive integer. Find n numbers whose sum is $8n$ and the sum of the squares is as small as possible.

23. Find the point on the level surface, $2x^2 + xy + z^2 = 16$ which is closest to $(0, 0, 0)$.

24. Find the point on $\frac{x^2}{4} + \frac{y^2}{9} + z^2 = 1$ closest to the plane $x + y + z = 10$.

25. Let $x_1, \cdots, x_5$ be 5 positive numbers. Maximize their product subject to the constraint that
$$x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 = 300.$$

26. Let $f(x_1, \cdots, x_n) = x_1^n x_2^{n-1} \cdots x_n^1$. Then f achieves a maximum on the set,
$$S \equiv \left\{ \mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n i x_i = 1 \text{ and each } x_i \geq 0 \right\}.$$
If $\mathbf{x} \in S$ is the point where this maximum is achieved, find x_1/x_n.

27. Let (x, y) be a point on the ellipse, $x^2/a^2 + y^2/b^2 = 1$ which is in the first quadrant. Extend the tangent line through (x, y) till it intersects the x and y axes and let $A(x, y)$ denote the area of the triangle formed by this line and the two coordinate axes. Find the maximum value of the area of this triangle as a function of a and b.

28. Maximize $\prod_{i=1}^n x_i^2$ $(\equiv x_1^2 \times x_2^2 \times x_3^2 \times \cdots \times x_n^2)$ subject to the constraint, $\sum_{i=1}^n x_i^2 = r^2$. Show the maximum is $\left(r^2/n\right)^n$. Now show from this that
$$\left(\prod_{i=1}^n x_i^2 \right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^n x_i^2$$
and finally, conclude that if each number $x_i \geq 0$, then
$$\left(\prod_{i=1}^n x_i \right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^n x_i$$
and there exist values of the x_i for which equality holds. This says the "geometric mean" is always smaller than the arithmetic mean.

29. Maximize $x^2 y^2$ subject to the constraint
$$\frac{x^{2p}}{p} + \frac{y^{2q}}{q} = r^2$$
where p, q are real numbers larger than 1 which have the property that
$$\frac{1}{p} + \frac{1}{q} = 1.$$
show the maximum is achieved when $x^{2p} = y^{2q}$ and equals r^2. Now conclude that if $x, y > 0$, then
$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}$$
and there are values of x and y where this inequality is an equation.

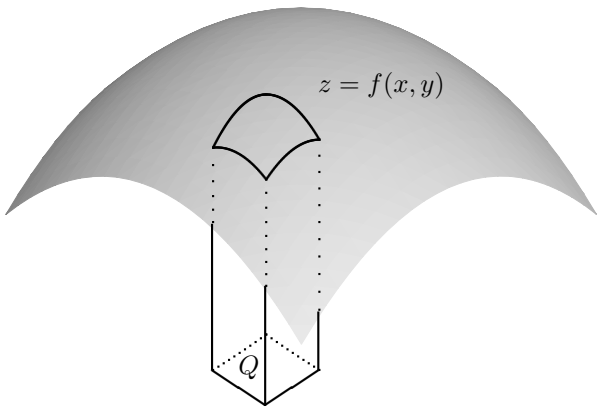
The Riemann Integral On $\mathbb{R}^n$

31.0.1 Outcomes

1. Recall and define the Riemann integral.
2. Recall the relation between iterated integrals and the Riemann integral.
3. Evaluate double integrals over simple regions.
4. Evaluate multiple integrals over simple regions.
5. Use multiple integrals to calculate the volume and mass.

31.1 Methods For Double Integrals

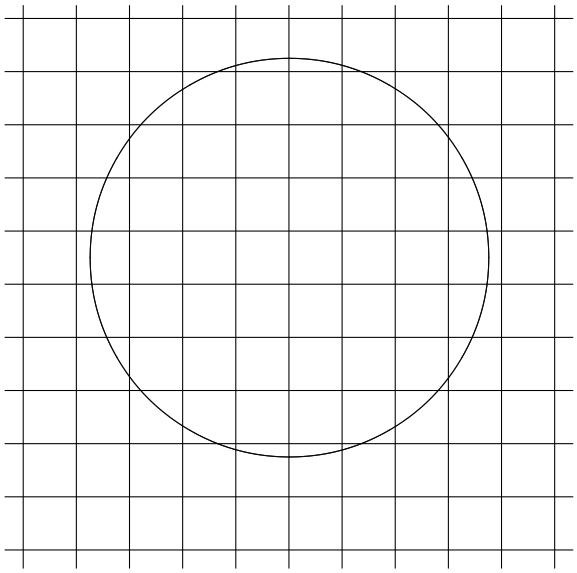
This chapter is on the Riemann integral for a function of n variables. It begins by introducing the basic concepts and applications of the integral. The proofs of the theorems involved are difficult and are left till the end. To begin with consider the problem of finding the volume under a surface of the form $z = f(x,y)$ where $f(x,y) \geq 0$ and $f(x,y) = 0$ for all (x,y) outside of some bounded set. To solve this problem, consider the following picture.



In this picture, the volume of the little prism which lies above the rectangle Q and the graph of the function would lie between $M_Q(f)\, v(Q)$ and $m_Q(f)\, v(Q)$ where

$$M_Q(f) \equiv \sup\{f(\mathbf{x}) : \mathbf{x} \in Q\},\ m_Q(f) \equiv \inf\{f(\mathbf{x}) : \mathbf{x} \in Q\}, \tag{31.1}$$

and $v(Q)$ is defined as the area of Q. Now consider the following picture.

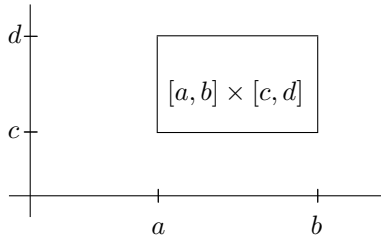


In this picture, it is assumed f equals zero outside the circle and f is a bounded nonnegative function. Then each of those little squares are the base of a prism of the sort in the previous picture and the sum of the volumes of those prisms should be the volume under the surface, $z = f(x,y)$. Therefore, the desired volume must lie between the two numbers,

$$\sum_Q M_Q(f)\, v(Q) \text{ and } \sum_Q m_Q(f)\, v(Q)$$

where the notation, $\sum_Q M_Q(f)\, v(Q)$, means for each Q, take $M_Q(f)$, multiply it by the area of $Q, v(Q)$, and then add all these numbers together. Thus in $\sum_Q M_Q(f)\, v(Q)$, adds numbers which are at least as large as what is desired while in $\sum_Q m_Q(f)\, v(Q)$ numbers are added which are at least as small as what is desired. Note this is a finite sum because by assumption, $f = 0$ except for finitely many Q, namely those which intersect the circle. The sum, $\sum_Q M_Q(f)\, v(Q)$ is called an upper sum, $\sum_Q m_Q(f)\, v(Q)$ is a lower sum, and the desired volume is caught between these upper and lower sums.

None of this depends in any way on the function being nonnegative. It also does not depend in any essential way on the function being defined on $\mathbb{R}^2$, although it is impossible to draw meaningful pictures in higher dimensional cases. To define the Riemann integral, it is necessary to first give a description of something called a grid. First you must understand that something like $[a,b] \times [c,d]$ is a rectangle in $\mathbb{R}^2$, having sides parallel to the axes. The situation is illustrated in the following picture.



$(x,y) \in [a,b] \times [c,d]$, means $x \in [a,b]$ and also $y \in [c,d]$ and the points which do this comprise the rectangle just as shown in the picture.

Definition 31.1.1 *For $i = 1,2$, let $\left\{\alpha_k^i\right\}_{k=-\infty}^{\infty}$ be points on $\mathbb{R}$ which satisfy*

$$\lim_{k\to\infty} \alpha_k^i = \infty,\ \lim_{k\to-\infty} \alpha_k^i = -\infty,\ \alpha_k^i < \alpha_{k+1}^i. \tag{31.2}$$

For such sequences, define a grid on $\mathbb{R}^2$ denoted by $\mathcal{G}$ or $\mathcal{F}$ as the collection of rectangles of the form

$$Q = \left[\alpha_k^1, \alpha_{k+1}^1\right] \times \left[\alpha_l^2, \alpha_{l+1}^2\right]. \tag{31.3}$$

If $\mathcal{G}$ is a grid, another grid, $\mathcal{F}$ is a refinement of $\mathcal{G}$ if every box of $\mathcal{G}$ is the union of boxes of $\mathcal{F}$.

For $\mathcal{G}$ a grid, the expression,

$$\sum_{Q\in\mathcal{G}} M_Q(f)\, v(Q)$$

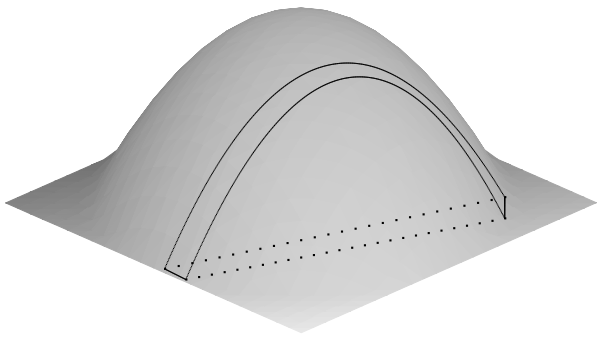
is called the upper sum associated with the grid, $\mathcal{G}$ as described above in the discussion of the volume under a surface. Again, this means to take a rectangle from $\mathcal{G}$ multiply $M_Q(f)$ defined in (31.1) by its area, $v(Q)$ and sum all these products for every $Q \in \mathcal{G}$. The symbol,

$$\sum_{Q\in\mathcal{G}} m_Q(f)\, v(Q),$$

called a lower sum, is defined similarly. With this preparation it is time to give a definition of the Riemann integral of a function of two variables.

Definition 31.1.2 *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a bounded function which equals zero for all (x,y) outside some bounded set. Then $\int f\,dV$ is defined to be the unique number which lies between all upper sums and all lower sums. In the case of $\mathbb{R}^2$, it is common to replace the V with A and write this symbol as $\int f\,dA$ where A stands for area.*

This definition begs a difficult question. For which functions does there exist a unique number between all the upper and lower sums? This interesting and fundamental question is discussed in any advanced calculus book and may be seen in the appendix on the theory of the Riemann integral. It is a hard problem which was only solved in the first part of the twentieth century. When it was solved, it was also realized that the Riemann integral was not the right integral to use. First consider the question: How can the Riemann integral be computed? Consider the following picture in which f equals zero outside the rectangle $[a,b] \times [c,d]$.



It depicts a slice taken from the solid defined by $\{(x,y) : 0 \leq y \leq f(x,y)\}$. You see these when you look at a loaf of bread. If you wanted to find the volume of the loaf of bread, and you knew the volume of each slice of bread, you could find the volume of the whole loaf by adding the volumes of individual slices. It is the same here. If you could find the volume of the slice represented in this picture, you could add these up and get the volume of the solid. The slice in the picture corresponds to constant y and is assumed to be very thin, having thickness equal to h. Denote the volume of the solid under the graph of $z = f(x,y)$ on $[a,b] \times [c,y]$ by $V(y)$. Then

$$V(y+h) - V(y) \approx h \int_a^b f(x,y)\, dx$$

where the integral is obtained by fixing y and integrating with respect to x. It is hoped that the approximation would be increasingly good as h gets smaller. Thus, dividing by h and taking a limit, it is expected that

$$V'(y) = \int_a^b f(x,y)\, dx,\ V(c) = 0.$$

Therefore, the volume of the solid under the graph of $z = f(x,y)$ is given by

$$\int_c^d \left(\int_a^b f(x,y)\, dx \right) dy \tag{31.4}$$

but this was also the result of $\int f\, dV$. Therefore, it is expected that this is a way to evaluate $\int f\, dV$. Note what has been gained here. A hard problem, finding $\int f\, dV$, is reduced to a sequence of easier problems. First do

$$\int_a^b f(x,y)\, dx$$

getting a function of y, say $F(y)$ and then do

$$\int_c^d \left(\int_a^b f(x,y)\, dx \right) dy = \int_c^d F(y)\, dy.$$

Of course there is nothing special about fixing y first. The same thing should be obtained from the integral,

$$\int_a^b \left(\int_c^d f(x,y)\, dy \right) dx \tag{31.5}$$

These expressions in (31.4) and (31.5) are called iterated integrals. They are tools for evaluating $\int f\, dV$ which would be hard to find otherwise. In practice, the parenthesis is usually omitted in these expressions. Thus

$$\int_a^b \left(\int_c^d f(x,y)\, dy \right) dx = \int_a^b \int_c^d f(x,y)\, dy\, dx$$

and it is understood that you are to do the inside integral first and then when you have done it, obtaining a function of x, you integrate this function of x.

I have presented this for the case where $f(x,y) \geq 0$ and the integral represents a volume, but there is no difference in the general case where f is not necessarily nonnegative.

Throughout, I have been assuming the notion of volume has some sort of independent meaning. This assumption is nonsense and is one of many reasons the above explanation does not rise to the level of a proof. It is only intended to make things plausible. A careful presentation which is not for the faint of heart is in an appendix.

Another aspect of this is the notion of integrating a function which is defined on some set, not on all $\mathbb{R}^2$. For example, suppose f is defined on the set, $S \subseteq \mathbb{R}^2$. What is meant by $\int_S f\,dV$?

Definition 31.1.3 *Let $f : S \to \mathbb{R}$ where S is a subset of $\mathbb{R}^2$. Then denote by f_1 the function defined by*

$$f_1(x,y) \equiv \begin{cases} f(x,y) & \text{if } (x,y) \in S \\ 0 & \text{if } (x,y) \notin S \end{cases}.$$

Then

$$\int_S f\,dV \equiv \int f_1\,dV.$$

Example 31.1.4 *Let $f(x,y) = x^2y + yx$ for $(x,y) \in [0,1] \times [0,2] \equiv R$. Find $\int_R f\,dV$.*

This is done using iterated integrals like those defined above. Thus

$$\int_R f\,dV = \int_0^1 \int_0^2 \left(x^2y + yx\right)\,dy\,dx.$$

The inside integral yields

$$\int_0^2 \left(x^2y + yx\right)\,dy = 2x^2 + 2x$$

and now the process is completed by doing $\int_0^1$ to what was just obtained. Thus

$$\int_0^1 \int_0^2 \left(x^2y + yx\right)\,dy\,dx = \int_0^1 \left(2x^2 + 2x\right)\,dx = \frac{5}{3}.$$

If the integration is done in the opposite order, the same answer should be obtained.

$$\int_0^2 \int_0^1 \left(x^2y + yx\right)\,dx\,dy$$

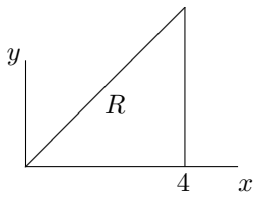
$$\int_0^1 \left(x^2y + yx\right)\,dx = \frac{5}{6}y$$

Now

$$\int_0^2 \int_0^1 \left(x^2y + yx\right)\,dx\,dy = \int_0^2 \left(\frac{5}{6}y\right)\,dy = \frac{5}{3}.$$

If a different answer had been obtained it would have been a sign that a mistake had been made.

Example 31.1.5 *Let $f(x,y) = x^2y + yx$ for $(x,y) \in R$ where R is the triangular region defined to be in the first quadrant, below the line $y = x$ and to the left of the line $x = 4$. Find $\int_R f\,dV$.*



Now from the above discussion,

$$\int_R f\,dV = \int_0^4 \int_0^x \left(x^2y + yx\right)\,dy\,dx$$

The reason for this is that x goes from 0 to 4 and for each fixed x between 0 and 4, y goes from 0 to the slanted line, $y = x$. Thus y goes from 0 to x. This explains the inside integral. Now $\int_0^x \left(x^2y + yx\right)\,dy = \frac{1}{2}x^4 + \frac{1}{2}x^3$ and so

$$\int_R f\,dV = \int_0^4 \left(\frac{1}{2}x^4 + \frac{1}{2}x^3\right)\,dx = \frac{672}{5}.$$

What of integration in a different order? Lets put the integral with respect to y on the outside and the integral with respect to x on the inside. Then

$$\int_R f\,dV = \int_0^4 \int_y^4 \left(x^2y + yx\right)\,dx\,dy$$

For each y between 0 and 4, the variable x, goes from y to 4.

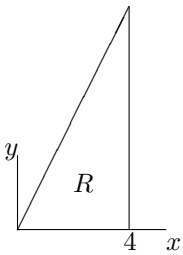
$$\int_y^4 \left(x^2y + yx\right)\,dx = \frac{88}{3}y - \frac{1}{3}y^4 - \frac{1}{2}y^3$$

Now

$$\int_R f\,dV = \int_0^4 \left(\frac{88}{3}y - \frac{1}{3}y^4 - \frac{1}{2}y^3\right)\,dy = \frac{672}{5}.$$

Here is a similar example.

Example 31.1.6 *Let $f(x,y) = x^2y$ for $(x,y) \in R$ where R is the triangular region defined to be in the first quadrant, below the line $y = 2x$ and to the left of the line $x = 4$. Find $\int_R f\,dV$.*



Put the integral with respect to x on the outside first. Then

$$\int_R f\,dV = \int_0^4 \int_0^{2x} \left(x^2y\right)\,dy\,dx$$

because for each $x \in [0,4]$, y goes from 0 to $2x$. Then

$$\int_0^{2x} \left(x^2 y\right)\, dy = 2x^4$$

and so

$$\int_R f\, dV = \int_0^4 \left(2x^4\right)\, dx = \frac{2048}{5}$$

Now do the integral in the other order. Here the integral with respect to y will be on the outside. What are the limits of this integral? Look at the triangle and note that x goes from 0 to 4 and so $2x = y$ goes from 0 to 8. Now for fixed y between 0 and 8, where does x go? It goes from the x coordinate on the line $y = 2x$ which corresponds to this y to 4. What is the x coordinate on this line which goes with y? It is $x = y/2$. Therefore, the iterated integral is

$$\int_0^8 \int_{y/2}^4 \left(x^2 y\right)\, dx\, dy.$$

Now

$$\int_{y/2}^4 \left(x^2 y\right)\, dx = \frac{64}{3} y - \frac{1}{24} y^4$$

and so

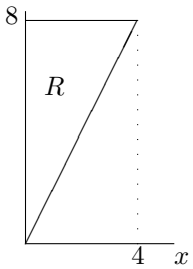
$$\int_R f\, dV = \int_0^8 \left(\frac{64}{3} y - \frac{1}{24} y^4\right)\, dy = \frac{2048}{5}$$

the same answer.

A few observations are in order here. In finding $\int_S f\, dV$ there is no problem in setting things up if S is a rectangle. However, if S is not a rectangle, the procedure **always** is agonizing. A good rule of thumb is that if what you do is easy it will be wrong. There are no shortcuts! There are no quick fixes which require no thought! Pain and suffering is inevitable and you must not expect it to be otherwise. Always draw a picture and then begin **agonizing** over the correct limits. Even when you are careful you will make lots of mistakes until you get used to the process.

Sometimes an integral can be evaluated in one order but not in another.

Example 31.1.7 *For R as shown below, find $\int_R \sin\left(y^2\right)\, dV$.*



Setting this up to have the integral with respect to y on the inside yields

$$\int_0^4 \int_{2x}^8 \sin\left(y^2\right)\, dy\, dx.$$

Unfortunately, there is no antiderivative in terms of elementary functions for $\sin\left(y^2\right)$ so there is an immediate problem in evaluating the inside integral. It doesn't work out so the

next step is to do the integration in another order and see if some progress can be made. This yields

$$\int_0^8 \int_0^{y/2} \sin\left(y^2\right)\, dx\, dy = \int_0^8 \frac{y}{2} \sin\left(y^2\right)\, dy$$

and $\int_0^8 \frac{y}{2} \sin\left(y^2\right)\, dy = -\frac{1}{4}\cos 64 + \frac{1}{4}$ which you can verify by making the substitution, $u = y^2$. Thus

$$\int_R \sin\left(y^2\right)\, dy = -\frac{1}{4}\cos 64 + \frac{1}{4}.$$

This illustrates an important idea. The integral $\int_R \sin\left(y^2\right)\, dV$ is defined as a number. It is the unique number between all the upper sums and all the lower sums. Finding it is another matter. In this case it was possible to find it using one order of integration but not the other. The iterated integral in this other order also is defined as a number but it can't be found directly without interchanging the order of integration. Of course sometimes nothing you try will work out.

31.1.1 Density And Mass

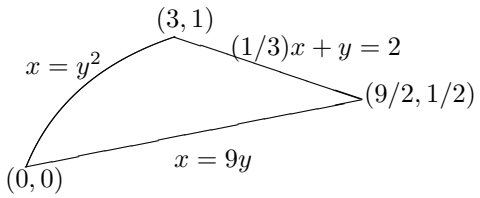
Consider a two dimensional material. Of course there is no such thing but a flat plate might be modeled as one. The density ρ is a function of position and is defined as follows. Consider a small chunk of area, dV located at the point whose Cartesian coordinates are (x, y). Then the mass of this small chunk of material is given by $\rho(x, y)\, dV$. Thus if the material occupies a region in two dimensional space, U, the total mass of this material would be

$$\int_U \rho\, dV$$

In other words you integrate the density to get the mass. Now by letting ρ depend on position, you can include the case where the material is not homogeneous. Here is an example.

Example 31.1.8 *Let $\rho(x, y)$ denote the density of the plane region determined by the curves $\frac{1}{3}x + y = 2, x = 3y^2$, and $x = 9y$. Find the total mass if $\rho(x, y) = y$.*

You need to first draw a picture of the region, R. A rough sketch follows.



This region is in two pieces, one having the graph of $x = 9y$ on the bottom and the graph of $x = 3y^2$ on the top and another piece having the graph of $x = 9y$ on the bottom and the graph of $\frac{1}{3}x + y = 2$ on the top. Therefore, in setting up the integrals, with the integral with respect to x on the outside, the double integral equals the following sum of iterated integrals.

$$\overbrace{\int_0^3 \int_{x/9}^{\sqrt{x/3}} y\, dy\, dx}^{\text{has } x=3y^2 \text{ on top}} + \overbrace{\int_3^{\frac{9}{2}} \int_{x/9}^{2-\frac{1}{3}x} y\, dy\, dx}^{\text{has } \frac{1}{3}x+y=2 \text{ on top}}$$

You notice it is not necessary to have a perfect picture, just one which is good enough to figure out what the limits should be. The dividing line between the two cases is $x = 3$ and this was shown in the picture. Now it is only a matter of evaluating the iterated integrals which in this case is routine and gives 1.

31.2 Exercises

1. Let $\rho(x,y)$ denote the density of the plane region determined by the curves $\frac{1}{4}x + y = 6, x = 4y^2$, and $x = 16y$. Find the total mass if $\rho(x,y) = y$. Your answer should be $\frac{1168}{75}$.

2. Let $\rho(x,y)$ denote the density of the plane region determined by the curves $\frac{1}{5}x + y = 6, x = 5y^2$, and $x = 25y$. Find the total mass if $\rho(x,y) = y + 2x$. Your answer should be $\frac{1735}{3}$.

3. Let $\rho(x,y)$ denote the density of the plane region determined by the curves $y = 3x, y = x, 3x + 3y = 9$. Find the total mass if $\rho(x,y) = y + 1$. Your answer should be $\frac{81}{32}$.

4. Let $\rho(x,y)$ denote the density of the plane region determined by the curves $y = 3x, y = x, 4x + 2y = 8$. Find the total mass if $\rho(x,y) = y + 1$.

5. Let $\rho(x,y)$ denote the density of the plane region determined by the curves $y = 3x, y = x, 2x + 2y = 4$. Find the total mass if $\rho(x,y) = x + 2y$.

6. Let $\rho(x,y)$ denote the density of the plane region determined by the curves $y = 3x, y = x, 5x + 2y = 10$. Find the total mass if $\rho(x,y) = y + 1$.

7. Find $\int_0^4 \int_{y/2}^2 \frac{1}{x} e^{2\frac{y}{x}}\, dx\, dy$. Your answer should be $e^4 - 1$. You might need to interchange the order of integration.

8. Find $\int_0^8 \int_{y/2}^4 \frac{1}{x} e^{3\frac{y}{x}}\, dx\, dy$.

9. Find $\int_0^8 \int_{y/2}^4 \frac{1}{x} e^{3\frac{y}{x}}\, dx\, dy$.

10. Find $\int_0^4 \int_{y/2}^2 \frac{1}{x} e^{3\frac{y}{x}}\, dx\, dy$.

11. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_0^4 \int_0^{3y} xy^3\, dx\, dy$. Your answer for the iterated integral should be $\int_0^{12} \int_{\frac{1}{3}x}^4 xy^3\, dy\, dx$.

12. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_0^3 \int_0^{3y} xy^3\, dx\, dy$.

13. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_0^2 \int_0^{2y} xy^2\, dx\, dy$.

14. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_0^3 \int_0^y xy^3\, dx\, dy$.

15. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_0^1 \int_0^y xy^2\, dx\, dy$.

16. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_0^5 \int_0^{3y} xy^2\, dx\, dy$.

17. Find $\int_0^{\frac{1}{3}\pi}\int_x^{\frac{1}{3}\pi}\frac{\sin y}{y}\,dy\,dx$. Your answer should be $\frac{1}{2}$.

18. Find $\int_0^{\frac{1}{2}\pi}\int_x^{\frac{1}{2}\pi}\frac{\sin y}{y}\,dy\,dx$.

19. Find $\int_0^{\pi}\int_x^{\pi}\frac{\sin y}{y}\,dy\,dx$

20. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_{-3}^{3}\int_{-x}^{x}x^2\,dy\,dx$

    Your answer for the iterated integral should be $\int_3^0\int_{-3}^{-y}x^2\,dx\,dy+\int_0^{-3}\int_{-3}^{y}x^2\,dx\,dy+\int_0^3\int_y^3x^2\,dx\,dy+\int_{-3}^0\int_{-y}^3x^2\,dx\,dy$. This is a very interesting example which shows that iterated integrals have a life of their own, not just as a method for evaluating double integrals.

21. Evaluate the iterated integral and then write the iterated integral with the order of integration reversed. $\int_{-2}^{2}\int_{-x}^{x}x^2\,dy\,dx$.

31.3 Methods For Triple Integrals

31.3.1 Definition Of The Integral

The integral of a function of three variables is defined similar to the integral of a function of two variables.

Definition 31.3.1 *For $i=1,2,3$ let $\left\{\alpha_k^i\right\}_{k=-\infty}^{\infty}$ be points on $\mathbb{R}$ which satisfy*

$$\lim_{k\to\infty}\alpha_k^i=\infty,\ \lim_{k\to-\infty}\alpha_k^i=-\infty,\ \alpha_k^i<\alpha_{k+1}^i. \tag{31.6}$$

For such sequences, define a grid on $\mathbb{R}^3$ denoted by $\mathcal{G}$ or $\mathcal{F}$ as the collection of boxes of the form

$$Q=\left[\alpha_k^1,\alpha_{k+1}^1\right]\times\left[\alpha_l^2,\alpha_{l+1}^2\right]\times\left[\alpha_p^3,\alpha_{p+1}^3\right]. \tag{31.7}$$

If $\mathcal{G}$ is a grid, $\mathcal{F}$ is called a refinement of $\mathcal{G}$ if every box of $\mathcal{G}$ is the union of boxes of $\mathcal{F}$.

For $\mathcal{G}$ a grid,

$$\sum_{Q\in\mathcal{G}}M_Q(f)\,v(Q)$$

is the upper sum associated with the grid, $\mathcal{G}$ where

$$M_Q(f)\equiv\sup\{f(\mathbf{x}):\mathbf{x}\in Q\}$$

and if $Q=[a,b]\times[c,d]\times[e,f]$, then $v(Q)$ is the volume of Q given by $(b-a)(d-c)(f-e)$. Letting

$$m_Q(f)\equiv\inf\{f(\mathbf{x}):\mathbf{x}\in Q\}$$

the lower sum associated with this partition is

$$\sum_{Q\in\mathcal{G}}m_Q(f)\,v(Q),$$

With this preparation it is time to give a definition of the Riemann integral of a function of three variables. This definition is just like the one for a function of two variables.

Definition 31.3.2 *Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a bounded function which equals zero outside of some bounded subset of $\mathbb{R}^3$. $\int f\,dV$ is defined as the unique number between all the upper sums and lower sums.*

As in the case of a function of two variables there are all sorts of mathematical questions which are dealt with later.

The way to think of integrals is as follows. Located at a point $\mathbf{x}$, there is an "infinitesimal" chunk of volume, dV. The integral involves taking this little chunk of volume, dV, multiplying it by $f(\mathbf{x})$ and then adding up all such products. Upper sums are too large and lower sums are too small but the unique number between all the lower and upper sums is just right and corresponds to the notion of adding up all the $f(\mathbf{x})\,dV$. Even the notation is suggestive of this concept of sum. It is a long thin S denoting sum. This is the fundamental concept for the integral in any number of dimensions and all the definitions and technicalities are designed to give precision and mathematical respectability to this notion.

To consider how to evaluate triple integrals, imagine a sum of the form $\sum_{ijk} a_{ijk}$ where there are only finitely many choices for i, j, and k and the symbol means you simply add up all the a_{ijk}. By the commutative law of addition, these may be added systematically in the form, $\sum_k \sum_j \sum_i a_{ijk}$. A similar process is used to evaluate triple integrals and since integrals are like sums, you might expect it to be valid. Specifically,

$$\int f\,dV = \int\int\int f(x,y,z)\,dx\,dy\,dz.$$

In words, sum with respect to x and then sum what you get with respect to y and finally, with respect to z. Of course this should hold in any other order such as

$$\int f\,dV = \int\int\int f(x,y,z)\,dz\,dy\,dx.$$

This is proved in an appendix[1].

Having discussed double and triple integrals, the definition of the integral of a function of n variables is accomplished in the same way.

Definition 31.3.3 *For $i = 1, \cdots, n$, let $\left\{\alpha_k^i\right\}_{k=-\infty}^{\infty}$ be points on $\mathbb{R}$ which satisfy*

$$\lim_{k\to\infty} \alpha_k^i = \infty,\ \lim_{k\to-\infty} \alpha_k^i = -\infty,\ \alpha_k^i < \alpha_{k+1}^i. \tag{31.8}$$

For such sequences, define a grid on $\mathbb{R}^n$ denoted by $\mathcal{G}$ or $\mathcal{F}$ as the collection of boxes of the form

$$Q = \prod_{i=1}^{n} \left[\alpha_{j_i}^i, \alpha_{j_i+1}^i\right]. \tag{31.9}$$

If $\mathcal{G}$ is a grid, $\mathcal{F}$ is called a refinement of $\mathcal{G}$ if every box of $\mathcal{G}$ is the union of boxes of $\mathcal{F}$.

Definition 31.3.4 *Let f be a bounded function which equals zero off a bounded set, D, and let $\mathcal{G}$ be a grid. For $Q \in \mathcal{G}$, define*

$$M_Q(f) \equiv \sup\{f(\mathbf{x}) : \mathbf{x} \in Q\},\ m_Q(f) \equiv \inf\{f(\mathbf{x}) : \mathbf{x} \in Q\}. \tag{31.10}$$

Also define for Q a box, the volume of Q, denoted by $v(Q)$ by

$$v(Q) \equiv \prod_{i=1}^{n} (b_i - a_i),\ Q \equiv \prod_{i=1}^{n} [a_i, b_i].$$

[1] All of these fundamental questions about integrals can be considered more easily in the context of the Lebesgue integral. However, this integral is more abstract than the Riemann integral.

Now define upper sums, $\mathcal{U}_{\mathcal{G}}(f)$ and lower sums, $\mathcal{L}_{\mathcal{G}}(f)$ with respect to the indicated grid, by the formulas

$$\mathcal{U}_{\mathcal{G}}(f) \equiv \sum_{Q\in\mathcal{G}} M_Q(f)\, v(Q),\ \mathcal{L}_{\mathcal{G}}(f) \equiv \sum_{Q\in\mathcal{G}} m_Q(f)\, v(Q).$$

Then a function of n variables is Riemann integrable if there is a unique number between all the upper and lower sums. This number is the value of the integral.

In this book most integrals will involve no more than three variables. However, this does not mean an integral of a function of more than three variables is unimportant. Therefore, I will begin to refer to the general case when theorems are stated.

Definition 31.3.5 *For $E \subseteq \mathbb{R}^n$,*

$$\mathcal{X}_E(\mathbf{x}) \equiv \begin{cases} 1 \text{ if } \mathbf{x}\in E \\ 0 \text{ if } \mathbf{x}\notin E \end{cases}.$$

Define $\int_E f\,dV \equiv \int \mathcal{X}_E f\,dV$ when $f\mathcal{X}_E \in \mathcal{R}(\mathbb{R}^n)$.

31.3.2 Iterated Integrals

As before, the integral is often computed by using an iterated integral.

Example 31.3.6 *Find $\int_2^3\int_3^x\int_{3y}^x (x-y)\,dz\,dy\,dx$.*

The inside integral yields $\int_{3y}^x (x-y)\,dz = x^2-4xy+3y^2$. Next this must be integrated with respect to y to give $\int_3^x \left(x^2-4xy+3y^2\right)dy = -3x^2+18x-27$. Finally the third integral gives

$$\int_2^3\int_3^x\int_{3y}^x (x-y)\,dz\,dy\,dx = \int_2^3 \left(-3x^2+18x-27\right)dx = -1.$$

Example 31.3.7 *Find $\int_0^\pi\int_0^{3y}\int_0^{y+z}\cos(x+y)\,dx\,dz\,dy$.*

The inside integral is $\int_0^{y+z}\cos(x+y)\,dx = 2\cos z\sin y\cos y+2\sin z\cos^2 y-\sin z-\sin y$. Now this has to be integrated.

$$\int_0^{3y}\int_0^{y+z}\cos(x+y)\,dx\,dz = \int_0^{3y}\left(2\cos z\sin y\cos y+2\sin z\cos^2 y-\sin z-\sin y\right)dz$$

$$= -1-16\cos^5 y+20\cos^3 y-5\cos y-3(\sin y)\,y+2\cos^2 y.$$

Finally, this last expression must be integrated from 0 to π. Thus

$$\int_0^\pi\int_0^{3y}\int_0^{y+z}\cos(x+y)\,dx\,dz\,dy$$

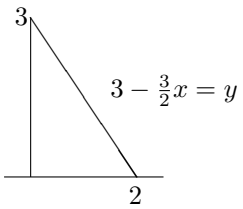
$$\begin{aligned} &= \int_0^\pi \left(-1-16\cos^5 y+20\cos^3 y-5\cos y-3(\sin y)\,y+2\cos^2 y\right)dy \\ &= -3\pi \end{aligned}$$

Example 31.3.8 *Here is an iterated integral: $\int_0^2\int_0^{3-\frac{3}{2}x}\int_0^{x^2} dz\,dy\,dx$. Write as an iterated integral in the order $dz\,dx\,dy$.*

The inside integral is just a function of x and y. (In fact, only a function of x.) The order of the last two integrals must be interchanged. Thus the iterated integral which needs to be done in a different order is

$$\int_0^2 \int_0^{3-\frac{3}{2}x} f(x,y)\, dy\, dx.$$

As usual, it is important to draw a picture and then go from there.



Thus this double integral equals

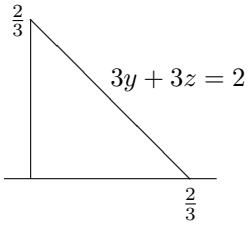
$$\int_0^3 \int_0^{\frac{2}{3}(3-y)} f(x,y)\, dx\, dy.$$

Now substituting in for $f(x,y)$,

$$\int_0^3 \int_0^{\frac{2}{3}(3-y)} \int_0^{x^2} dz\, dx\, dy.$$

Example 31.3.9 *Find the volume of the bounded region determined by* $3y+3z=2, x=16-y^2, y=0, x=0$.

In the yz plane, the following picture corresponds to $x=0$.



Therefore, the outside integrals taken with respect to z and y are of the form $\int_0^{\frac{2}{3}} \int_0^{\frac{2}{3}-y} dz\, dy$ and now for any choice of (y,z) in the above triangular region, x goes from 0 to $16-y^2$. Therefore, the iterated integral is

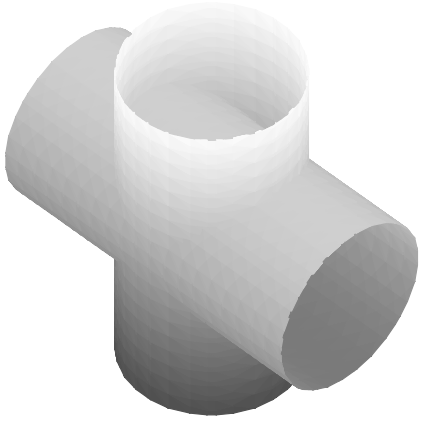
$$\int_0^{\frac{2}{3}} \int_0^{\frac{2}{3}-y} \int_0^{16-y^2} dx\, dz\, dy = \frac{860}{243}$$

Example 31.3.10 *Find the volume of the region determined by the intersection of the two cylinders,* $x^2+y^2 \leq 9$ *and* $y^2+z^2 \leq 9$.

The first listed cylinder intersects the xy plane in the disk, $x^2+y^2 \leq 9$. What is the volume of the three dimensional region which is between this disk and the two surfaces, $z=\sqrt{9-y^2}$ and $z=-\sqrt{9-y^2}$? An iterated integral for the volume is

$$\int_{-3}^3 \int_{-\sqrt{9-y^2}}^{\sqrt{9-y^2}} \int_{-\sqrt{9-y^2}}^{\sqrt{9-y^2}} dz\, dx\, dy = 144.$$

Note I drew no picture of the three dimensional region. If you are interested, here it is.



One of the cylinders is parallel to the z axis, $x^2 + y^2 \leq 9$ and the other is parallel to the x axis, $y^2 + z^2 \leq 9$. I did not need to be able to draw such a nice picture in order to work this problem. This is the key to doing these. Draw pictures in two dimensions and reason from the two dimensional pictures rather than attempt to wax artistic and consider all three dimensions at once. These problems are hard enough without making them even harder by attempting to be an artist.

31.3.3 Mass And Density

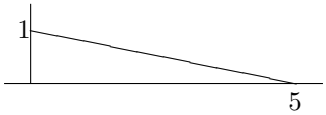
As an example of the use of triple integrals, consider a solid occupying a set of points, $U \subseteq \mathbb{R}^3$ having density ρ. Thus ρ is a function of position and the total mass of the solid equals

$$\int_U \rho \, dV.$$

This is just like the two dimensional case. The mass of an infinitesimal chunk of the solid located at $\mathbf{x}$ would be $\rho(\mathbf{x})\, dV$ and so the total mass is just the sum of all these, $\int_U \rho(\mathbf{x})\, dV$.

Example 31.3.11 *Find the volume of R where R is the bounded region formed by the plane $\frac{1}{5}x + y + \frac{1}{5}z = 1$ and the planes $x = 0, y = 0, z = 0$.*

When $z = 0$, the plane becomes $\frac{1}{5}x + y = 1$. Thus the intersection of this plane with the xy plane is this line shown in the following picture.



Therefore, the bounded region is between the triangle formed in the above picture by the x axis, the y axis and the above line and the surface given by $\frac{1}{5}x + y + \frac{1}{5}z = 1$ or $z = 5\left(1 - \left(\frac{1}{5}x + y\right)\right) = 5 - x - 5y$. Therefore, an iterated integral which yields the volume is

$$\int_0^5 \int_0^{1-\frac{1}{5}x} \int_0^{5-x-5y} dz\, dy\, dx = \frac{25}{6}.$$

Example 31.3.12 *Find the mass of the bounded region, R formed by the plane $\frac{1}{3}x+\frac{1}{3}y+\frac{1}{5}z=1$ and the planes $x=0, y=0, z=0$ if the density is $\rho(x,y,z)=z$.*

This is done just like the previous example except in this case there is a function to integrate. Thus the answer is

$$\int_0^3 \int_0^{3-x} \int_0^{5-\frac{5}{3}x-\frac{5}{3}y} z \; dz\, dy\, dx = \frac{75}{8}.$$

Example 31.3.13 *Find the total mass of the bounded solid determined by $z=9-x^2-y^2$ and $x,y,z\geq 0$ if the mass is given by $\rho(x,y,z)=z$*

When $z=0$ the surface, $z=9-x^2-y^2$ intersects the xy plane in a circle of radius 3 centered at $(0,0)$. Since $x,y\geq 0$, it is only a quarter of a circle of interest, the part where both these variables are nonnegative. For each (x,y) inside this quarter circle, z goes from 0 to $9-x^2-y^2$. Therefore, the iterated integral is of the form,

$$\int_0^3 \int_0^{\sqrt{(9-x^2)}} \int_0^{9-x^2-y^2} z \; dz\, dy\, dx = \frac{243}{8}\pi$$

Example 31.3.14 *Find the volume of the bounded region determined by $x\geq 0, y\geq 0, z\geq 0$, and $\frac{1}{7}x+y+\frac{1}{4}z=1$, and $x+\frac{1}{7}y+\frac{1}{4}z=1$.*

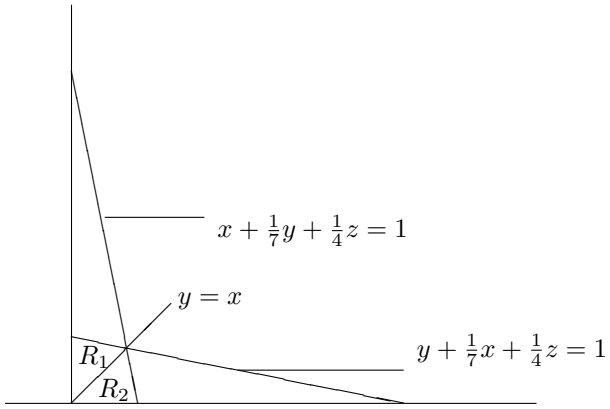
When $z=0$, the plane $\frac{1}{7}x+y+\frac{1}{4}z=1$ intersects the xy plane in the line whose equation is

$$\frac{1}{7}x+y=1$$

while the plane, $x+\frac{1}{7}y+\frac{1}{4}z=1$ intersects the xy plane in the line whose equation is

$$x+\frac{1}{7}y=1.$$

Furthermore, the two planes intersect when $x=y$ as can be seen from the equations, $x+\frac{1}{7}y=1-\frac{z}{4}$ and $\frac{1}{7}x+y=1-\frac{z}{4}$ which imply $x=y$. Thus the two dimensional picture to look at is depicted in the following picture.



You see in this picture, the base of the region in the xy plane is the union of the two triangles, R_1 and R_2. For $(x,y)\in R_1$, z goes from 0 to what it needs to be to be on the

plane, $\frac{1}{7}x+y+\frac{1}{4}z=1$. Thus z goes from 0 to $4\left(1-\frac{1}{7}x-y\right)$. Similarly, on R_2, z goes from 0 to $4\left(1-\frac{1}{7}y-x\right)$. Therefore, the integral needed is

$$\int_{R_1}\int_0^{4\left(1-\frac{1}{7}x-y\right)} dz\,dV+\int_{R_2}\int_0^{4\left(1-\frac{1}{7}y-x\right)} dz\,dV$$

and now it only remains to consider $\int_{R_1} dV$ and $\int_{R_2} dV$. The point of intersection of these lines shown in the above picture is $\left(\frac{7}{8},\frac{7}{8}\right)$ and so an iterated integral is

$$\int_0^{7/8}\int_x^{1-\frac{x}{7}}\int_0^{4\left(1-\frac{1}{7}x-y\right)} dz\,dy\,dx+\int_0^{7/8}\int_y^{1-\frac{y}{7}}\int_0^{4\left(1-\frac{1}{7}x-y\right)} dz\,dx\,dy=\frac{7}{6}.$$

31.4 Exercises With Answers

The evaluation of integrals by setting up appropriate iterated integrals and then evaluating these requires a lot of practice. Therefore, I have included exercises with answers. Each of these exercises corresponds to one which does not have answers in the next section.

1. Evaluate the integral $\int_4^7\int_5^{3x}\int_{5y}^{x} dz\,dy\,dx$
   Answer:
   $-\frac{3417}{2}$

2. Find $\int_0^4\int_0^{2-5x}\int_0^{4-2x-y}(2x)\,dz\,dy\,dx$
   Answer:
   $-\frac{2464}{3}$

3. Find $\int_0^2\int_0^{2-5x}\int_0^{1-4x-3y}(2x)\,dz\,dy\,dx$
   Answer:
   $-\frac{196}{3}$

4. Evaluate the integral $\int_5^8\int_4^{3x}\int_{4y}^{x}(x-y)\,dz\,dy\,dx$
   Answer:
   $\frac{114\,607}{8}$

5. Evaluate the integral $\int_0^{\pi}\int_0^{4y}\int_0^{y+z}\cos(x+y)\,dx\,dz\,dy$
   Answer:
   -4π

6. Evaluate the integral $\int_0^{\pi}\int_0^{2y}\int_0^{y+z}\sin(x+y)\,dx\,dz\,dy$
   Answer:
   $-\frac{19}{4}$

7. Fill in the missing limits. $\int_0^1\int_0^z\int_0^z f(x,y,z)\,dx\,dy\,dz=\int_?^?\int_?^?\int_?^? f(x,y,z)\,dx\,dz\,dy$,
   $\int_0^1\int_0^z\int_0^{2z} f(x,y,z)\,dx\,dy\,dz=\int_?^?\int_?^?\int_?^? f(x,y,z)\,dy\,dz\,dx$,
   $\int_0^1\int_0^z\int_0^z f(x,y,z)\,dx\,dy\,dz=\int_?^?\int_?^?\int_?^? f(x,y,z)\,dz\,dy\,dx$,
   $\int_0^1\int_{z/2}^{\sqrt{z}}\int_0^{y+z} f(x,y,z)\,dx\,dy\,dz=\int_?^?\int_?^?\int_?^? f(x,y,z)\,dx\,dz\,dy$,

$\int_5^7 \int_2^5 \int_0^3 f(x,y,z)\, dx\, dy\, dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\, dz\, dy\, dx.$

Answer:

$\int_0^1 \int_0^z \int_0^z f(x,y,z)\, dx\, dy\, dz = \int_0^1 \int_y^1 \int_0^z f(x,y,z)\, dx\, dz\, dy,$

$\int_0^1 \int_0^z \int_0^{2z} f(x,y,z)\, dx\, dy\, dz = \int_0^2 \int_{x/2}^1 \int_0^z f(x,y,z)\, dy\, dz\, dx,$

$\int_0^1 \int_0^z \int_0^z f(x,y,z)\, dx\, dy\, dz = \int_0^1 \left[\int_0^x \int_x^1 f(x,y,z)\, dz\, dy + \int_x^1 \int_y^1 f(x,y,z)\, dz\, dy \right] dx,$

$\int_0^1 \int_{z/2}^{\sqrt{z}} \int_0^{y+z} f(x,y,z)\, dx\, dy\, dz =$

$\int_0^{1/2} \int_{y^2}^{2y} \int_0^{y+z} f(x,y,z)\, dx\, dz\, dy + \int_{1/2}^1 \int_{y^2}^1 \int_0^{y+z} f(x,y,z)\, dx\, dz\, dy$

$\int_5^7 \int_2^5 \int_0^3 f(x,y,z)\, dx\, dy\, dz = \int_0^3 \int_2^5 \int_5^7 f(x,y,z)\, dz\, dy\, dx$

8. Find the volume of R where R is the bounded region formed by the plane $\frac{1}{5}x+y+\frac{1}{4}z = 1$ and the planes $x = 0, y = 0, z = 0$.

   Answer:

   $\int_0^5 \int_0^{1-\frac{1}{5}x} \int_0^{4-\frac{4}{5}x-4y} dz\, dy\, dx = \frac{10}{3}$

9. Find the volume of R where R is the bounded region formed by the plane $\frac{1}{5}x+\frac{1}{2}y+\frac{1}{4}z = 1$ and the planes $x = 0, y = 0, z = 0$.

   Answer:

   $\int_0^5 \int_0^{2-\frac{2}{5}x} \int_0^{4-\frac{4}{5}x-2y} dz\, dy\, dx = \frac{20}{3}$

10. Find the mass of the bounded region, R formed by the plane $\frac{1}{4}x + \frac{1}{2}y + \frac{1}{3}z = 1$ and the planes $x = 0, y = 0, z = 0$ if the density is $\rho(x,y,z) = y$

    Answer:

    $\int_0^4 \int_0^{2-\frac{1}{2}x} \int_0^{3-\frac{3}{4}x-\frac{3}{2}y} (y)\, dz\, dy\, dx = 2$

11. Find the mass of the bounded region, R formed by the plane $\frac{1}{2}x + \frac{1}{2}y + \frac{1}{4}z = 1$ and the planes $x = 0, y = 0, z = 0$ if the density is $\rho(x,y,z) = z^2$

    Answer:

    $\int_0^2 \int_0^{2-x} \int_0^{4-2x-2y} \left(z^2\right) dz\, dy\, dx = \frac{64}{15}$

12. Here is an iterated integral: $\int_0^3 \int_0^{3-x} \int_0^{x^2} dz\, dy\, dx$. Write as an iterated integral in the following orders: $dz\, dx\, dy$, $dx\, dz\, dy$, $dx\, dy\, dz$, $dy\, dx\, dz$, $dy\, dz\, dx$.

    Answer:

    $$\int_0^3 \int_0^{x^2} \int_0^{3-x} dy\, dz\, dx, \int_0^9 \int_{\sqrt{z}}^3 \int_0^{3-x} dy\, dx\, dz, \int_0^9 \int_0^{3-\sqrt{z}} \int_{\sqrt{z}}^{3-y} dx\, dy\, dz,$$
    $$\int_0^3 \int_0^{3-y} \int_0^{x^2} dz\, dx\, dy, \int_0^3 \int_0^{(3-y)^2} \int_{\sqrt{z}}^{3-y} dx\, dz\, dy$$

13. Find the volume of the bounded region determined by $5y + 2z = 4, x = 4 - y^2, y = 0, x = 0$.

    Answer:

    $\int_0^{\frac{4}{5}} \int_0^{2-\frac{5}{2}y} \int_0^{4-y^2} dx\, dz\, dy = \frac{1168}{375}$

14. Find the volume of the bounded region determined by $4y + 3z = 3, x = 4 - y^2, y = 0, x = 0$.

    Answer:
    $\int_0^{\frac{3}{4}} \int_0^{1-\frac{4}{3}y} \int_0^{4-y^2} dx\, dz\, dy = \frac{375}{256}$

15. Find the volume of the bounded region determined by $3y + z = 3, x = 4 - y^2, y = 0, x = 0$.

    Answer:
    $\int_0^1 \int_0^{3-3y} \int_0^{4-y^2} dx\, dz\, dy = \frac{23}{4}$

16. Find the volume of the region bounded by $x^2 + y^2 = 16, z = 3x, z = 0$, and $x \geq 0$.

    Answer:
    $\int_0^4 \int_{-\sqrt{(16-x^2)}}^{\sqrt{(16-x^2)}} \int_0^{3x} dz\, dy\, dx = 128$

17. Find the volume of the region bounded by $x^2 + y^2 = 25, z = 2x, z = 0$, and $x \geq 0$.

    Answer:
    $\int_0^5 \int_{-\sqrt{(25-x^2)}}^{\sqrt{(25-x^2)}} \int_0^{2x} dz\, dy\, dx = \frac{500}{3}$

18. Find the volume of the region determined by the intersection of the two cylinders, $x^2 + y^2 \leq 9$ and $y^2 + z^2 \leq 9$.

    Answer:
    $8 \int_0^3 \int_0^{\sqrt{(9-y^2)}} \int_0^{\sqrt{(9-y^2)}} dz\, dx\, dy = 144$

19. Find the total mass of the bounded solid determined by $z = a^2 - x^2 - y^2$ and $x, y, z \geq 0$ if the mass is given by $\rho(x, y, z) = z$

    Answer:
    $\int_0^4 \int_0^{\sqrt{(16-x^2)}} \int_0^{16-x^2-y^2} (z)\, dz\, dy\, dx = \frac{512}{3}\pi$

20. Find the total mass of the bounded solid determined by $z = a^2 - x^2 - y^2$ and $x, y, z \geq 0$ if the mass is given by $\rho(x, y, z) = x + 1$

    Answer:
    $\int_0^5 \int_0^{\sqrt{(25-x^2)}} \int_0^{25-x^2-y^2} (x+1)\, dz\, dy\, dx = \frac{625}{8}\pi + \frac{1250}{3}$

21. Find the volume of the region bounded by $x^2 + y^2 = 9, z = 0, z = 5 - y$

    Answer:
    $\int_{-3}^3 \int_{-\sqrt{(9-x^2)}}^{\sqrt{(9-x^2)}} \int_0^{5-y} dz\, dy\, dx = 45\pi$

22. Find the volume of the bounded region determined by $x \geq 0, y \geq 0, z \geq 0$, and $\frac{1}{2}x + y + \frac{1}{2}z = 1$, and $x + \frac{1}{2}y + \frac{1}{2}z = 1$.

    Answer:
    $\int_0^{\frac{2}{3}} \int_x^{1-\frac{1}{2}x} \int_0^{2-x-2y} dz\, dy\, dx + \int_0^{\frac{2}{3}} \int_y^{1-\frac{1}{2}y} \int_0^{2-2x-y} dz\, dx\, dy = \frac{4}{9}$

23. Find the volume of the bounded region determined by $x \geq 0, y \geq 0, z \geq 0$, and $\frac{1}{7}x + y + \frac{1}{3}z = 1$, and $x + \frac{1}{7}y + \frac{1}{3}z = 1$.

    Answer:
    $\int_0^{\frac{7}{8}} \int_x^{1-\frac{1}{7}x} \int_0^{3-\frac{3}{7}x-3y} dz\, dy\, dx + \int_0^{\frac{7}{8}} \int_y^{1-\frac{1}{7}y} \int_0^{3-3x-\frac{3}{7}y} dz\, dx\, dy = \frac{7}{8}$

24. Find the mass of the solid determined by $25x^2 + 4y^2 \leq 9, z \geq 0$, and $z = x + 2$ if the density is $\rho(x, y, z) = x$.

    Answer:

    $\int_{-\frac{3}{5}}^{\frac{3}{5}} \int_{-\frac{1}{2}\sqrt{(9-25x^2)}}^{\frac{1}{2}\sqrt{(9-25x^2)}} \int_0^{x+2} (x)\ dz\, dy\, dx = \frac{81}{1000}\pi$

25. Find $\int_0^1 \int_0^{35-5z} \int_{\frac{1}{5}x}^{7-z} (7-z)\cos\left(y^2\right)\ dy\, dx\, dz$.

    Answer:

    You need to interchange the order of integration. $\int_0^1 \int_0^{7-z} \int_0^{5y} (7-z)\cos\left(y^2\right)\ dx\, dy\, dz = \frac{5}{4}\cos 36 - \frac{5}{4}\cos 49$

26. Find $\int_0^2 \int_0^{12-3z} \int_{\frac{1}{3}x}^{4-z} (4-z)\exp\left(y^2\right)\ dy\, dx\, dz$.

    Answer:

    You need to interchange the order of integration. $\int_0^2 \int_0^{4-z} \int_0^{3y} (4-z)\exp\left(y^2\right)\ dx\, dy\, dz = -\frac{3}{4}e^4 - 9 + \frac{3}{4}e^{16}$

27. Find $\int_0^2 \int_0^{25-5z} \int_{\frac{1}{5}y}^{5-z} (5-z)\exp\left(x^2\right)\ dx\, dy\, dz$.

    Answer:

    You need to interchange the order of integration.

    $$\int_0^2 \int_0^{5-z} \int_0^{5x} (5-z)\exp\left(x^2\right)\ dy\, dx\, dz = -\frac{5}{4}e^9 - 20 + \frac{5}{4}e^{25}$$

28. Find $\int_0^1 \int_0^{10-2z} \int_{\frac{1}{2}y}^{5-z} \frac{\sin x}{x}\ dx\, dy\, dz$.

    Answer:

    You need to interchange the order of integration.

    $$\int_0^1 \int_0^{5-z} \int_0^{2x} \frac{\sin x}{x}\ dy\, dx\, dz =$$

    $$-2\sin 1\cos 5 + 2\cos 1\sin 5 + 2 - 2\sin 5$$

29. Find $\int_0^{20} \int_0^2 \int_{\frac{1}{5}y}^{6-z} \frac{\sin x}{x}\ dx\, dz\, dy + \int_{20}^{30} \int_0^{6-\frac{1}{5}y} \int_{\frac{1}{5}y}^{6-z} \frac{\sin x}{x}\ dx\, dz\, dy$.

    Answer:

    You need to interchange the order of integration.

    $$\int_0^2 \int_0^{30-5z} \int_{\frac{1}{5}y}^{6-z} \frac{\sin x}{x}\ dx\, dy\, dz = \int_0^2 \int_0^{6-z} \int_0^{5x} \frac{\sin x}{x}\ dy\, dx\, dz$$

    $$= -5\sin 2\cos 6 + 5\cos 2\sin 6 + 10 - 5\sin 6$$

31.5 Exercises

1. Evaluate the integral $\int_2^4 \int_2^{2x} \int_{2y}^{x} dz\,dy\,dx$
2. Find $\int_0^3 \int_0^{2-5x} \int_0^{2-x-2y} 2x\ dz\,dy\,dx$
3. Find $\int_0^2 \int_0^{1-3x} \int_0^{3-3x-2y} x\ dz\,dy\,dx$
4. Evaluate the integral $\int_2^5 \int_4^{3x} \int_{4y}^{x} (x-y)\ dz\,dy\,dx$
5. Evaluate the integral $\int_0^{\pi} \int_0^{3y} \int_0^{y+z} \cos(x+y)\ dx\,dz\,dy$
6. Evaluate the integral $\int_0^{\pi} \int_0^{4y} \int_0^{y+z} \sin(x+y)\ dx\,dz\,dy$
7. Fill in the missing limits. $\int_0^1 \int_0^z \int_0^z f(x,y,z)\ dx\,dy\,dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\ dx\,dz\,dy$,
   $\int_0^1 \int_0^z \int_0^{2z} f(x,y,z)\ dx\,dy\,dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\ dy\,dz\,dx$,
   $\int_0^1 \int_0^z \int_0^z f(x,y,z)\ dx\,dy\,dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\ dz\,dy\,dx$,
   $\int_0^1 \int_{z/2}^{\sqrt{z}} \int_0^{y+z} f(x,y,z)\ dx\,dy\,dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\ dx\,dz\,dy$,
   $\int_4^6 \int_2^6 \int_0^4 f(x,y,z)\ dx\,dy\,dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\ dz\,dy\,dx$.
8. Find the volume of R where R is the bounded region formed by the plane $\frac{1}{5}x+\frac{1}{3}y+\frac{1}{4}z = 1$ and the planes $x=0, y=0, z=0$.
9. Find the volume of R where R is the bounded region formed by the plane $\frac{1}{4}x+\frac{1}{2}y+\frac{1}{4}z = 1$ and the planes $x=0, y=0, z=0$.
10. Find the mass of the bounded region, R formed by the plane $\frac{1}{4}x + \frac{1}{3}y + \frac{1}{2}z = 1$ and the planes $x=0, y=0, z=0$ if the density is $\rho(x,y,z) = y+z$
11. Find the mass of the bounded region, R formed by the plane $\frac{1}{4}x + \frac{1}{2}y + \frac{1}{5}z = 1$ and the planes $x=0, y=0, z=0$ if the density is $\rho(x,y,z) = y$
12. Here is an iterated integral: $\int_0^2 \int_0^{1-\frac{1}{2}x} \int_0^{x^2} dz\,dy\,dx$. Write as an iterated integral in the following orders: $dz\,dx\,dy$, $dx\,dz\,dy$, $dx\,dy\,dz$, $dy\,dx\,dz$, $dy\,dz\,dx$.
13. Find the volume of the bounded region determined by $2y + z = 3, x = 9 - y^2, y = 0, x = 0$.
14. Find the volume of the bounded region determined by $3y + 2z = 5, x = 9 - y^2, y = 0, x = 0$.

    Your answer should be $\frac{11\,525}{648}$
15. Find the volume of the bounded region determined by $5y + 2z = 3, x = 9 - y^2, y = 0, x = 0$.
16. Find the volume of the region bounded by $x^2 + y^2 = 25, z = x, z = 0$, and $x \geq 0$.

    Your answer should be $\frac{250}{3}$.
17. Find the volume of the region bounded by $x^2 + y^2 = 9, z = 3x, z = 0$, and $x \geq 0$.
18. Find the volume of the region determined by the intersection of the two cylinders, $x^2 + y^2 \leq 16$ and $y^2 + z^2 \leq 16$.

19. Find the total mass of the bounded solid determined by $z = 4 - x^2 - y^2$ and $x, y, z \geq 0$ if the mass is given by $\rho(x, y, z) = y$

20. Find the total mass of the bounded solid determined by $z = 9 - x^2 - y^2$ and $x, y, z \geq 0$ if the mass is given by $\rho(x, y, z) = z^2$

21. Find the volume of the region bounded by $x^2 + y^2 = 4, z = 0, z = 5 - y$

22. Find the volume of the bounded region determined by $x \geq 0, y \geq 0, z \geq 0$, and $\frac{1}{7}x + \frac{1}{3}y + \frac{1}{3}z = 1$, and $\frac{1}{3}x + \frac{1}{7}y + \frac{1}{3}z = 1$.

23. Find the volume of the bounded region determined by $x \geq 0, y \geq 0, z \geq 0$, and $\frac{1}{5}x + \frac{1}{3}y + z = 1$, and $\frac{1}{3}x + \frac{1}{5}y + z = 1$.

24. Find the mass of the solid determined by $16x^2 + 4y^2 \leq 9, z \geq 0$, and $z = x + 2$ if the density is $\rho(x, y, z) = z$.

25. Find $\int_0^2 \int_0^{6-2z} \int_{\frac{1}{2}x}^{3-z} (3 - z) \cos\left(y^2\right)\, dy\, dx\, dz$.

26. Find $\int_0^1 \int_0^{18-3z} \int_{\frac{1}{3}x}^{6-z} (6 - z) \exp\left(y^2\right)\, dy\, dx\, dz$.

27. Find $\int_0^2 \int_0^{24-4z} \int_{\frac{1}{4}y}^{6-z} (6 - z) \exp\left(x^2\right)\, dx\, dy\, dz$.

28. Find $\int_0^1 \int_0^{12-4z} \int_{\frac{1}{4}y}^{3-z} \frac{\sin x}{x}\, dx\, dy\, dz$.

29. Find $\int_0^{20} \int_0^1 \int_{\frac{1}{5}y}^{5-z} \frac{\sin x}{x}\, dx\, dz\, dy + \int_{20}^{25} \int_0^{5-\frac{1}{5}y} \int_{\frac{1}{5}y}^{5-z} \frac{\sin x}{x}\, dx\, dz\, dy$. **Hint:** You might try doing it in the order, $dy\, dx\, dz$

The Integral In Other Coordinates

32.0.1 Outcomes

1. Represent a region in polar coordinates and use to evaluate integrals.
2. Represent a region in spherical or cylindrical coordinates and use to evaluate integrals.
3. Convert integrals in rectangular coordinates to integrals in polar coordinates and use to evaluate the integral.
4. Evaluate integrals in any coordinate system using the Jacobian.
5. Evaluate areas and volumes using another coordinate system.
6. Understand the transformation equations between spherical, polar and cylindrical coordinates and be able to change algebraic expressions from one system to another.
7. Use multiple integrals in an appropriate coordinate system to calculate the volume, mass, moments, center of gravity and moment of inertia.

32.1 Different Coordinates

As mentioned above, the fundamental concept of an integral is a sum of things of the form $f(\mathbf{x})\,dV$ where dV is an "infinitesimal" chunk of volume located at the point, $\mathbf{x}$. Up to now, this infinitesimal chunk of volume has had the form of a box with sides $dx_1, \cdots, dx_n$ so $dV = dx_1\,dx_2 \cdots dx_n$ but its form is not important. It could just as well be an infinitesimal parallelepiped for example. In what follows, this is what it will be.

First recall the following fundamental definition on Page 518.

Definition 32.1.1 *Let $\mathbf{u}_1, \cdots, \mathbf{u}_p$ be vectors in $\mathbb{R}^k$. The parallelepiped determined by these vectors will be denoted by $P(\mathbf{u}_1, \cdots, \mathbf{u}_p)$ and it is defined as*

$$P(\mathbf{u}_1, \cdots, \mathbf{u}_p) \equiv \left\{ \sum_{j=1}^{p} s_j \mathbf{u}_j : s_j \in [0,1] \right\}.$$

Now define the volume of this parallelepiped.

$$\text{volume of } P(\mathbf{u}_1, \cdots, \mathbf{u}_p) \equiv \left(\det\left(\mathbf{u}_i \cdot \mathbf{u}_j\right)\right)^{1/2}.$$

The dot product is used to determine this volume of a parallelepiped spanned by the given vectors and you should note that it is only the dot product that matters. Now consider spherical coordinates, ρ, ϕ, and θ. Recall there is a relationship between these coordinates and rectangular coordinates given by

$$x = \rho \sin\phi \cos\theta, y = \rho \sin\phi \sin\theta, \ z = \rho \cos\phi \tag{32.1}$$

where $\phi \in [0, \pi], \theta \in [0, 2\pi)$, and $\rho > 0$. Thus (ρ, ϕ, θ) is a point in $\mathbb{R}^3$, more specifically in the set

$$U = (0, \infty) \times [0, \pi] \times [0, 2\pi)$$

and corresponding to such a $(\rho, \phi, \theta) \in U$ there exists a unique point, $(x, y, z) \in V$ where V consists of all points of $\mathbb{R}^3$ other than the origin, $(0, 0, 0)$. This (x, y, z) determines a unique point in three dimensional space as mentioned earlier. Suppose at the point $(\rho_0, \phi_0, \theta_0) \in U$, there is an infinitesimal box having sides $d\rho$, $d\phi$, $d\theta$. Then this little box would correspond to something in V. What? Consider the mapping from U to V defined by

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \rho \sin\phi \cos\theta \\ \rho \sin\phi \sin\theta \\ \rho \cos\phi \end{pmatrix} = \mathbf{f}(\rho, \phi, \theta) \tag{32.2}$$

which takes a point, (ρ, ϕ, θ) in U and sends it to the point in V which is identified as $(x, y, z)^T \equiv \mathbf{x}$. What happens to a point of the infinitesimal box, located at $(\rho_0, \phi_0, \theta_0)$? Such a point is of the form

$$(\rho_0 + s_1 d\rho, \phi_0 + s_2\, d\phi, \theta_0 + s_3 d\theta),$$

where $s_i \geq 0$ and $\sum_i s_i \leq 1$. Also, from the definition of the derivative,

$$\mathbf{f}(\rho_0 + s_1 d\rho, \phi_0 + s_2 d\phi, \theta_0 + s_3 d\theta) - \mathbf{f}(\rho_0, \phi_0, \theta_0) =$$

$$D\mathbf{f}(\rho_0, \phi_0, \theta_0) \begin{pmatrix} s_1 d\rho \\ s_2 d\phi \\ s_3 d\theta \end{pmatrix} + \mathbf{o} \begin{pmatrix} s_1 d\rho \\ s_2 d\phi \\ s_3 d\theta \end{pmatrix}$$

where the last term may be taken equal to $\mathbf{0}$ because the vector, $(s_1 d\rho, s_2 d\phi, s_3 d\theta)^T$ is infinitesimal meaning nothing precise but conveying the idea that it is surpassingly small. Therefore, a point of this infinitesimal box is sent to the vector,

$$\overbrace{\left(\frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \rho}, \frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \phi}, \frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \theta}\right)}^{=D\mathbf{f}(\rho_0, \phi_0, \theta_0)} \begin{pmatrix} s_1 d\rho \\ s_2 d\phi \\ s_3 d\theta \end{pmatrix} =$$

$$s_1 \frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \rho} d\rho + s_2 \frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \phi} d\phi + s_3 \frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \theta} d\theta$$

a point of the infinitesimal parallelepiped determined by the vectors

$$\left\{ \frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \rho} d\rho, \frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \phi} d\phi, \frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \theta} d\theta \right\}.$$

The situation is no different for general coordinate systems. In general, $\mathbf{x} = \mathbf{f}(\mathbf{u})$ where $\mathbf{u} \in U$, a subset of $\mathbb{R}^n$ and $\mathbf{x}$ is a point in V, a subset of n dimensional space. Thus, letting the Cartesian coordinates of $\mathbf{x}$ be given by $\mathbf{x} = (x_1, \cdots, x_n)^T$, each x_i being a function of

$\mathbf{u}$, an infinitesimal box located at $\mathbf{u}_0$ corresponds to an infinitesimal parallelepiped located at $\mathbf{f}(\mathbf{u}_0)$ which is determined by the n vectors $\left\{\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} du_i\right\}_{i=1}^{n}$. From Definition 32.1.1, the volume of this infinitesimal parallelepiped located at $\mathbf{f}(\mathbf{u}_0)$ is given by

$$\det\left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} du_i \cdot \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_j} du_j\right)^{1/2} \tag{32.3}$$

in which there is no sum on the repeated index. Now in general if there are n vectors in $\mathbb{R}^n$, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$,

$$\det(\mathbf{v}_i \cdot \mathbf{v}_j)^{1/2} = |\det(\mathbf{v}_1, \cdots, \mathbf{v}_n)| \tag{32.4}$$

where this last matrix is the $n \times n$ matrix which has the i^{th} column equal to $\mathbf{v}_i$. The reason for this is that the matrix whose ij^{th} entry is $\mathbf{v}_i \cdot \mathbf{v}_j$ is just the product of the two matrices,

$$\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{pmatrix} (\mathbf{v}_1, \cdots, \mathbf{v}_n)$$

where the first on the left is the matrix having the i^{th} row equal to $\mathbf{v}_i^T$ while the matrix on the right is just the matrix having the i^{th} column equal to $\mathbf{v}_i$. Therefore, since the determinant of a matrix equals the determinant of its transpose,

$$\begin{aligned} \det(\mathbf{v}_i \cdot \mathbf{v}_j) &= \det\left(\begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T \end{pmatrix} (\mathbf{v}_1, \cdots, \mathbf{v}_n)\right) \\ &= \det(\mathbf{v}_1, \cdots, \mathbf{v}_n)^2 \end{aligned}$$

and so taking square roots yields (32.4). Therefore, from the properties of determinants, (32.3) equals

$$\left|\det\left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1} du_1, \cdots, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_n} du_n\right)\right| =$$

$$\left|\det\left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1}, \cdots, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_n}\right)\right| du_1 \cdots du_n$$

and this is the infinitesimal chunk of volume corresponding to the point $\mathbf{f}(\mathbf{u}_0)$ in V.

Definition 32.1.2 *Let $\mathbf{x} = \mathbf{f}(\mathbf{u})$ be as described above. Then the symbol, $\frac{\partial(x_1, \cdots x_n)}{\partial(u_1, \cdots, u_n)}$, called the Jacobian determinant, is defined by*

$$\det\left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1}, \cdots, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_n}\right) \equiv \frac{\partial(x_1, \cdots x_n)}{\partial(u_1, \cdots, u_n)}.$$

Also, the symbol, $\left|\frac{\partial(x_1, \cdots x_n)}{\partial(u_1, \cdots, u_n)}\right| du_1 \cdots du_n$ is called the volume element.

This has given motivation for the following fundamental procedure often called the change of variables formula which holds under fairly general conditions.

Procedure 32.1.3 *Suppose U is a subset of $\mathbb{R}^n$ and suppose $\mathbf{f}: U \to V$ is a C^1 function which is one to one.*[1] *Then if $h: V \to \mathbb{R}$,*

$$\int_U h(\mathbf{f}(\mathbf{u})) \left| \frac{\partial(x_1, \cdots x_n)}{\partial(u_1, \cdots, u_n)} \right| dV = \int_V h(\mathbf{x})\, dV.$$

Now return to Spherical coordinates. In this case, it is necessary to find the absolute value of

$$\det\left(\frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \rho}, \frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \phi}, \frac{\partial \mathbf{x}(\rho_0, \phi_0, \theta_0)}{\partial \theta} \right)$$

which equals

$$\det \begin{pmatrix} \sin\phi\cos\theta & \rho\cos\phi\cos\theta & -\rho\sin\phi\sin\theta \\ \sin\phi\sin\theta & \rho\cos\phi\sin\theta & \rho\sin\phi\cos\theta \\ \cos\phi & -\rho\sin\phi & 0 \end{pmatrix} = \rho^2 \sin\phi$$

which is positive because $\phi \in [0, \pi]$.

Example 32.1.4 *Find the volume of a ball, B_R of radius R.*

In this case, $U = (0, R] \times [0, \pi] \times [0, 2\pi)$ and use spherical coordinates. Then (32.2) yields a set in $\mathbb{R}^3$ which clearly differs from the ball of radius R only by a set having volume equal to zero. It leaves out the point at the origin is all. Therefore, the volume of the ball is

$$\begin{aligned} \int_{B_R} 1\, dV &= \int_U \rho^2 \sin\phi\, dV \\ &= \int_0^R \int_0^\pi \int_0^{2\pi} \rho^2 \sin\phi\ d\theta\, d\phi\, d\rho = \frac{4}{3} R^3 \pi. \end{aligned}$$

The reason this was effortless, is that the ball, B_R is realized as a box in terms of the spherical coordinates. Remember what was pointed out earlier about setting up iterated integrals over boxes.

Example 32.1.5 *Find the volume element for cylindrical coordinates.*

In cylindrical coordinates,

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r\cos\theta \\ r\sin\theta \\ z \end{pmatrix}$$

Therefore, the Jacobian determinant is

$$\det \begin{pmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = r.$$

It follows the volume element in cylindrical coordinates is $r\, d\theta\, dr\, dz$.

[1] This will cause non overlapping infinitesimal boxes in U to be mapped to non overlapping infinitesimal parallelepipeds in V.

Also, in the context of the Riemann integral we should say more about the sets, U and V in any case the function, h. These conditions are mainly technical however, and since a mathematically respectable treatment will not be attempted for this theorem, I think it best to give a memorable version of it which is essentially correct in all examples of interest.

Example 32.1.6 *This example uses spherical coordinates to verify an important conclusion about gravitational force. Let the hollow sphere, H be defined by $a^2 \leq x^2+y^2+z^2 \leq b^2$ and suppose this hollow sphere has constant density taken to equal 1. Now place a unit mass at the point $(0,0,z_0)$ where $|z_0| \in [a,b]$. Show the force of gravity acting on this unit mass is $\left(\alpha G \int\int\int_H \frac{(z-z_0)}{\left[x^2+y^2+(z-z_0)^2\right]^{3/2}}\,dV\right)\mathbf{k}$ and then show that if $|z_0|>b$ then the force of gravity acting on this point mass is the same as if the entire mass of the hollow sphere were placed at the origin, while if $|z_0|<a$, the total force acting on the point mass from gravity equals zero. Here G is the gravitation constant and α is the density. In particular, this shows that the force a planet exerts on an object is as though the entire mass of the planet were situated at its center*[2].

Without loss of generality, assume $z_0>0$. Let dV be a little chunk of material located at the point (x,y,z) of H the hollow sphere. Then according to Newton's law of gravity, the force this small chunk of material exerts on the given point mass equals

$$\frac{x\mathbf{i}+y\mathbf{j}+(z-z_0)\,\mathbf{k}}{|x\mathbf{i}+y\mathbf{j}+(z-z_0)\,\mathbf{k}|}\frac{1}{\left(x^2+y^2+(z-z_0)^2\right)}G\alpha\,dV=$$

$$(x\mathbf{i}+y\mathbf{j}+(z-z_0)\,\mathbf{k})\frac{1}{\left(x^2+y^2+(z-z_0)^2\right)^{3/2}}G\alpha\,dV$$

Therefore, the total force is

$$\int\int\int_H (x\mathbf{i}+y\mathbf{j}+(z-z_0)\,\mathbf{k})\frac{1}{\left(x^2+y^2+(z-z_0)^2\right)^{3/2}}G\alpha\,dV.$$

By the symmetry of the sphere, the $\mathbf{i}$ and $\mathbf{j}$ components will cancel out when the integral is taken. This is because there is the same amount of stuff for negative x and y as there is for positive x and y. Hence what remains is

$$\alpha G\mathbf{k}\int\int\int_H \frac{(z-z_0)}{\left[x^2+y^2+(z-z_0)^2\right]^{3/2}}\,dV$$

as claimed. Now for the interesting part, the integral is evaluated. In spherical coordinates this integral is.

$$\int_0^{2\pi}\int_a^b\int_0^{\pi}\frac{(\rho\cos\phi-z_0)\,\rho^2\sin\phi}{\left(\rho^2+z_0^2-2\rho z_0\cos\phi\right)^{3/2}}\,d\phi\,d\rho\,d\theta. \tag{32.5}$$

Rewrite the inside integral and use integration by parts to obtain this inside integral equals

$$\frac{1}{2z_0}\int_0^{\pi}\left(\rho^2\cos\phi-\rho z_0\right)\frac{(2z_0\rho\sin\phi)}{\left(\rho^2+z_0^2-2\rho z_0\cos\phi\right)^{3/2}}\,d\phi=$$

$$\frac{1}{2z_0}\left(-2\frac{-\rho^2-\rho z_0}{\sqrt{(\rho^2+z_0^2+2\rho z_0)}}+2\frac{\rho^2-\rho z_0}{\sqrt{(\rho^2+z_0^2-2\rho z_0)}}-\int_0^{\pi}2\rho^2\frac{\sin\phi}{\sqrt{(\rho^2+z_0^2-2\rho z_0\cos\phi)}}\,d\phi\right). \tag{32.6}$$

There are some cases to consider here.

[2]This was shown by Newton in 1685 and allowed him to assert his law of gravitation applied to the planets as though they were point masses. It was a major accomplishment.

First suppose $z_0 < a$ so the point is on the inside of the hollow sphere and it is always the case that $\rho > z_0$. Then in this case, the two first terms reduce to

$$\frac{2\rho\left(\rho+z_0\right)}{\sqrt{\left(\rho+z_0\right)^2}}+\frac{2\rho\left(\rho-z_0\right)}{\sqrt{\left(\rho-z_0\right)^2}}=\frac{2\rho\left(\rho+z_0\right)}{\left(\rho+z_0\right)}+\frac{2\rho\left(\rho-z_0\right)}{\rho-z_0}=4\rho$$

and so the expression in (32.6) equals

$$\begin{aligned}&\frac{1}{2z_0}\left(4\rho-\int_0^\pi 2\rho^2\frac{\sin\phi}{\sqrt{\left(\rho^2+z_0^2-2\rho z_0\cos\phi\right)}}\,d\phi\right)\\ =&\frac{1}{2z_0}\left(4\rho-\frac{1}{z_0}\int_0^\pi \rho\frac{2\rho z_0\sin\phi}{\sqrt{\left(\rho^2+z_0^2-2\rho z_0\cos\phi\right)}}\,d\phi\right)\\ =&\frac{1}{2z_0}\left(4\rho-\frac{2\rho}{z_0}\left(\rho^2+z_0^2-2\rho z_0\cos\phi\right)^{1/2}|_0^\pi\right)\\ =&\frac{1}{2z_0}\left(4\rho-\frac{2\rho}{z_0}\left[\left(\rho+z_0\right)-\left(\rho-z_0\right)\right]\right)=0.\end{aligned}$$

Therefore, in this case the inner integral of (32.5) equals zero and so the original integral will also be zero.

The other case is when $z_0 > b$ and so it is always the case that $z_0 > \rho$. In this case the first two terms of (32.6) are

$$\frac{2\rho\left(\rho+z_0\right)}{\sqrt{\left(\rho+z_0\right)^2}}+\frac{2\rho\left(\rho-z_0\right)}{\sqrt{\left(\rho-z_0\right)^2}}=\frac{2\rho\left(\rho+z_0\right)}{\left(\rho+z_0\right)}+\frac{2\rho\left(\rho-z_0\right)}{z_0-\rho}=0.$$

Therefore in this case, (32.6) equals

$$\begin{aligned}&\frac{1}{2z_0}\left(-\int_0^\pi 2\rho^2\frac{\sin\phi}{\sqrt{\left(\rho^2+z_0^2-2\rho z_0\cos\phi\right)}}\,d\phi\right)\\ =\;&\frac{-\rho}{2z_0^2}\left(\int_0^\pi\frac{2\rho z_0\sin\phi}{\sqrt{\left(\rho^2+z_0^2-2\rho z_0\cos\phi\right)}}\,d\phi\right)\end{aligned}$$

which equals

$$\begin{aligned}&\frac{-\rho}{z_0^2}\left(\left(\rho^2+z_0^2-2\rho z_0\cos\phi\right)^{1/2}|_0^\pi\right)\\ =\;&\frac{-\rho}{z_0^2}\left[\left(\rho+z_0\right)-\left(z_0-\rho\right)\right]=-\frac{2\rho^2}{z_0^2}.\end{aligned}$$

Thus the inner integral of (32.5) reduces to the above simple expression. Therefore, (32.5) equals

$$\int_0^{2\pi}\int_a^b\left(-\frac{2}{z_0^2}\rho^2\right)\,d\rho\,d\theta=-\frac{4}{3}\pi\frac{b^3-a^3}{z_0^2}$$

and so

$$\alpha G\mathbf{k}\int\int\int_H\frac{\left(z-z_0\right)}{\left[x^2+y^2+\left(z-z_0\right)^2\right]^{3/2}}\,dV=\alpha G\mathbf{k}\left(-\frac{4}{3}\pi\frac{b^3-a^3}{z_0^2}\right)=-\mathbf{k}G\frac{\text{total mass}}{z_0^2}.$$

32.2 Exercises With Answers

1. Find the area of the bounded region, R, determined by $3x+3y=1, 3x+3y=8, y=3x$, and $y=4x$.

   Answer:

   Let $u=\frac{y}{x}, v=3x+3y$. Then solving these equations for x and y yields

   $$\left\{x=\frac{1}{3}\frac{v}{1+u}, y=\frac{1}{3}u\frac{v}{1+u}\right\}.$$

   Now

   $$\frac{\partial(x,y)}{\partial(u,v)}=\det\begin{pmatrix} -\frac{1}{3}\frac{v}{(1+u)^2} & \frac{1}{3+3u} \\ \frac{1}{3}\frac{v}{(1+u)^2} & \frac{1}{3}\frac{u}{1+u}\end{pmatrix}=-\frac{1}{9}\frac{v}{(1+u)^2}.$$

   Also, $u\in[3,4]$ while $v\in[1,8]$. Therefore,

   $$\int_R dV=\int_3^4\int_1^8\left|-\frac{1}{9}\frac{v}{(1+u)^2}\right|dv\,du=$$

   $$\int_3^4\int_1^8\frac{1}{9}\frac{v}{(1+u)^2}\,dv\,du=\frac{7}{40}$$

2. Find the area of the bounded region, R, determined by $5x+y=1, 5x+y=9, y=2x$, and $y=5x$.

   Answer:

   Let $u=\frac{y}{x}, v=5x+y$. Then solving these equations for x and y yields

   $$\left\{x=\frac{v}{5+u}, y=u\frac{v}{5+u}\right\}.$$

   Now

   $$\frac{\partial(x,y)}{\partial(u,v)}=\det\begin{pmatrix} -\frac{v}{(5+u)^2} & \frac{1}{5+u} \\ 5\frac{v}{(5+u)^2} & \frac{u}{5+u}\end{pmatrix}=-\frac{v}{(5+u)^2}.$$

   Also, $u\in[2,5]$ while $v\in[1,9]$. Therefore,

   $$\int_R dV=\int_2^5\int_1^9\left|-\frac{v}{(5+u)^2}\right|dv\,du=\int_2^5\int_1^9\frac{v}{(5+u)^2}\,dv\,du=\frac{12}{7}$$

3. A solid, R is determined by $5x+3y=4, 5x+3y=9, y=2x$, and $y=5x$ and the density is $\rho=x$. Find the total mass of R.

   Answer:

   Let $u=\frac{y}{x}, v=5x+3y$. Then solving these equations for x and y yields

   $$\left\{x=\frac{v}{5+3u}, y=u\frac{v}{5+3u}\right\}.$$

   Now

$$\frac{\partial(x,y)}{\partial(u,v)} = \det\begin{pmatrix} -3\frac{v}{(5+3u)^2} & \frac{1}{5+3u} \\ 5\frac{v}{(5+3u)^2} & \frac{u}{5+3u} \end{pmatrix} = -\frac{v}{(5+3u)^2}.$$

Also, $u \in [2,5]$ while $v \in [4,9]$. Therefore,

$$\int_R \rho\, dV = \int_2^5 \int_4^9 \frac{v}{5+3u}\left|-\frac{v}{(5+3u)^2}\right| dv\, du =$$

$$\int_2^5 \int_4^9 \left(\frac{v}{5+3u}\right)\left(\frac{v}{(5+3u)^2}\right) dv\, du = \frac{4123}{19\,360}.$$

4. A solid, R is determined by $2x+2y=1, 2x+2y=10, y=4x$, and $y=5x$ and the density is $\rho = x+1$. Find the total mass of R.

   Answer:

   Let $u=\frac{y}{x}, v=2x+2y$. Then solving these equations for x and y yields

   $$\left\{x = \frac{1}{2}\frac{v}{1+u}, y = \frac{1}{2}u\frac{v}{1+u}\right\}.$$

   Now

   $$\frac{\partial(x,y)}{\partial(u,v)} = \det\begin{pmatrix} -\frac{1}{2}\frac{v}{(1+u)^2} & \frac{1}{2+2u} \\ \frac{1}{2}\frac{v}{(1+u)^2} & \frac{1}{2}\frac{u}{1+u} \end{pmatrix} = -\frac{1}{4}\frac{v}{(1+u)^2}.$$

   Also, $u \in [4,5]$ while $v \in [1,10]$. Therefore,

   $$\begin{aligned} \int_R \rho\, dV &= \int_4^5 \int_1^{10} (x+1)\left|-\frac{1}{4}\frac{v}{(1+u)^2}\right| dv\, du \\ &= \int_4^5 \int_1^{10} (x+1)\left(\frac{1}{4}\frac{v}{(1+u)^2}\right) dv\, du \end{aligned}$$

5. A solid, R is determined by $4x+2y=1, 4x+2y=9, y=x$, and $y=6x$ and the density is $\rho = y^{-1}$. Find the total mass of R.

   Answer:

   Let $u=\frac{y}{x}, v=4x+2y$. Then solving these equations for x and y yields

   $$\left\{x = \frac{1}{2}\frac{v}{2+u}, y = \frac{1}{2}u\frac{v}{2+u}\right\}.$$

   Now

   $$\frac{\partial(x,y)}{\partial(u,v)} = \det\begin{pmatrix} -\frac{1}{2}\frac{v}{(2+u)^2} & \frac{1}{4+2u} \\ \frac{v}{(2+u)^2} & \frac{1}{2}\frac{u}{2+u} \end{pmatrix} = -\frac{1}{4}\frac{v}{(2+u)^2}.$$

   Also, $u \in [1,6]$ while $v \in [1,9]$. Therefore,

   $$\int_R \rho\, dV = \int_1^6 \int_1^9 \left(\frac{1}{2}u\frac{v}{2+u}\right)^{-1}\left|-\frac{1}{4}\frac{v}{(2+u)^2}\right| dv\, du = -4\ln 2 + 4\ln 3$$

6. Find the volume of the region, E, bounded by the ellipsoid, $\frac{1}{4}x^2+\frac{1}{9}y^2+\frac{1}{49}z^2=1$.

   Answer:

   Let $u=\frac{1}{2}x, v=\frac{1}{3}y, w=\frac{1}{7}z$. Then (u,v,w) is a point in the unit ball, B. Therefore,

$$\int_B \frac{\partial(x,y,z)}{\partial(u,v,w)}\,dV=\int_E dV.$$

   But $\frac{\partial(x,y,z)}{\partial(u,v,w)}=42$ and so the answer is

$$(volume\ of\ B)\times 42=\frac{4}{3}\pi 42=56\pi.$$

7. Here are three vectors. $(4,1,4)^T,(5,0,4)^T$, and$(3,1,5)^T$. These vectors determine a parallelepiped, R, which is occupied by a solid having density $\rho=x$. Find the mass of this solid.

   Answer:

   Let $\begin{pmatrix}4&5&3\\1&0&1\\4&4&5\end{pmatrix}\begin{pmatrix}u\\v\\w\end{pmatrix}=\begin{pmatrix}x\\y\\z\end{pmatrix}$. Then this maps the unit cube,

$$Q\equiv[0,1]\times[0,1]\times[0,1]$$

   onto R and

$$\frac{\partial(x,y,z)}{\partial(u,v,w)}=\left|\det\begin{pmatrix}4&5&3\\1&0&1\\4&4&5\end{pmatrix}\right|=|-9|=9$$

   so the mass is

$$\int_R x\,dV=\int_Q (4u+5v+3w)\,(9)\,\,dV$$

$$=\int_0^1\int_0^1\int_0^1 (4u+5v+3w)\,(9)\,\,du\,dv\,dw=54$$

8. Here are three vectors. $(3,2,6)^T,(4,1,6)^T$, and $(2,2,7)^T$. These vectors determine a parallelepiped, R, which is occupied by a solid having density $\rho=y$. Find the mass of this solid.

   Answer:

   Let $\begin{pmatrix}3&4&2\\2&1&2\\6&6&7\end{pmatrix}\begin{pmatrix}u\\v\\w\end{pmatrix}=\begin{pmatrix}x\\y\\z\end{pmatrix}$. Then this maps the unit cube,

$$Q\equiv[0,1]\times[0,1]\times[0,1]$$

   onto R and

$$\frac{\partial(x,y,z)}{\partial(u,v,w)}=\left|\det\begin{pmatrix}3&4&2\\2&1&2\\6&6&7\end{pmatrix}\right|=|-11|=11$$

and so the mass is

$$\int_R x\,dV = \int_Q (2u+v+2w)\,(11)\;dV$$

$$= \int_0^1 \int_0^1 \int_0^1 (2u+v+2w)\,(11)\;du\,dv\,dw = \frac{55}{2}.$$

9. Here are three vectors. $(2,2,4)^T, (3,1,4)^T$, and $(1,2,5)^T$. These vectors determine a parallelepiped, R, which is occupied by a solid having density $\rho = y+x$. Find the mass of this solid.

Answer:

Let $\begin{pmatrix} 2 & 3 & 1 \\ 2 & 1 & 2 \\ 4 & 4 & 5 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$. Then this maps the unit cube,

$$Q \equiv [0,1] \times [0,1] \times [0,1]$$

onto R and

$$\frac{\partial(x,y,z)}{\partial(u,v,w)} = \left| \det \begin{pmatrix} 2 & 3 & 1 \\ 2 & 1 & 2 \\ 4 & 4 & 5 \end{pmatrix} \right| = |-8| = 8$$

and so the mass is $2u+3v+w$

$$\int_R x\,dV = \int_Q (4u+4v+3w)\,(8)\;dV$$

$$= \int_0^1 \int_0^1 \int_0^1 (4u+4v+3w)\,(8)\;du\,dv\,dw = 44.$$

10. Let $D = \left\{(x,y) : x^2+y^2 \leq 25\right\}$. Find $\int_D e^{36x^2+36y^2}\,dx\,dy$.

Answer:

This is easy in polar coordinates. $x = r\cos\theta, y = r\sin\theta$. Thus $\frac{\partial(x,y)}{\partial(r,\theta)} = r$ and in terms of these new coordinates, the disk, D, is the rectangle,

$$R = \{(r,\theta) \in [0,5] \times [0,2\pi]\}.$$

Therefore,

$$\int_D e^{36x^2+36y^2}\,dV = \int_R e^{36r^2} r\,dV =$$

$$\int_0^5 \int_0^{2\pi} e^{36r^2} r\,d\theta\,dr = \frac{1}{36}\pi\left(e^{900}-1\right).$$

Note you wouldn't get very far without changing the variables in this.

11. Let $D = \left\{(x,y) : x^2 + y^2 \leq 9\right\}$. Find $\int_D \cos\left(36x^2 + 36y^2\right)\, dx\, dy$.

    Answer:

    This is easy in polar coordinates. $x = r\cos\theta, y = r\sin\theta$. Thus $\frac{\partial(x,y)}{\partial(r,\theta)} = r$ and in terms of these new coordinates, the disk, D, is the rectangle,

    $$R = \{(r,\theta) \in [0,3] \times [0,2\pi]\}.$$

    Therefore,

    $$\int_D \cos\left(36x^2 + 36y^2\right)\, dV = \int_R \cos\left(36r^2\right) r\, dV =$$

    $$\int_0^3 \int_0^{2\pi} \cos\left(36r^2\right) r\, d\theta\, dr = \frac{1}{36}\left(\sin 324\right)\pi.$$

12. The ice cream in a sugar cone is described in spherical coordinates by $\rho \in [0,8], \phi \in \left[0, \frac{1}{4}\pi\right], \theta \in [0,2\pi]$. If the units are in centimeters, find the total volume in cubic centimeters of this ice cream.

    Answer:

    Remember that in spherical coordinates, the volume element is $\rho^2 \sin\phi\, dV$ and so the total volume of this is $\int_0^8 \int_0^{\frac{1}{4}\pi} \int_0^{2\pi} \rho^2 \sin\phi\, d\theta\, d\phi\, d\rho = -\frac{512}{3}\sqrt{2}\pi + \frac{1024}{3}\pi$.

13. Find the volume between $z = 5 - x^2 - y^2$ and $z = \sqrt{(x^2 + y^2)}$.

    Answer:

    Use cylindrical coordinates. In terms of these coordinates the shape is

    $$h - r^2 \geq z \geq r, r \in \left[0, \frac{1}{2}\sqrt{21} - \frac{1}{2}\right], \theta \in [0,2\pi].$$

    Also, $\frac{\partial(x,y,z)}{\partial(r,\theta,z)} = r$. Therefore, the volume is

    $$\int_0^{2\pi} \int_0^{\frac{1}{2}\sqrt{21}-\frac{1}{2}} \int_0^{5-r^2} r\, dz\, dr\, d\theta = \frac{39}{4}\pi + \frac{1}{4}\pi\sqrt{21}$$

14. A ball of radius 12 is placed in a drill press and a hole of radius 4 is drilled out with the center of the hole a diameter of the ball. What is the volume of the material which remains?

    Answer:

    You know the formula for the volume of a sphere and so if you find out how much stuff is taken away, then it will be easy to find what is left. To find the volume of what is removed, it is easiest to use cylindrical coordinates. This volume is

    $$\int_0^4 \int_0^{2\pi} \int_{-\sqrt{(144-r^2)}}^{\sqrt{(144-r^2)}} r\, dz\, d\theta\, dr = -\frac{4096}{3}\sqrt{2}\pi + 2304\pi.$$

    Therefore, the volume of what remains is $\frac{4}{3}\pi\left(12\right)^3$ minus the above. Thus the volume of what remains is

    $$\frac{4096}{3}\sqrt{2}\pi.$$

15. A ball of radius 11 has density equal to $\sqrt{x^2+y^2+z^2}$ in rectangular coordinates. The top of this ball is sliced off by a plane of the form $z = 1$. What is the mass of what remains?

    Answer:

$$\int_0^{2\pi}\int_0^{\arcsin\left(\frac{2}{11}\sqrt{30}\right)}\int_0^{\sec\phi}\rho^3\sin\phi\,d\rho\,d\phi\,d\theta+\int_0^{2\pi}\int_{\arcsin\left(\frac{2}{11}\sqrt{30}\right)}^{\pi}\int_0^{11}\rho^3\sin\phi\,d\rho\,d\phi\,d\theta$$

$$=\frac{24\,623}{3}\pi$$

16. Find $\int\int_S \frac{y}{x}\,dV$ where S is described in polar coordinates as $1 \le r \le 2$ and $0 \le \theta \le \pi/4$.

    Answer:

    Use $x = r\cos\theta$ and $y = r\sin\theta$. Then the integral in polar coordinates is

$$\int_0^{\pi/4}\int_1^2 (r\tan\theta)\;dr\,d\theta = \frac{3}{4}\ln 2.$$

17. Find $\int\int_S \left(\left(\frac{y}{x}\right)^2+1\right)\,dV$ where S is given in polar coordinates as $1 \le r \le 2$ and $0 \le \theta \le \frac{1}{4}\pi$.

    Answer:

    Use $x = r\cos\theta$ and $y = r\sin\theta$. Then the integral in polar coordinates is

$$\int_0^{\frac{1}{4}\pi}\int_1^2 \left(1+\tan^2\theta\right) r\,dr\,d\theta.$$

18. Use polar coordinates to evaluate the following integral. Here S is given in terms of the polar coordinates. $\int\int_S \sin\left(4x^2+4y^2\right)\,dV$ where $r \le 2$ and $0 \le \theta \le \frac{1}{6}\pi$.

    Answer:

$$\int_0^{\frac{1}{6}\pi}\int_0^2 \sin\left(4r^2\right) r\,dr\,d\theta.$$

19. Find $\int\int_S e^{2x^2+2y^2}\,dV$ where S is given in terms of the polar coordinates, $r \le 2$ and $0 \le \theta \le \frac{1}{3}\pi$.

    Answer:

    The integral is

$$\int_0^{\frac{1}{3}\pi}\int_0^2 re^{2r^2}\,dr\,d\theta = \frac{1}{12}\pi\left(e^8-1\right).$$

20. Compute the volume of a sphere of radius R using cylindrical coordinates.

    Answer:

    Using cylindrical coordinates, the integral is $\int_0^{2\pi}\int_0^R\int_{-\sqrt{R^2-r^2}}^{\sqrt{R^2-r^2}} r\,dz\,dr\,d\theta = \frac{4}{3}\pi R^3$.

32.3 Exercises

1. Find the area of the bounded region, R, determined by $5x+y=2, 5x+y=8, y=2x$, and $y=6x$.

2. Find the area of the bounded region, R, determined by $y+2x=6, y+2x=10, y=3x$, and $y=4x$.

3. A solid, R is determined by $3x+y=2, 3x+y=4, y=2x$, and $y=6x$ and the density is $\rho=x$. Find the total mass of R.

4. A solid, R is determined by $4x+2y=5, 4x+2y=6, y=5x$, and $y=7x$ and the density is $\rho=y$. Find the total mass of R.

5. A solid, R is determined by $3x+y=3, 3x+y=10, y=3x$, and $y=5x$ and the density is $\rho=y^{-1}$. Find the total mass of R.

6. Find the volume of the region, E, bounded by the ellipsoid, $\frac{1}{4}x^2+y^2+z^2=1$.

7. Here are three vectors. $(4,1,2)^T, (5,0,2)^T$, and $(3,1,3)^T$. These vectors determine a parallelepiped, R, which is occupied by a solid having density $\rho=x$. Find the mass of this solid.

8. Here are three vectors. $(5,1,6)^T, (6,0,6)^T$, and $(4,1,7)^T$. These vectors determine a parallelepiped, R, which is occupied by a solid having density $\rho=y$. Find the mass of this solid.

9. Here are three vectors. $(5,2,9)^T, (6,1,9)^T$, and $(4,2,10)^T$. These vectors determine a parallelepiped, R, which is occupied by a solid having density $\rho=y+x$. Find the mass of this solid.

10. Let $D=\{(x,y): x^2+y^2\leq 25\}$. Find $\int_D e^{25x^2+25y^2}\,dx\,dy$.

11. Let $D=\{(x,y): x^2+y^2\leq 16\}$. Find $\int_D \cos\left(9x^2+9y^2\right)\,dx\,dy$.

12. The ice cream in a sugar cone is described in spherical coordinates by $\rho\in[0,10], \phi\in\left[0,\frac{1}{3}\pi\right], \theta\in[0,2\pi]$. If the units are in centimeters, find the total volume in cubic centimeters of this ice cream.

13. Find the volume between $z=5-x^2-y^2$ and $z=2\sqrt{(x^2+y^2)}$.

14. A ball of radius 3 is placed in a drill press and a hole of radius 2 is drilled out with the center of the hole a diameter of the ball. What is the volume of the material which remains?

15. A ball of radius 9 has density equal to $\sqrt{x^2+y^2+z^2}$ in rectangular coordinates. The top of this ball is sliced off by a plane of the form $z=2$. What is the mass of what remains?

16. Find $\int\int_S \frac{y}{x}\,dV$ where S is described in polar coordinates as $1\leq r\leq 2$ and $0\leq\theta\leq\pi/4$.

17. Find $\int\int_S \left(\left(\frac{y}{x}\right)^2+1\right)\,dV$ where S is given in polar coordinates as $1\leq r\leq 2$ and $0\leq\theta\leq\frac{1}{6}\pi$.

18. Use polar coordinates to evaluate the following integral. Here S is given in terms of the polar coordinates. $\int\int_S \sin\left(2x^2+2y^2\right)\,dV$ where $r\leq 2$ and $0\leq\theta\leq\frac{3}{2}\pi$.

19. Find $\int\int_S e^{2x^2+2y^2}\,dV$ where S is given in terms of the polar coordinates, $r \leq 2$ and $0 \leq \theta \leq \pi$.

20. Compute the volume of a sphere of radius R using cylindrical coordinates.

21. In Example 32.1.6 on Page 727 check out all the details by working the integrals to be sure the steps are right.

22. What if the hollow sphere in Example 32.1.6 were in two dimensions and everything, including Newton's law still held? Would similar conclusions hold? Explain.

23. Fill in all details for the following argument that $\int_0^\infty e^{-x^2}dx = \frac{1}{2}\sqrt{\pi}$. Let $I = \int_0^\infty e^{-x^2}dx$. Then

$$I^2 = \int_0^\infty \int_0^\infty e^{-\left(x^2+y^2\right)}dx\,dy = \int_0^{\pi/2} \int_0^\infty re^{-r^2}dr\,d\theta = \frac{1}{4}\pi$$

from which the result follows.

24. Show $\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx = 1$. Here σ is a positive number called the standard deviation and μ is a number called the mean.

25. Show using Problem 23 $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$,.

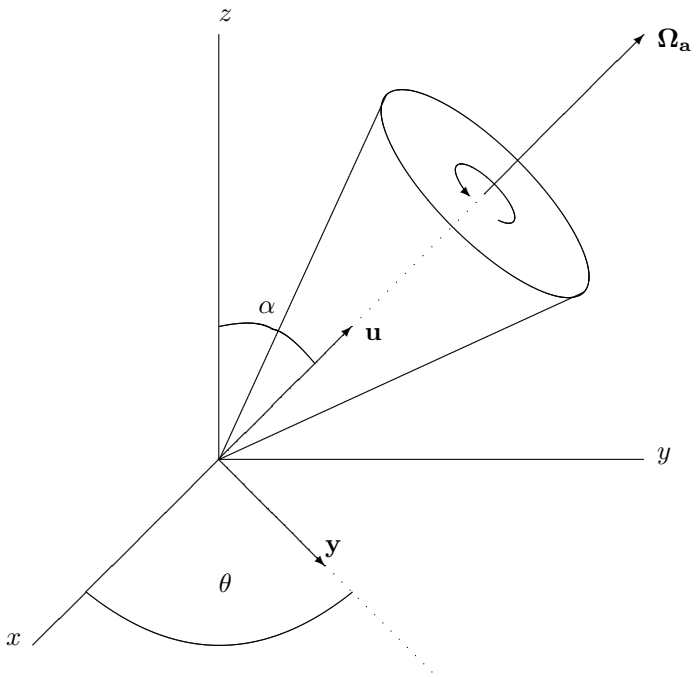
26. Let $p, q > 0$ and define $B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}$. Show $\Gamma(p)\Gamma(q) = B(p,q)\Gamma(p+q)$. **Hint:** It is fairly routine if you start with the left side and proceed to change variables.

32.4 The Moment Of Inertia

In order to appreciate the importance of this concept, it is necessary to discuss its physical significance.

32.4.1 The Spinning Top

To begin with consider a spinning top as illustrated in the following picture.



For the purpose of this discussion, consider the top as a large number of point masses, m_i, located at the positions, $\mathbf{r}_i(t)$ for $i = 1, 2, \cdots, N$ and these masses are symmetrically arranged relative to the axis of the top. As the top spins, the axis of symmetry is observed to move around the z axis. This is called precession and you will see it occur whenever you spin a top. What is the speed of this precession? In other words, what is θ'? The following discussion follows one given in Sears and Zemansky [26].

Imagine a coordinate system which is fixed relative to the moving top. Thus in this coordinate system the points of the top are fixed. Let the standard unit vectors of the coordinate system moving with the top be denoted by $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$. From Theorem 24.4.2 on Page 571, there exists an angular velocity vector $\boldsymbol{\Omega}(t)$ such that if $\mathbf{u}(t)$ is the position vector of a point fixed in the top, $(\mathbf{u}(t) = u_1\mathbf{i}(t) + u_2\mathbf{j}(t) + u_3\mathbf{k}(t))$,

$$\mathbf{u}'(t) = \boldsymbol{\Omega}(t) \times \mathbf{u}(t).$$

The vector $\boldsymbol{\Omega}_a$ shown in the picture is the vector for which

$$\mathbf{r}_i'(t) \equiv \boldsymbol{\Omega}_a \times \mathbf{r}_i(t)$$

is the velocity of the i^{th} point mass due to rotation about the axis of the top. Thus $\boldsymbol{\Omega}(t) = \boldsymbol{\Omega}_a(t) + \boldsymbol{\Omega}_p(t)$ and it is assumed $\boldsymbol{\Omega}_p(t)$ is very small relative to $\boldsymbol{\Omega}_a$. In other words, it is assumed the axis of the top moves very slowly relative to the speed of the points in the top which are spinning very fast around the axis of the top. The angular momentum, $\mathbf{L}$ is defined by

$$\mathbf{L} \equiv \sum_{i=1}^{N} \mathbf{r}_i \times m_i\mathbf{v}_i \tag{32.7}$$

where $\mathbf{v}_i$ equals the velocity of the i^{th} point mass. Thus $\mathbf{v}_i = \boldsymbol{\Omega}(t) \times \mathbf{r}_i$ and from the above

assumption, $\mathbf{v}_i$ may be taken equal to $\mathbf{\Omega}_a \times \mathbf{r}_i$. Therefore, $\mathbf{L}$ is essentially given by

$$\begin{aligned}\mathbf{L} &\equiv \sum_{i=1}^{N} m_i \mathbf{r}_i \times (\mathbf{\Omega}_a \times \mathbf{r}_i) \\ &= \sum_{i=1}^{N} m_i \left(|\mathbf{r}_i|^2 \mathbf{\Omega}_a - (\mathbf{r}_i \cdot \mathbf{\Omega}_a) \mathbf{r}_i \right).\end{aligned}$$

By symmetry of the top, this last expression equals a multiple of $\mathbf{\Omega}_a$. Thus $\mathbf{L}$ is parallel to $\mathbf{\Omega}_a$. Also,

$$\begin{aligned}\mathbf{L} \cdot \Omega_a &= \sum_{i=1}^{N} m_i \mathbf{\Omega}_a \cdot \mathbf{r}_i \times (\mathbf{\Omega}_a \times \mathbf{r}_i) \\ &= \sum_{i=1}^{N} m_i (\mathbf{\Omega}_a \times \mathbf{r}_i) \cdot (\mathbf{\Omega}_a \times \mathbf{r}_i) \\ &= \sum_{i=1}^{N} m_i |\mathbf{\Omega}_a \times \mathbf{r}_i|^2 = \sum_{i=1}^{N} m_i |\mathbf{\Omega}_a|^2 |\mathbf{r}_i|^2 \sin^2(\beta_i)\end{aligned}$$

where β_i denotes the angle between the position vector of the i^{th} point mass and the axis of the top. Since this expression is positive, this also shows $\mathbf{L}$ has the same direction as $\mathbf{\Omega}_a$. Let $\omega \equiv |\mathbf{\Omega}_a|$. Then the above expression is of the form

$$\mathbf{L} \cdot \mathbf{\Omega}_a = I\omega^2,$$

where

$$I \equiv \sum_{i=1}^{N} m_i |\mathbf{r}_i|^2 \sin^2(\beta_i).$$

Thus, to get I you take the mass of the i^{th} point mass, multiply it by the square of its distance to the axis of the top and add all these up. This is defined as the moment of inertia of the top about the axis of the top. Letting $\mathbf{u}$ denote a unit vector in the direction of the axis of the top, this implies

$$\mathbf{L} = I\omega\mathbf{u}. \tag{32.8}$$

Note the simple description of the angular momentum in terms of the moment of inertia. Referring to the above picture, define the vector, $\mathbf{y}$ to be the projection of the vector, $\mathbf{u}$ on the xy plane. Thus

$$\mathbf{y} = \mathbf{u} - (\mathbf{u} \cdot \mathbf{k})\mathbf{k}$$

and

$$(\mathbf{u} \cdot \mathbf{i}) = (\mathbf{y} \cdot \mathbf{i}) = \sin\alpha \cos\theta. \tag{32.9}$$

Now also from (32.7),

$$\begin{aligned}\frac{d\mathbf{L}}{dt} &= \sum_{i=1}^{N} m_i \overbrace{\mathbf{r}_i' \times \mathbf{v}_i}^{=0} + \mathbf{r}_i \times m_i \mathbf{v}_i' \\ &= \sum_{i=1}^{N} \mathbf{r}_i \times m_i \mathbf{v}_i' = -\sum_{i=1}^{N} \mathbf{r}_i \times m_i g \mathbf{k}\end{aligned}$$

where g is the acceleration of gravity. From (32.8), (32.9), and the above,

$$\begin{aligned}\frac{d\mathbf{L}}{dt}\cdot\mathbf{i} &= I\omega\left(\frac{d\mathbf{u}}{dt}\cdot\mathbf{i}\right) = I\omega\left(\frac{d\mathbf{y}}{dt}\cdot\mathbf{i}\right)\\ &= \left(-I\omega\sin\alpha\sin\theta\right)\theta' = -\sum_{i=1}^{N}\mathbf{r}_i\times m_i g\mathbf{k}\cdot\mathbf{i}\\ &= -\sum_{i=1}^{N} m_i g\mathbf{r}_i\cdot\mathbf{k}\times\mathbf{i} = -\sum_{i=1}^{N} m_i g\mathbf{r}_i\cdot\mathbf{j}. \end{aligned} \tag{32.10}$$

To simplify this further, recall the following definition of the center of mass.

Definition 32.4.1 *Define the total mass, M by*

$$M = \sum_{i=1}^{N} m_i$$

and the center of mass, $\mathbf{r}_0$ by

$$\mathbf{r}_0 \equiv \frac{\sum_{i=1}^{N}\mathbf{r}_i m_i}{M}. \tag{32.11}$$

In terms of the center of mass, the last expression equals

$$\begin{aligned}-Mg\mathbf{r}_0\cdot\mathbf{j} &= -Mg\left(\mathbf{r}_0 - (\mathbf{r}_0\cdot\mathbf{k})\,\mathbf{k} + (\mathbf{r}_0\cdot\mathbf{k})\,\mathbf{k}\right)\cdot\mathbf{j}\\ &= -Mg\left(\mathbf{r}_0 - (\mathbf{r}_0\cdot\mathbf{k})\,\mathbf{k}\right)\cdot\mathbf{j}\\ &= -Mg\left|\mathbf{r}_0 - (\mathbf{r}_0\cdot\mathbf{k})\,\mathbf{k}\right|\cos\theta\\ &= -Mg\left|\mathbf{r}_0\right|\sin\alpha\cos\left(\frac{\pi}{2}-\theta\right).\end{aligned}$$

Note that by symmetry, $\mathbf{r}_0(t)$ is on the axis of the top, is in the same direction as $\mathbf{L}, \mathbf{u}$, and $\boldsymbol{\Omega}_a$, and also $|\mathbf{r}_0|$ is independent of t. Therefore, from the second line of (32.10),

$$\left(-I\omega\sin\alpha\sin\theta\right)\theta' = -Mg\left|\mathbf{r}_0\right|\sin\alpha\sin\theta.$$

which shows

$$\theta' = \frac{Mg\left|\mathbf{r}_0\right|}{I\omega}. \tag{32.12}$$

From (32.12), the angular velocity of precession does not depend on α in the picture. It also is slower when ω is large and I is large.

The above discussion is a considerable simplification of the problem of a spinning top obtained from an assumption that $\boldsymbol{\Omega}_a$ is approximately equal to $\boldsymbol{\Omega}$. It also leaves out all considerations of friction and the observation that the axis of symmetry wobbles. This is wobbling is called nutation. The full mathematical treatment of this problem involves the Euler angles and some fairly complicated differential equations obtained using techniques discussed in advanced physics classes. Lagrange studied these types of problems back in the 1700's.

32.4.2 Kinetic Energy

The next problem is that of understanding the total kinetic energy of a collection of moving point masses. Consider a possibly large number of point masses, m_i located at the positions $\mathbf{r}_i$ for $i = 1, 2, \cdots, N$. Thus the velocity of the i^{th} point mass is $\mathbf{r}_i' = \mathbf{v}_i$. The kinetic energy of the mass m_i is defined by

$$\frac{1}{2} m_i \left|\mathbf{r}_i'\right|^2.$$

(This is a very good time to review the presentation on kinetic energy given on Page 577.) The total kinetic energy of the collection of masses is then

$$E = \sum_{i=1}^{N} \frac{1}{2} m_i \left|\mathbf{r}_i'\right|^2. \tag{32.13}$$

As these masses move about, so does the center of mass, $\mathbf{r}_0$. Thus $\mathbf{r}_0$ is a function of t just as the other $\mathbf{r}_i$. From (32.13) the total kinetic energy is

$$\begin{aligned} E &= \sum_{i=1}^{N} \frac{1}{2} m_i \left|\mathbf{r}_i' - \mathbf{r}_0' + \mathbf{r}_0'\right|^2 \\ &= \sum_{i=1}^{N} \frac{1}{2} m_i \left[\left|\mathbf{r}_i' - \mathbf{r}_0'\right|^2 + \left|\mathbf{r}_0'\right|^2 + 2\left(\mathbf{r}_i' - \mathbf{r}_0' \cdot \mathbf{r}_0'\right)\right]. \end{aligned} \tag{32.14}$$

Now

$$\begin{aligned} \sum_{i=1}^{N} m_i \left(\mathbf{r}_i' - \mathbf{r}_0' \cdot \mathbf{r}_0'\right) &= \left(\sum_{i=1}^{N} m_i \left(\mathbf{r}_i - \mathbf{r}_0\right)\right)' \cdot \mathbf{r}_0' \\ &= 0 \end{aligned}$$

because from (32.11)

$$\begin{aligned} \sum_{i=1}^{N} m_i \left(\mathbf{r}_i - \mathbf{r}_0\right) &= \sum_{i=1}^{N} m_i \mathbf{r}_i - \sum_{i=1}^{N} m_i \mathbf{r}_0 \\ &= \sum_{i=1}^{N} m_i \mathbf{r}_i - \sum_{i=1}^{N} m_i \left(\frac{\sum_{i=1}^{N} \mathbf{r}_i m_i}{\sum_{i=1}^{N} m_i}\right) = \mathbf{0}. \end{aligned}$$

Let $M \equiv \sum_{i=1}^{N} m_i$ be the total mass. Then (32.14) reduces to

$$\begin{aligned} E &= \sum_{i=1}^{N} \frac{1}{2} m_i \left[\left|\mathbf{r}_i' - \mathbf{r}_0'\right|^2 + \left|\mathbf{r}_0'\right|^2\right] \\ &= \frac{1}{2} M \left|\mathbf{r}_0'\right|^2 + \sum_{i=1}^{N} \frac{1}{2} m_i \left|\mathbf{r}_i' - \mathbf{r}_0'\right|^2. \end{aligned} \tag{32.15}$$

The first term is just the kinetic energy of a point mass equal to the sum of all the masses involved, located at the center of mass of the system of masses while the second term represents kinetic energy which comes from the relative velocities of the masses taken with respect to the center of mass. It is this term which is considered more carefully in the case where the system of masses maintain distance between each other.

To illustrate the contrast between the case where the masses maintain a constant distance and on in which they don't, take a hard boiled egg and spin it and then take a raw egg and give it a spin. You will certainly feel a big difference in the way the two eggs respond. Incidentally, this is a good way to tell whether the egg has been hard boiled or is raw and can be used to prevent messiness which could occur if you think it is hard boiled and it really isn't.

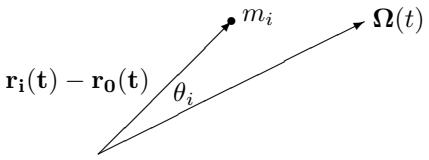
Now let $\mathbf{e}_1(t), \mathbf{e}_2(t)$, and $\mathbf{e}_3(t)$ be an orthonormal set of vectors which is fixed in the body undergoing rigid body motion. This means that $\mathbf{r}_i(t) - \mathbf{r}_0(t)$ has components which are constant in t with respect to the vectors, $\mathbf{e}_i(t)$. By Theorem 24.4.2 on Page 571 there exists a vector, $\mathbf{\Omega}(t)$ which does not depend on i such that

$$\mathbf{r}_i'(t) - \mathbf{r}_0'(t) = \mathbf{\Omega}(t) \times (\mathbf{r}_i(t) - \mathbf{r}_0(t)).$$

Now using this in (32.15),

$$\begin{aligned} E &= \frac{1}{2}M|\mathbf{r}_0'|^2 + \sum_{i=1}^{N}\frac{1}{2}m_i|\mathbf{\Omega}(t) \times (\mathbf{r}_i(t) - \mathbf{r}_0(t))|^2 \\ &= \frac{1}{2}M|\mathbf{r}_0'|^2 + \frac{1}{2}\left(\sum_{i=1}^{N} m_i|\mathbf{r}_i(t) - \mathbf{r}_0(t)|^2 \sin^2\theta_i\right)|\mathbf{\Omega}(t)|^2 \\ &= \frac{1}{2}M|\mathbf{r}_0'|^2 + \frac{1}{2}\left(\sum_{i=1}^{N} m_i|\mathbf{r}_i(0) - \mathbf{r}_0(0)|^2 \sin^2\theta_i\right)|\mathbf{\Omega}(t)|^2 \end{aligned}$$

where θ_i is the angle between $\mathbf{\Omega}(t)$ and the vector, $\mathbf{r}_i(t) - \mathbf{r}_0(t)$. Therefore, $|\mathbf{r}_i(t) - \mathbf{r}_0(t)| \sin\theta_i$ is the distance between the point mass, m_i located at $\mathbf{r}_i$ and a line through the center of mass, $\mathbf{r}_0$ with direction, $\mathbf{\Omega}$ as indicated in the following picture.



Thus the expression, $\sum_{i=1}^{N} m_i|\mathbf{r}_i(0) - \mathbf{r}_0(0)|^2 \sin^2\theta_i$ plays the role of a mass in the definition of kinetic energy except instead of the speed, substitute the angular speed, $|\mathbf{\Omega}(t)|$. It is this expression which is called the moment of inertia about the line whose direction is $\mathbf{\Omega}(t)$.

In both of these examples, the center of mass and the moment of inertia occurred in a natural way.

32.4.3 Finding The Moment Of Inertia And Center Of Mass

The methods used to evaluate multiple integrals make possible the determination of centers of mass and moments of inertia. In the case of a solid material rather than finitely many point masses, you replace the sums with integrals. The sums are essentially approximations of the integrals which result.

Example 32.4.2 *Let a solid occupy the three dimensional region R and suppose the density is ρ. What is the moment of inertia of this solid about the z axis? What is the center of mass?*

Here the little masses would be of the form $\rho(\mathbf{x})\,dV$ where $\mathbf{x}$ is a point of R. Therefore, the contribution of this mass to the moment of inertia would be

$$\left(x^2+y^2\right)\rho(\mathbf{x})\,dV$$

where the Cartesian coordinates of the point $\mathbf{x}$ are (x,y,z). Then summing these up as an integral, yields the following for the moment of inertia.

$$\int_R \left(x^2+y^2\right)\rho(\mathbf{x})\;dV. \tag{32.16}$$

To find the center of mass, sum up $\mathbf{r}\rho\,dV$ for the points in R and divide by the total mass. In Cartesian coordinates, where $\mathbf{r}=(x,y,z)$, this means to sum up vectors of the form $(x\rho\,dV, y\rho\,dV, z\rho\,dV)$ and divide by the total mass. Thus the Cartesian coordinates of the center of mass are

$$\left(\frac{\int_R x\rho\,dV}{\int_R \rho\,dV}, \frac{\int_R y\rho\,dV}{\int_R \rho\,dV}, \frac{\int_R z\rho\,dV}{\int_R \rho\,dV}\right) \equiv \frac{\int_R \mathbf{r}\rho\,dV}{\int_R \rho\,dV}.$$

Here is a specific example.

Example 32.4.3 *Find the moment of inertia about the z axis and center of mass of the solid which occupies the region, R defined by $9-\left(x^2+y^2\right)\geq z\geq 0$ if the density is $\rho(x,y,z)=\sqrt{x^2+y^2}$.*

This moment of inertia is $\int_R \left(x^2+y^2\right)\sqrt{x^2+y^2}\,dV$ and the easiest way to find this integral is to use cylindrical coordinates. Thus the answer is

$$\int_0^{2\pi}\int_0^3\int_0^{9-r^2} r^3 r\,dz\,dr\,d\theta = \frac{8748}{35}\pi.$$

To find the center of mass, note the x and y coordinates of the center of mass,

$$\frac{\int_R x\rho\,dV}{\int_R \rho\,dV}, \frac{\int_R y\rho\,dV}{\int_R \rho\,dV}$$

both equal zero because the above shape is symmetric about the z axis and ρ is also symmetric in its values. Thus $x\rho\,dV$ will cancel with $-x\rho\,dV$ and a similar conclusion will hold for the y coordinate. It only remains to find the z coordinate of the center of mass, $\overline{z}$. In polar coordinates, $\rho=r$ and so,

$$\overline{z}=\frac{\int_R z\rho\,dV}{\int_R \rho\,dV}=\frac{\int_0^{2\pi}\int_0^3\int_0^{9-r^2} zr^2\,dz\,dr\,d\theta}{\int_0^{2\pi}\int_0^3\int_0^{9-r^2} r^2\,dz\,dr\,d\theta}=\frac{18}{7}.$$

Thus the center of mass will be $\left(0,0,\frac{18}{7}\right)$.

32.5 Exercises

1. Let R denote the finite region bounded by $z=4-x^2-y^2$ and the xy plane. Find z_c, the z coordinate of the center of mass if the density, σ is a constant.

2. Let R denote the finite region bounded by $z=4-x^2-y^2$ and the xy plane. Find z_c, the z coordinate of the center of mass if the density, σ is equals $\sigma(x,y,z)=z$.

3. Find the mass and center of mass of the region between the surfaces $z = -y^2 + 8$ and $z = 2x^2 + y^2$ if the density equals $\sigma = 1$.

4. Find the mass and center of mass of the region between the surfaces $z = -y^2 + 8$ and $z = 2x^2 + y^2$ if the density equals $\sigma(x, y, z) = x^2$.

5. The two cylinders, $x^2 + y^2 = 4$ and $y^2 + z^2 = 4$ intersect in a region, set, R. Find the mass and center of mass if the density, σ, is given by $\sigma(x, y, z) = z^2$.

6. The two cylinders, $x^2 + y^2 = 4$ and $y^2 + z^2 = 4$ intersect in a region, set, R. Find the mass and center of mass if the density, σ, is given by $\sigma(x, y, z) = 4 + z$.

7. Find the mass and center of mass of the set, (x, y, z) such that $\frac{x^2}{4} + \frac{y^2}{9} + z^2 \leq 1$ if the density is $\sigma(x, y, z) = 4 + y + z$.

8. Let R denote the finite region bounded by $z = 9 - x^2 - y^2$ and the xy plane. Find the moment of inertia of this shape about the z axis.

9. Let R denote the finite region bounded by $z = 9 - x^2 - y^2$ and the xy plane. Find the moment of inertia of this shape about the x axis.

10. Let B be a solid ball of constant density and radius R. Find the moment of inertia about a line through a diameter of the ball. You should get $\frac{2}{5}R^2M$.

11. Let B be a solid ball of density, $\sigma = \rho$ where ρ is the distance to the center of the ball which has radius R. Find the moment of inertia about a line through a diameter of the ball. Write your answer in terms of the total mass and the radius as was done in the constant density case.

12. Let C be a solid cylinder of constant density and radius R. Find the moment of inertia about the axis of the cylinder

    You should get $\frac{1}{2}R^2M$.

13. Let C be a solid cylinder of constant density and radius R and mass M and let B be a solid ball of radius R and mass M. The cylinder and the sphere are placed on the top of an inclined plane and allowed to roll to the bottom. Which one will arrive first and why?

14. Suppose a solid of mass M occupying the region, B has moment of inertia, I_l about a line, l which passes through the center of mass of M and let l_1 be another line parallel to l and at a distance of a from l. Then the parallel axis theorem states $I_{l_1} = I_l + a^2M$. Prove the parallel axis theorem. **Hint:** Choose axes such that the z axis is l and l_1 passes through the point $(a, 0)$ in the xy plane.

15. Using the parallel axis theorem find the moment of inertia of a solid ball of radius R and mass M about an axis located at a distance of a from the center of the ball. Your answer should be $Ma^2 + \frac{2}{5}MR^2$.

16. Consider all axes in computing the moment of inertia of a solid. Will the smallest possible moment of inertia always result from using an axis which goes through the center of mass?

17. Find the moment of inertia of a solid thin rod of length l, mass M, and constant density about an axis through the center of the rod perpendicular to the axis of the rod. You should get $\frac{1}{12}l^2M$.

18. Using the parallel axis theorem, find the moment of inertia of a solid thin rod of length l, mass M, and constant density about an axis through an end of the rod perpendicular to the axis of the rod. You should get $\frac{1}{3}l^2M$.

19. Let the angle between the z axis and the sides of a right circular cone be α. Also assume the height of this cone is h. Find the z coordinate of the center of mass of this cone in terms of α and h assuming the density is constant.

20. Let the angle between the z axis and the sides of a right circular cone be α. Also assume the height of this cone is h. Assuming the density is $\sigma = 1$, find the moment of inertia about the z axis in terms of α and h.

21. Let R denote the part of the solid ball, $x^2 + y^2 + z^2 \leq R^2$ which lies in the first octant. That is $x, y, z \geq 0$. Find the coordinates of the center of mass if the density is constant. Your answer for one of the coordinates for the center of mass should be $(3/8)\, R$.

22. Show that in general for $\mathbf{L}$ angular momentum,

$$\frac{d\mathbf{L}}{dt} = \mathbf{\Gamma}$$

where $\mathbf{\Gamma}$ is the total torque,

$$\mathbf{\Gamma} \equiv \sum \mathbf{r}_i \times \mathbf{F}_i$$

where $\mathbf{F}_i$ is the force on the i^{th} point mass.

The Integral On Other Sets

33.0.1 Outcomes

1. Define the p dimensional volume.
2. Find the area of a surface.
3. Define and compute integrals over surfaces given parametrically.

33.1 The p Dimensional Volume In $\mathbb{R}^n$

Consider the boundary of some three dimensional region such that a function, f is defined on this boundary. Imagine taking the value of this function at a point, multiplying this value by the area of an infinitesimal chunk of area located at this point and then adding these up. This is just the notion of the integral presented earlier only now there is a difference because this infinitesimal chunk of area should be considered as two dimensional even though it is in three dimensions. However, it is not really all that different from what was done earlier. As before, it all depends on the following fundamental definition on Page 518.

Definition 33.1.1 *Let $\mathbf{u}_1, \cdots, \mathbf{u}_p$ be vectors in $\mathbb{R}^n$. The p dimensional parallelepiped determined by these vectors will be denoted by $P(\mathbf{u}_1, \cdots, \mathbf{u}_p)$ and it is defined as*

$$P(\mathbf{u}_1, \cdots, \mathbf{u}_p) \equiv \left\{ \sum_{j=1}^{p} s_j \mathbf{u}_j : s_j \in [0,1] \right\}.$$

Define the volume of this parallelepiped by

$$\text{volume of } P(\mathbf{u}_1, \cdots, \mathbf{u}_p) \equiv \left(\det\left(\mathbf{u}_i \cdot \mathbf{u}_j\right)\right)^{1/2}.$$

Suppose then that $\mathbf{x} = \mathbf{f}(\mathbf{u})$ where $\mathbf{u} \in U$, a subset of $\mathbb{R}^p$ and $\mathbf{x}$ is a point in V, a subset of n dimensional space where $n \geq p$. Thus, letting the Cartesian coordinates of $\mathbf{x}$ be given by $\mathbf{x} = (x_1, \cdots, x_n)^T$, each x_i being a function of $\mathbf{u}$, an infinitesimal box located at $\mathbf{u}_0$ corresponds to an infinitesimal parallelepiped located at $\mathbf{f}(\mathbf{u}_0)$ which is determined by the p vectors $\left\{ \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} du_i \right\}_{i=1}^{p}$, each of which is tangent to the surface defined by $\mathbf{x} = \mathbf{f}(\mathbf{u})$. (No sum on the repeated index.) From Definition 33.1.1, the volume of this infinitesimal parallelepiped located at $\mathbf{f}(\mathbf{u}_0)$ is given by

$$\det\left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} du_i \cdot \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_j} du_j \right)^{1/2}. \tag{33.1}$$

I like to think of this in the case where $p = 2$. In this case the infinitesimal parallelepiped is an infinitesimal parallelogram tangent to the surface defined by $\mathbf{x} = \mathbf{f}(\mathbf{u})$ like a very small scale on a lizard. This is the essence of the idea. To define the area of the lizard sum up areas of individual scales[1]. If the scales are small enough, their sum would serve as a good approximation to the area of the lizard.

Now, continuing with the general case, the matrix in the above formula is a $p \times p$ matrix. Denoting

$$\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} = \mathbf{x}_{,i}$$

to save space, this matrix is of the form

$$\overbrace{\begin{pmatrix} du_1 & 0 & \cdots & 0 \\ 0 & du_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & du_p \end{pmatrix}}^{p\times p} \overbrace{\begin{pmatrix} \mathbf{x}_{,1}^T \\ \mathbf{x}_{,2}^T \\ \vdots \\ \mathbf{x}_{,p}^T \end{pmatrix}}^{p\times n}.$$

[1]This beautiful lizard is a *Sceloporus magister*. It was photographed by C. Riley Nelson who is in the Zoology department at Brigham Young University © 2004 in Kane Co. Utah. The lizard is a little less than one foot in length.

$$\overbrace{\begin{pmatrix} \mathbf{x}_{,1} & \mathbf{x}_{,2} & \cdots & \mathbf{x}_{,p} \end{pmatrix}}^{n\times p} \overbrace{\begin{pmatrix} du_1 & 0 & \cdots & 0 \\ 0 & du_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & du_p \end{pmatrix}}^{p\times p}$$

Therefore, by the theorem which says the determinant of a product equals the product of the determinants, the determinant of the above product equals

$$\det \begin{pmatrix} du_1 & 0 & \cdots & 0 \\ 0 & du_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & du_p \end{pmatrix}^2 \det \left(\begin{pmatrix} \mathbf{x}_{,1}^T \\ \mathbf{x}_{,2}^T \\ \vdots \\ \mathbf{x}_{,p}^T \end{pmatrix} \begin{pmatrix} \mathbf{x}_{,1} & \mathbf{x}_{,2} & \cdots & \mathbf{x}_{,p} \end{pmatrix} \right) =$$

$$\det \left(\begin{pmatrix} \mathbf{x}_{,1}^T \\ \mathbf{x}_{,2}^T \\ \vdots \\ \mathbf{x}_{,p}^T \end{pmatrix} \begin{pmatrix} \mathbf{x}_{,1} & \mathbf{x}_{,2} & \cdots & \mathbf{x}_{,p} \end{pmatrix} \right) (du_1 du_2 \cdots du_p)^2$$

and so taking the square root implies the volume of the infinitesimal parallelepiped at $\mathbf{x} = \mathbf{f}(\mathbf{u}_0)$ is

$$\det \left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} \cdot \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_j} \right)^{1/2} du_1\, du_2 \cdots\, du_p =$$

$$\det \left(D\mathbf{f}(\mathbf{u})^T D\mathbf{f}(\mathbf{u}) \right)^{1/2} du_1\, du_2 \cdots\, du_p$$

Definition 33.1.2 *Let $\mathbf{x} = \mathbf{f}(\mathbf{u})$ be as described above. Then the symbol, $\frac{\partial(x_1,\cdots x_n)}{\partial(u_1,\cdots,u_p)}$, is defined by*

$$\det \left(\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} \cdot \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_j} \right)^{1/2} \equiv \frac{\partial(x_1, \cdots x_n)}{\partial(u_1, \cdots, u_p)}.$$

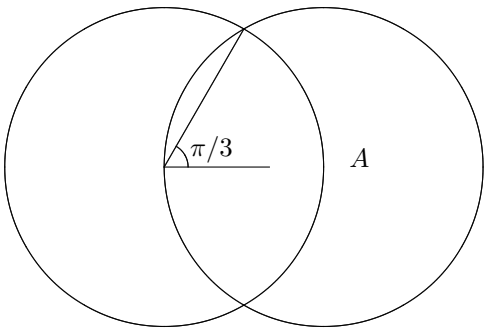
Also, the symbol, $dV_p \equiv \frac{\partial(x_1,\cdots x_n)}{\partial(u_1,\cdots,u_p)}\, du_1 \cdots du_p$ is called the volume element or area element.Note the use of the subscript, p. This indicates the p dimensional volume element. When $p = 2$ it is customary to write dA. Also, continue referring to $\frac{\partial(x_1,\cdots x_n)}{\partial(u_1,\cdots,u_p)}$ as the Jacobian.

This motivates the following fundamental procedure which I hope is extremely familiar from the earlier material.

Procedure 33.1.3 *Suppose U is a subset of $\mathbb{R}^p$ and suppose $\mathbf{f} : U \to \mathbf{f}(U) \subseteq \mathbb{R}^n$ is a one to one and C^1 function. Then if $h : \mathbf{f}(U) \to \mathbb{R}$, define the p dimensional surface integral, $\int_{\mathbf{f}(U)} h(\mathbf{x})\, dV_p$ according to the following formula.*

$$\int_{\mathbf{f}(U)} h(\mathbf{x})\, dV_p \equiv \int_U h(\mathbf{f}(\mathbf{u})) \frac{\partial(x_1, \cdots x_n)}{\partial(u_1, \cdots, u_p)}\, dV.$$

Example 33.1.4 *Find the area of the region labeled A in the following picture. The two circles are of radius 1, one has center $(0,0)$ and the other has center $(1,0)$.*



The circles bounding these disks are $x^2+y^2=1$ and $(x-1)^2+y^2=x^2+y^2-2x+1=1$. Therefore, in polar coordinates these are of the form $r=1$ and $r=2\cos\theta$.

The set A corresponds to the set U, in the (θ,r) plane determined by $\theta\in\left[-\frac{\pi}{3},\frac{\pi}{3}\right]$ and for each value of θ in this interval, r goes from 1 up to $2\cos\theta$. Therefore, the area of this region is of the form,

$$\int_U 1\,dV=\int_{-\pi/3}^{\pi/3}\int_1^{2\cos\theta}\frac{\partial(x_1,x_2)}{\partial(\theta,r)}\,dr\,d\theta.$$

It is necessary to find $\frac{\partial(x_1,x_2)}{\partial(\theta,r)}$. The mapping $\mathbf{f}:U\to\mathbb{R}^2$ takes the form $\mathbf{f}(\theta,r)=(r\cos\theta,r\sin\theta)^T$ and so

$$D\mathbf{f}(\theta,r)=\begin{pmatrix}-r\sin\theta & \cos\theta\\ r\cos\theta & \sin\theta\end{pmatrix}$$

and so

$$D\mathbf{f}(\theta,r)^T\,D\mathbf{f}(\theta,r)=\begin{pmatrix}r^2 & 0\\ 0 & 1\end{pmatrix}$$

which implies

$$\frac{\partial(x_1,x_2)}{\partial(\theta,r)}=\det\left(D\mathbf{f}(\theta,r)^T\,D\mathbf{f}(\theta,r)\right)^{1/2}=r.$$

Therefore, the area element is $r\,dr\,d\theta$. It follows the desired area is

$$\int_{-\pi/3}^{\pi/3}\int_1^{2\cos\theta}r\,dr\,d\theta=\frac{1}{2}\sqrt{3}+\frac{1}{3}\pi.$$

Example 33.1.5 *Consider the surface given by $z=x^2$ for $(x,y)\in[0,1]\times[0,1]=U$. Find the surface area of this surface.*

The first step in using the above is to write this surface in the form $\mathbf{x}=\mathbf{f}(\mathbf{u})$. This is easy to do if you let $\mathbf{u}=(x,y)$. Then $\mathbf{f}(x,y)=\left(x,y,x^2\right)$. If you like, let $x=u_1$ and $y=u_2$. What is $\frac{\partial(x_1,x_2,x_3)}{\partial(x,y)}$?

$$D\mathbf{f}(x,y)=\begin{pmatrix}1 & 0\\ 0 & 1\\ 2x & 0\end{pmatrix}$$

and so

$$D\mathbf{f}(x,y)^T=\begin{pmatrix}1 & 0 & 2x\\ 0 & 1 & 0\end{pmatrix}$$

Thus in this case,

$$D\mathbf{f}(\mathbf{u})^T D\mathbf{f}(\mathbf{u}) = \begin{pmatrix} 1+4x^2 & 0 \\ 0 & 1 \end{pmatrix}$$

and so the area element is $\sqrt{1+4x^2}\,dx\,dy$ and the surface area is obtained by integrating the function, $h(\mathbf{x}) \equiv 1$. Therefore, this area is

$$\int_U dV = \int_0^1 \int_0^1 \sqrt{1+4x^2}\,dx\,dy = \frac{1}{2}\sqrt{5} - \frac{1}{4}\ln\left(-2+\sqrt{5}\right)$$

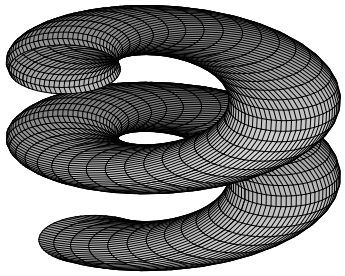
which can be obtained by using the trig. substitution, $2x = \tan\theta$ on the inside integral.

Note this all depends on being able to write the surface in the form, $\mathbf{x} = \mathbf{f}(\mathbf{u})$ for $\mathbf{u} \in U \subseteq \mathbb{R}^p$. Surfaces obtained in this form are called parametrically defined surfaces. These are best but sometimes you have some other description of a surface and in these cases things can get pretty intractable. For example, you might have a level surface of the form $3x^2 + 4y^4 + z^6 = 10$. In this case, you could solve for z using methods of algebra. Thus $z = \sqrt[6]{10 - 3x^2 - 4y^4}$ and a parametric description of part of this level surface is $\left(x, y, \sqrt[6]{10-3x^2-4y^4}\right)$ for $(x,y) \in U$ where $U = \left\{(x,y) : 3x^2 + 4y^4 \leq 10\right\}$. But what if the level surface was something like

$$\sin\left(x^2 + \ln\left(7 + y^2\sin x\right)\right) + \sin(zx)\,e^z = 11\sin(xyz)?$$

I really don't see how to use methods of algebra to solve for some variable in terms of the others. It isn't even clear to me whether there are any points $(x,y,z) \in \mathbb{R}^3$ satisfying this particular relation. However, if a point satisfying this relation can be identified, the implicit function theorem from advanced calculus can usually be used to assert one of the variables is a function of the others, proving the existence of a parameterization at least locally. However, this theorem doesn't give us the answer in terms of known functions so this isn't much help. Finding a parametric description of a surface is a hard problem and there are no easy answers.

Example 33.1.6 *Let $U = [0,12] \times [0,2\pi]$ and let $\mathbf{f} : U \to \mathbb{R}^3$ be given by $\mathbf{f}(t,s) \equiv (2\cos t + \cos s, 2\sin t + \sin s, t)^T$. Find a double integral for the surface area. A graph of this surface is drawn below.*



It looks like something you would use to make sausages[2]. Anyway,

$$D\mathbf{f}(t,s) = \begin{pmatrix} -2\sin t & -\sin s \\ 2\cos t & \cos s \\ 1 & 0 \end{pmatrix}$$

[2]At Volwerth's in Hancock Michigan, they make excellent sausages and hot dogs. The best are made from "natural casings" which are the linings of intestines.

and so

$$D\mathbf{f}(t,s)^T D\mathbf{f}(t,s) = \begin{pmatrix} 5 & 2\sin t\sin s + 2\cos t\cos s \\ 2\sin t\sin s + 2\cos t\cos s & 1 \end{pmatrix}$$

and

$$\begin{aligned} \left(\frac{\partial(x_1,x_2,x_3)}{\partial(t,s)}\right)^2 &= \det\begin{pmatrix} 5 & 2\sin t\sin s + 2\cos t\cos s \\ 2\sin t\sin s + 2\cos t\cos s & 1 \end{pmatrix} \\ &= 5 - 4\sin^2 t\sin^2 s - 8\sin t\sin s\cos t\cos s - 4\cos^2 t\cos^2 s \end{aligned}$$

which implies the area equals

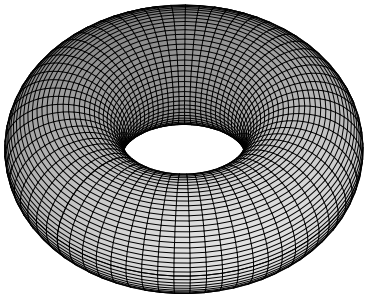
$$\int_0^{2\pi}\int_0^{12} \sqrt{5 - 4\sin^2 t\sin^2 s - 8\sin t\sin s\cos t\cos s - 4\cos^2 t\cos^2 s}\, dt\, ds.$$

If you really needed to find the number this equals, how would you go about finding it? This is an interesting question and there is no single right answer. You should think about this. Here is an example for which you will be able to find the integrals.

Example 33.1.7 *Let $U = [0,2\pi]\times[0,2\pi]$ and for $(t,s)\in U$, let*

$$\mathbf{f}(t,s) = (2\cos t + \cos t\cos s, -2\sin t - \sin t\cos s, \sin s)^T.$$

Find the area of $\mathbf{f}(U)$. This is the surface of a donut shown below. The fancy name for this shape is a torus.



To find its area,

$$D\mathbf{f}(t,s) = \begin{pmatrix} -2\sin t - \sin t\cos s & -\cos t\sin s \\ -2\cos t - \cos t\cos s & \sin t\sin s \\ 0 & \cos s \end{pmatrix}$$

and so

$$D\mathbf{f}(t,s)^T D\mathbf{f}(t,s) = \begin{pmatrix} 4 + 4\cos s + \cos^2 s & 0 \\ 0 & 1 \end{pmatrix}$$

which implies the area element is

$$\begin{aligned} \det\begin{pmatrix} 4 + 4\cos s + \cos^2 s & 0 \\ 0 & 1 \end{pmatrix}^{1/2} ds\, dt &= \left(4 + 4\cos s + \cos^2 s\right)^{1/2} ds\, dt \\ &= (\cos s + 2)\, ds\, dt \end{aligned}$$

and the area is

$$\int_0^{2\pi}\int_0^{2\pi}(\cos s+2)\,ds\,dt = 8\pi^2$$

Example 33.1.8 *Let $U=[0,2\pi]\times[0,2\pi]$ and for $(t,s)\in U$, let*

$$\mathbf{f}(t,s) = (2\cos t+\cos t\cos s, -2\sin t-\sin t\cos s, \sin s)^T.$$

Find

$$\int_{\mathbf{f}(U)} h\,dV$$

where $h(x,y,z)=x^2$.

Everything is the same as the preceding example except this time it is an integral of a function. The area element is $(\cos s+2)\,ds\,dt$ and so the integral called for is

$$\int_{\mathbf{f}(U)} h\,dV = \int_0^{2\pi}\int_0^{2\pi}\left(\overbrace{2\cos t+\cos t\cos s}^{x \text{ on the surface}}\right)^2(\cos s+2)\,ds\,dt = 22\pi^2$$

Example 33.1.9 *Let $U=\left\{(x,y,z): x^2+y^2+z^2\le 4\right\}$ and for $(x,y,z)\in U$ let $\mathbf{f}(x,y,z) = (x,y,x+y,z)$. Find the three dimensional volume of $\mathbf{f}(U)$.*

Note there is no picture here because I am unable to draw one in four dimensions. Nevertheless it is a three dimensional volume which is being computed. Everything is done the same as before.

$$D\mathbf{f}(x,y,z) = \begin{pmatrix} 1 & 0 & 0\\ 0 & 1 & 0\\ 1 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}$$

and so

$$D\mathbf{f}(x,y,z)^T D\mathbf{f}(x,y,z) = \begin{pmatrix} 2 & 1 & 0\\ 1 & 2 & 0\\ 0 & 0 & 1\end{pmatrix}$$

and so the volume element is $3\,dx\,dy\,dz$. Therefore, the volume of $\mathbf{f}(U)$ is

$$\int_U 3\,dx\,dy\,dz = 3\left(\frac{4}{3}\pi(8)\right) = 32\pi.$$

The special case where a surface is in the form $z=f(x,y), (x,y)\in U$, yields a simple formula which is used most often in this situation. You write the surface parametrically in the form $\mathbf{f}(x,y) = (x,y,f(x,y))^T : (x,y)\in U$. Then

$$D\mathbf{f}(x,y) = \begin{pmatrix} 1 & 0\\ 0 & 1\\ f_x & f_y\end{pmatrix}$$

and so

$$D\mathbf{f}(x,y,z)^T D\mathbf{f}(x,y,z) = \begin{pmatrix} 1+f_x^2 & f_xf_y\\ f_xf_y & 1+f_y^2\end{pmatrix}.$$

Thus,

$$\det\begin{pmatrix} 1+f_x^2 & f_xf_y\\ f_xf_y & 1+f_y^2\end{pmatrix} = 1+f_y^2+f_x^2$$

and so the area element is

$$\sqrt{1+f_y^2+f_x^2}\,dx\,dy.$$

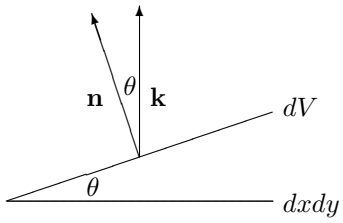
When the surface of interest comes in this simple form, people generally use this area element directly rather than worrying about a parameterization and taking determinants and finding matrices.

In the case where the surface is of the form $x = f(y,z)$ for $(y,z) \in U$, the area element is obtained similarly and is

$$\sqrt{1+f_y^2+f_z^2}\,dy\,dz.$$

I think you can guess what the area element is if $y = f(x,z)$.

There is also a simple geometric description of these area elements. Consider the surface $z = f(x,y)$. This is a level surface of the function of three variables $z - f(x,y)$. In fact the surface is simply $z - f(x,y) = 0$. Now consider the gradient of this function of three variables. The gradient is perpendicular to the surface and the third component is positive in this case. This gradient is $(-f_x, -f_y, 1)$ and so the unit upward normal is just $\frac{1}{\sqrt{1+f_x^2+f_y^2}}(-f_x,-f_y,1)$. Now consider the following picture.



In this picture, you are looking at a chunk of area on the surface seen on edge and so it seems reasonable to expect to have $dx\,dy = dV\cos\theta$. But it is easy to find $\cos\theta$ from the picture and the properties of the dot product.

$$\cos\theta = \frac{\mathbf{n}\cdot\mathbf{k}}{|\mathbf{n}|\,|\mathbf{k}|} = \frac{1}{\sqrt{1+f_x^2+f_y^2}}.$$

Therefore, $dV = \sqrt{1+f_x^2+f_y^2}\,dx\,dy$ as claimed. In this context, the surface involved is referred to as S because the vector valued function, $\mathbf{f}$ giving the parameterization will not have been identified.

Example 33.1.10 *Let $z = \sqrt{x^2+y^2}$ where $(x,y) \in U$ for $U = \{(x,y) : x^2+y^2 \le 4\}$ Find*

$$\int_S h\,dV$$

where $h(x,y,z) = x+z$ and S is the surface described as $\left(x, y, \sqrt{x^2+y^2}\right)$ for $(x,y) \in U$.

Here you can see directly the angle in the above picture is $\frac{\pi}{4}$ and so $dV = \sqrt{2}\,dx\,dy$. If you don't see this or if it is unclear, simply compute $\sqrt{1+f_x^2+f_y^2}$ and you will find it is

$\sqrt{2}$. Therefore, using polar coordinates,

$$\begin{aligned}\int_S h\,dV &= \int_U \left(x+\sqrt{x^2+y^2}\right)\sqrt{2}\,dV \\ &= \sqrt{2}\int_0^{2\pi}\int_0^2 (r\cos\theta + r)\,r\,dr\,d\theta \\ &= \frac{16}{3}\sqrt{2}\pi.\end{aligned}$$

One other issue is worth mentioning. Suppose $\mathbf{f}_i : U_i \to \mathbb{R}^n$ where U_i are sets in $\mathbb{R}^p$ and suppose $\mathbf{f}_1(U_1)$ intersects $\mathbf{f}_2(U_2)$ along C where $C = \mathbf{h}(V)$ for $V \subseteq \mathbb{R}^k$ for $k < p$. Then define integrals and areas over $\mathbf{f}_1(U_1) \cup \mathbf{f}_2(U_2)$ as follows.

$$\int_{\mathbf{f}_1(U_1)\cup\mathbf{f}_2(U_2)} g\,dV_p \equiv \int_{\mathbf{f}_1(U_1)} g\,dV_p + \int_{\mathbf{f}_2(U_2)} g\,dV_p.$$

Admittedly, the set C gets added in twice but this doesn't matter because its p dimensional volume equals zero and therefore, the integrals over this set will also be zero. Why is this? To find the p dimensional volume element on C, it is necessary to find a function, $\mathbf{f}$, mapping $U \subseteq \mathbb{R}^p$ to C. Let $\mathbf{f}(\mathbf{v}, s_1, \cdots, s_{p-k}) \equiv \mathbf{h}(\mathbf{v})$. Then $D\mathbf{f}(\mathbf{v}, s_1, \cdots, s_{p-k})$ has at least one column of zeros and so $\det\left(D\mathbf{f}^T D\mathbf{f}\right) = 0$ showing the p dimensional volume element is zero and so this makes no contribution to the integral as claimed. Clearly something similar holds in the case of many surfaces joined in this way.

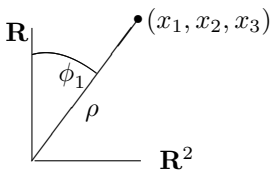
I have been purposely vague about precise mathematical conditions necessary for the above procedures. This is because the precise mathematical conditions which are usually cited are very technical and at the same time far too restrictive. The most general conditions under which these sorts of procedures are valid include things like Lipschitz functions defined on very general sets. These are functions satisfying a Lipschitz condition of the form $|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq K|\mathbf{x}-\mathbf{y}|$. For example, $y = |x|$ is Lipschitz continuous. However, this function does not have a derivative at every point. So it is with Lipschitz functions. However, it turns out these functions have derivatives at enough points to push everything through but this requires considerations involving the Lebesgue integral. Lipschitz functions are also not the most general kind of function for which the above is valid.

33.2 Spherical Coordinates In $\mathbb{R}^n$

Recall polar coordinates are of the form

$$\begin{aligned}x_1 &= \rho\cos\theta \\ x_2 &= \rho\sin\theta\end{aligned}$$

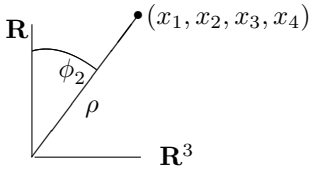
where $\rho > 0$ and $\theta \in [0, 2\pi)$. Here I am writing ρ in place of r to emphasize a pattern which is about to emerge. I will consider polar coordinates as spherical coordinates in two dimensions. This is also the reason I am writing x_1 and x_2 instead of the more usual x and y. Now consider what happens when you go to three dimensions. The situation is depicted in the following picture.



From this picture, you see that $x_3 = \rho\cos\phi_1$. Also the distance between (x_1, x_2) and $(0,0)$ is $\rho\sin(\phi_1)$. Therefore, using polar coordinates to write (x_1, x_2) in terms of θ and this distance,

$$\begin{aligned} x_1 &= \rho\sin\phi_1\cos\theta, \\ x_2 &= \rho\sin\phi_1\sin\theta, \\ x_3 &= \rho\cos\phi_1. \end{aligned}$$

where $\phi_1 \in [0,\pi]$. What was done is to replace ρ with $\rho\sin\phi_1$ and then to add in $x_3 = \rho\cos\phi_1$. Having done this, there is no reason to stop with three dimensions. Consider the following picture:



From this picture, you see that $x_4 = \rho\cos\phi_2$. Also the distance between (x_1, x_2, x_3) and $(0,0,0)$ is $\rho\sin(\phi_2)$. Therefore, using polar coordinates to write (x_1, x_2, x_3) in terms of θ, ϕ_1, and this distance,

$$\begin{aligned} x_1 &= \rho\sin\phi_2\sin\phi_1\cos\theta, \\ x_2 &= \rho\sin\phi_2\sin\phi_1\sin\theta, \\ x_3 &= \rho\sin\phi_2\cos\phi_1, \\ x_4 &= \rho\cos\phi_2 \end{aligned}$$

where $\phi_2 \in [0,\pi]$.

Continuing this way, given spherical coordinates in $\mathbb{R}^n$, to get the spherical coordinates in $\mathbb{R}^{n+1}$, you let $x_{n+1} = \rho\cos\phi_{n-1}$ and then replace every occurance of ρ with $\rho\sin\phi_{n-1}$ to obtain $x_1\cdots x_n$ in terms of $\phi_1, \phi_2, \cdots, \phi_{n-1}$,$\theta$, and ρ.

For spherical coordinates, it is always the case that ρ measures the distance from the point in $\mathbb{R}^n$ to the origin in $\mathbb{R}^n$, $\mathbf{0}$. Each $\phi_i \in [0,\pi]$, and $\theta \in [0, 2\pi)$. I leave it as an exercise using math induction to prove that these coordinates map $\prod_{i=1}^{n-2}[0,\pi]\times[0,2\pi)\times(0,\infty)$ one to one onto $\mathbb{R}^n \setminus \{\mathbf{0}\}$.

It is customary to write S^{n-1} for the set $\{\mathbf{x}\in\mathbb{R}^n : |\mathbf{x}| = 1\}$. Thus a parameterization for this level surface is given by letting $\rho = 1$ in spherical coordinates. I will denote by $S^{n-1}(a)$ the sphere having radius $a > 0$. What would the $n-1$ dimensional volume element on $S^{n-1}(a)$ be? For $S^1(a)$, there is only one parameter, θ. Therefore, the one dimensional volume element is

$$\det\left((-a\sin\theta, a\cos\theta)\begin{pmatrix} -a\sin\theta \\ a\cos\theta \end{pmatrix}\right)^{1/2} d\theta = a\,d\theta$$

where $\theta \in [0, 2\pi)$.

Next consider S^2. In this case the two dimensional volume element is

$$\begin{aligned} &\det\left(\begin{array}{c} \begin{pmatrix} -a\sin\phi_1\sin\theta & a\sin\phi_1\cos\theta & 0 \\ a\cos\phi_1\cos\theta & a\cos\phi_1\sin\theta & -a\sin\phi_1 \end{pmatrix} \\ \cdot \begin{pmatrix} -a\sin\phi_1\sin\theta & a\cos\phi_1\cos\theta \\ a\sin\phi_1\cos\theta & a\cos\phi_1\sin\theta \\ 0 & -a\sin\phi_1 \end{pmatrix} \end{array}\right)^{1/2} d\phi_1 d\theta \\ &= \det\begin{pmatrix} a^2\sin^2\phi_1 & 0 \\ 0 & a^2 \end{pmatrix}^{1/2} d\phi_1 d\theta = a^2(\sin\phi_1)\,d\phi_1 d\theta \end{aligned}$$

What of $S^3(a) \equiv \{\mathbf{x} \in \mathbb{R}^4 : |\mathbf{x}| = a\}$ and $S^n(a) = \{\mathbf{x} \in \mathbb{R}^{n+1} : |\mathbf{x}| = a\}$. Let $\mathbf{x}_n$ denote the vector $(x_1, \cdots, x_n)$. That is, $\mathbf{x}_n$ consists of the first n components of $\mathbf{x}$. Let $D_{\theta\cdots\phi_{n-2}}$ denote the derivative with respect to the vector, $(\theta, \phi_1 \cdots \phi_{n-2})$. Then the volume element on $S^n(a)$ is of the form

$$\det\left(\begin{array}{c}\left(\begin{array}{cc}\left(\sin(\phi_{n-1})\left(D_{\theta\cdots\phi_{n-2}}\mathbf{x}_n\right)_{n\times n-1}\right)^T_{(n-1)\times n} & (0)_{(n-1)\times 1}\\ (*)_{1\times n} & -a\sin\phi_{n-1}\end{array}\right)\\ \cdot\left(\begin{array}{cc}\sin(\phi_{n-1})\left(D_{\theta\cdots\phi_{n-2}}\mathbf{x}_n\right)_{n\times(n-1)} & (*)_{n\times 1}\\ (0)_{1\times n-1} & -a\sin\phi_{n-1}\end{array}\right)\end{array}\right)^{1/2} d\phi_{n-1}\cdots d\phi_1 d\theta$$

Now using block multiplication, this reduces to

$$a\sin(\phi_{n-1})\det\left(\left(\left(D_{\theta\cdots\phi_{n-2}}\mathbf{x}_n\right)_{n\times n-1}\right)^T_{(n-1)\times n}\left(D_{\theta\cdots\phi_{n-2}}\mathbf{x}_n\right)_{n\times(n-1)}\right)^{1/2} d\phi_{n-1}\cdots d\phi_1 d\theta.$$

That is, to get the volume element in $S^n(a)$, you multiply the volume element on $S^{n-1}(a)$ by $a\sin(\phi_{n-1})\,d\phi_{n-1}$. Consequently, beginning with the volume element on $S^1(a)$, you obtain the succession of volume elements for $S^1(a), S^2(a), S^3(a), S^4(a)$.

$$ad\theta, a^2\sin\phi_1 d\phi_1 d\theta, a^3\sin\phi_2\sin\phi_1 d\phi_2 d\phi_1 d\theta,$$
$$a^4\sin\phi_3\sin\phi_2\sin\phi_1 d\phi_3 d\phi_2 d\phi_1 d\theta, \text{ etc.}$$

In general, the n dimensional volume element on $S^n(a)$ is

$$a^n\left(\prod_{i=1}^{n-1}\sin\phi_i\right)\left(\prod_{i=1}^{n-1}d\phi_i\right)d\theta \tag{33.2}$$

Using similar reasoning, the n dimensional volume element in terms of the spherical coordinates is

$$\rho^{n-1}\left(\prod_{i=1}^{n-2}\sin\phi_i\right)\left(\prod_{i=1}^{n-2}d\phi_i\right)d\theta d\rho. \tag{33.3}$$

Formulas (33.2) and (33.3) are very useful in estimating integrals.

Example 33.2.1 *For what values of s is the integral $\int_{B(\mathbf{0},R)}\left(1+|\mathbf{x}|^2\right)^s dV$ bounded independent of R? Here $B(\mathbf{0},R)$ is the ball, $\{\mathbf{x}\in\mathbb{R}^n : |\mathbf{x}|\le R\}$.*

I think you can see immediately that s must be negative but exactly how negative? It turns out it depends on n and using spherical coordinates, you can find just exactly what is needed. It is very hard to overstate the importance of the technique I am about to show you. Write the above integral in n dimensional spherical coordinates.

$$\begin{aligned}&\int_0^R\int_0^\pi\cdots\int_0^\pi\int_0^{2\pi}\rho^{n-1}\left(1+\rho^2\right)^s\left(\prod_{i=1}^{n-2}\sin\phi_i\right)\left(\prod_{i=1}^{n-2}d\phi_i\right)d\theta d\rho\\ =\;&\int_0^R\rho^{n-1}\left(1+\rho^2\right)^s\int_{S^{n-1}}dS^{n-1}d\rho=\omega_n\int_0^R\rho^{n-1}\left(1+\rho^2\right)^s d\rho\end{aligned}$$

where $dS^{n-1} = \left(\prod_{i=1}^{n-2} \sin\phi_i\right)\left(\prod_{i=1}^{n-2} d\phi_i\right) d\theta$.

$$\omega_n \equiv \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} \left(\prod_{i=1}^{n-2} \sin\phi_i\right)\left(\prod_{i=1}^{n-2} d\phi_i\right) d\theta$$

and from the above explanation this equals the area of S^{n-1}, the unit sphere, $\{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x}| = 1\}$. Now the very hard problem has been reduced to considering an easy one variable problem of finding when

$$\int_0^R \rho^{n-1}\left(1+\rho^2\right)^s d\rho$$

is bounded independent of R. I leave it to you to verify using standard one variable calculus that you need $2s + (n-1) < -1$ so you need $s < -n/2$.

33.3 Exercises With Answers

1. Find a parameterization for the intersection of the planes $x + y + 2z = -3$ and $2x - y + z = -4$.

   Answer:

   $(x, y, z) = \left(-t - \frac{7}{3}, -t - \frac{2}{3}, t\right)$

2. Find a parameterization for the intersection of the plane $4x + 2y + 4z = 0$ and the circular cylinder $x^2 + y^2 = 16$.

   Answer:

   The cylinder is of the form $x = 4\cos t, y = 4\sin t$ and $z = z$. Therefore, from the equation of the plane, $16\cos t + 8\sin t + 4z = 0$. Therefore, $z = -16\cos t - 8\sin t$ and this shows the parameterization is of the form $(x, y, z) = (4\cos t, 4\sin t, -16\cos t - 8\sin t)$ where $t \in [0, 2\pi]$.

3. Find a parameterization for the intersection of the plane $3x + 2y + z = 4$ and the elliptic cylinder $x^2 + 4z^2 = 1$.

   Answer:

   The cylinder is of the form $x = \cos t, 2z = \sin t$ and $y = y$. Therefore, from the equation of the plane, $3\cos t + 2y + \frac{1}{2}\sin t = 4$. Therefore, $y = 2 - \frac{3}{2}\cos t - \frac{1}{4}\sin t$ and this shows the parameterization is of the form $(x, y, z) = \left(\cos t, 2 - \frac{3}{2}\cos t - \frac{1}{4}\sin t, \frac{1}{2}\sin t\right)$ where $t \in [0, 2\pi]$.

4. Find a parameterization for the straight line joining $(4, 3, 2)$ and $(1, 7, 6)$.

   Answer:

   $(x, y, z) = (4, 3, 2) + t(-3, 4, 4) = (4 - 3t, 3 + 4t, 2 + 4t)$ where $t \in [0, 1]$.

5. Find a parameterization for the intersection of the surfaces $y + 3z = 4x^2 + 4$ and $4y + 4z = 2x + 4$.

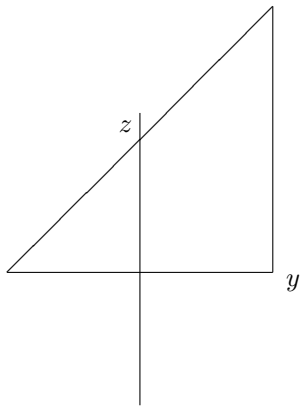
   Answer:

   This is an application of Cramer's rule. $y = -2x^2 - \frac{1}{2} + \frac{3}{4}x, z = -\frac{1}{4}x + \frac{3}{2} + 2x^2$. Therefore, the parameterization is $(x, y, z) = \left(t, -2t^2 - \frac{1}{2} + \frac{3}{4}t, -\frac{1}{4}t + \frac{3}{2} + 2t^2\right)$.

6. Find the area of S if S is the part of the circular cylinder $x^2 + y^2 = 16$ which lies between $z = 0$ and $z = 4 + y$.

   Answer:

   Use the parameterization, $x = 4\cos v, y = 4\sin v$ and $z = u$ with the parameter domain described as follows. The parameter, v goes from $-\frac{\pi}{2}$ to $\frac{3\pi}{2}$ and for each v in this interval, u should go from 0 to $4+4\sin v$. To see this observe that the cylinder has its axis parallel to the z axis and if you look at a side view of the surface you would see something like this:



   The positive x axis is coming out of the paper toward you in the above picture and the angle v is the usual angle measured from the positive x axis. Therefore, the area is just $A = \int_{-\pi/2}^{3\pi/2}\int_0^{4+4\sin v} 4\,du\,dv = 32\pi$.

7. Find the area of S if S is the part of the cone $x^2 + y^2 = 9z^2$ between $z = 0$ and $z = h$.

   Answer:

   When $z = h$, $x^2 + y^2 = 9h^2$ which is the boundary of a circle of radius ah. A parameterization of this surface is $x = u, y = v, z = \frac{1}{3}\sqrt{(u^2+v^2)}$ where $(u,v) \in D$, a disk centered at the origin having radius ha. Therefore, the volume is just $\int\int_D \sqrt{1+z_u^2+z_v^2}\,dA = \int_{-ha}^{ha}\int_{-\sqrt{(9h^2-u^2)}}^{\sqrt{(9h^2-u^2)}} \frac{1}{3}\sqrt{10}\,dv\,du = 3\pi h^2\sqrt{10}$

8. Parametrizing the cylinder $x^2 + y^2 = 4$ by $x = 2\cos v, y = 2\sin v, z = u$, show that the area element is $dA = 2\,du\,dv$

   Answer:

   It is necessary to compute $\frac{\partial(x,y,z)}{\partial(u,v)} = \det\left(D\mathbf{f}^T D\mathbf{f}\right)$.

$$D\mathbf{f}(u,v) = \begin{pmatrix} 0 & -2\sin v \\ 0 & 2\cos v \\ 1 & 0 \end{pmatrix}$$

   and so $D\mathbf{f}^T D\mathbf{f} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$ and so the area element is as described.

9. Find the area enclosed by the limacon $r = 2 + \cos\theta$.

   Answer:

You can graph this region and you see it is sort of an oval shape and that $\theta \in [0, 2\pi]$ while r goes from 0 up to $2 + \cos\theta$. Now $x = r\cos\theta$ and $y = r\sin\theta$ are the x and y coordinates corresponding to r and θ in the above parameter domain. Therefore, the area of the limacon equals $\int\int_P \left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| dr\,d\theta = \int_0^{2\pi}\int_0^{2+\cos\theta} r\,dr\,d\theta$ because the Jacobian equals r in this case. Therefore, the area equals $\int_0^{2\pi}\int_0^{2+\cos\theta} r\,dr\,d\theta = \frac{9}{2}\pi$.

10. Find the surface area of the paraboloid $z = h\left(1 - x^2 - y^2\right)$ between $z = 0$ and $z = h$.

    Answer:

    Let R denote the unit circle. Then the area of the surface above this circle would be $\int\int_R \sqrt{1 + 4x^2h^2 + 4y^2h^2}\,dA$. Changing to polar coordinates, this becomes

    $\int_0^{2\pi}\int_0^1 \sqrt{1 + 4h^2r^2}r\,dr\,d\theta = \frac{1}{6}\pi\frac{\sqrt{(1+4h^2)}+4\sqrt{(1+4h^2)}h^2-1}{h^2}$.

11. Evaluate $\int\int_S (1 + x)\,dA$ where S is the part of the plane $2x + 3y + 3z = 18$ which is in the first octant.

    Answer:

    $\int_0^6\int_0^{6-\frac{2}{3}x}(1 + x)\,\frac{1}{3}\sqrt{22}\,dy\,dx = 28\sqrt{22}$

12. Evaluate $\int\int_S (1 + x)\,dA$ where S is the part of the cylinder $x^2 + y^2 = 16$ between $z = 0$ and $z = h$.

    Answer:

    Parametrize the cylinder as $x = 4\cos\theta$ and $y = 4\sin\theta$ while $z = t$ and the parameter domain is just $[0, 2\pi] \times [0, h]$. Then the integral to evaluate would be

    $$\int_0^{2\pi}\int_0^h (1 + 4\cos\theta)\,4\,dt\,d\theta = 8h\pi.$$

    Note how $4\cos\theta$ was substituted for x and the area element is $4\,dt\,d\theta$.

13. Evaluate $\int\int_S (1 + x)\,dA$ where S is the hemisphere $x^2 + y^2 + z^2 = 16$ between $x = 0$ and $x = 4$.

    Answer:

    Parametrize the sphere as $x = 4\sin\phi\cos\theta, y = 4\sin\phi\sin\theta$, and $z = 4\cos\phi$ and consider the values of the parameters. Since it is referred to as a hemisphere and involves $x > 0$, $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and $\phi \in [0, \pi]$. Then the area element is $\sqrt{a^4\sin\phi}\,d\theta\,d\phi$ and so the integral to evaluate is

    $$\int_0^{\pi}\int_{-\pi/2}^{\pi/2} (1 + 4\sin\phi\cos\theta)\,16\sin\phi\,d\theta\,d\phi = 96\pi$$

14. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let

    $$\mathbf{f}(\theta, \alpha) \equiv (\cos\theta(2 + \cos\alpha), -\sin\theta(2 + \cos\alpha), \sin\alpha)^T.$$

    Find the area of $\mathbf{f}([0, 2\pi] \times [0, 2\pi])$.

    Answer:

    $D\mathbf{f}(\theta, \alpha) = \begin{pmatrix} -\sin(\theta)(2 + \cos\alpha) & -\cos\theta\sin\alpha \\ -\cos(\theta)(2 + \cos\alpha) & \sin\theta\sin\alpha \\ 0 & \cos\alpha \end{pmatrix}$ and so the area element is

    $$\det\left(D\mathbf{f}^T D\mathbf{f}\right)^{1/2} d\theta\,d\alpha = \left(4 + 4\cos\alpha + \cos^2\alpha\right)^{1/2} d\theta\,d\alpha.$$

Therefore, the area is

$$\int_0^{2\pi}\int_0^{2\pi}\left(4+4\cos\alpha+\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha=\int_0^{2\pi}\int_0^{2\pi}\left(2+\cos\alpha\right)\,d\theta\,d\alpha=8\pi^2.$$

15. For $(\theta,\alpha)\in[0,2\pi]\times[0,2\pi]$, let

$$\mathbf{f}(\theta,\alpha)\equiv\left(\cos\theta\left(4+2\cos\alpha\right),-\sin\theta\left(4+2\cos\alpha\right),2\sin\alpha\right)^T.$$

Also let $h(\mathbf{x})=\cos\alpha$ where α is such that

$$\mathbf{x}=\left(\cos\theta\left(4+2\cos\alpha\right),-\sin\theta\left(4+2\cos\alpha\right),2\sin\alpha\right)^T.$$

Find $\int_{\mathbf{f}([0,2\pi]\times[0,2\pi])}h\,dV$.

Answer:

$D\mathbf{f}(\theta,\alpha)=\begin{pmatrix}-\sin(\theta)(4+2\cos\alpha) & -2\cos\theta\sin\alpha\\ -\cos(\theta)(4+2\cos\alpha) & 2\sin\theta\sin\alpha\\ 0 & 2\cos\alpha\end{pmatrix}$ and so the area element is

$$\det\left(D\mathbf{f}^TD\mathbf{f}\right)^{1/2}\,d\theta\,d\alpha=\left(64+64\cos\alpha+16\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha.$$

Therefore, the desired integral is

$$\int_0^{2\pi}\int_0^{2\pi}\left(\cos\alpha\right)\left(64+64\cos\alpha+16\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha$$

$$=\int_0^{2\pi}\int_0^{2\pi}\left(\cos\alpha\right)\left(8+4\cos\alpha\right)\,d\theta\,d\alpha=8\pi^2$$

16. For $(\theta,\alpha)\in[0,2\pi]\times[0,2\pi]$, let

$$\mathbf{f}(\theta,\alpha)\equiv\left(\cos\theta\left(3+\cos\alpha\right),-\sin\theta\left(3+\cos\alpha\right),\sin\alpha\right)^T.$$

Also let $h(\mathbf{x})=\cos^2\theta$ where θ is such that

$$\mathbf{x}=\left(\cos\theta\left(3+\cos\alpha\right),-\sin\theta\left(3+\cos\alpha\right),\sin\alpha\right)^T.$$

Find $\int_{\mathbf{f}([0,2\pi]\times[0,2\pi])}h\,dV$.

Answer:

$D\mathbf{f}(\theta,\alpha)=\begin{pmatrix}-\sin(\theta)(3+\cos\alpha) & -\cos\theta\sin\alpha\\ -\cos(\theta)(3+\cos\alpha) & \sin\theta\sin\alpha\\ 0 & \cos\alpha\end{pmatrix}$ and so the area element is

$$\det\left(D\mathbf{f}^TD\mathbf{f}\right)^{1/2}\,d\theta\,d\alpha=\left(9+6\cos\alpha+\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha.$$

Therefore, the desired integral is

$$\int_0^{2\pi}\int_0^{2\pi}\left(\cos^2\theta\right)\left(9+6\cos\alpha+\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha$$

$$=\int_0^{2\pi}\int_0^{2\pi}\left(\cos^2\theta\right)\left(3+\cos\alpha\right)\,d\theta\,d\alpha=6\pi^2$$

17. For $(\theta,\alpha)\in[0,25]\times[0,2\pi]$, let

$$\mathbf{f}(\theta,\alpha)\equiv(\cos\theta(4+2\cos\alpha),-\sin\theta(4+2\cos\alpha),2\sin\alpha+\theta)^T.$$

Find a double integral which gives the area of $\mathbf{f}([0,25]\times[0,2\pi])$.

Answer:

In this case, $D\mathbf{f}(\theta,\alpha)=\begin{pmatrix}-\sin(\theta)(4+2\cos\alpha) & -2\cos\theta\sin\alpha\\ -\cos(\theta)(4+2\cos\alpha) & 2\sin\theta\sin\alpha\\ 1 & 2\cos\alpha\end{pmatrix}$ and so the area element is

$\det\left(D\mathbf{f}^TD\mathbf{f}\right)\,d\theta\,d\alpha=\left(68+64\cos\alpha+12\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha$ and so the surface area is

$\int_0^{2\pi}\int_0^{2\pi}\left(68+64\cos\alpha+12\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha$.

18. For $(\theta,\alpha)\in[0,2\pi]\times[0,2\pi]$, and β a fixed real number, define $\mathbf{f}(\theta,\alpha)\equiv$

$$(\cos\theta(2+\cos\alpha),-\cos\beta\sin\theta(2+\cos\alpha)+\sin\beta\sin\alpha,\sin\beta\sin\theta(2+\cos\alpha)+\cos\beta\sin\alpha)^T.$$

Find a double integral which gives the area of $\mathbf{f}([0,2\pi]\times[0,2\pi])$.

Answer:

$D\mathbf{f}=$

$\begin{pmatrix}-\sin(2\theta+\theta\cos\alpha) & -\sin\alpha\cos\theta\\ -2\cos\beta\cos\theta-\cos\beta\cos\theta\cos\alpha & \cos\beta\sin\theta\sin\alpha+\sin\beta\cos\alpha\\ 2\sin\beta\cos\theta+\sin\beta\cos\theta\cos\alpha & -\sin(\beta\sin\theta\sin\alpha)+\cos\beta\cos\alpha\end{pmatrix}$ and so after many computations, the area element is $\left(4+4\cos\alpha+\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha$. Therefore, the area is $\int_0^{2\pi}\int_0^{2\pi}(2+\cos\alpha)\,d\theta\,d\alpha=8\pi^2$.

33.4 Exercises

1. Find a parameterization for the intersection of the planes $4x+2y+4z=3$ and $6x-2y=-1$.
2. Find a parameterization for the intersection of the plane $3x+y+z=1$ and the circular cylinder $x^2+y^2=1$.
3. Find a parameterization for the intersection of the plane $3x+2y+4z=4$ and the elliptic cylinder $x^2+4z^2=16$.
4. Find a parameterization for the straight line joining $(1,3,1)$ and $(-2,5,3)$.
5. Find a parameterization for the intersection of the surfaces $4y+3z=3x^2+2$ and $3y+2z=-x+3$.
6. Find the area of S if S is the part of the circular cylinder $x^2+y^2=4$ which lies between $z=0$ and $z=2+y$.
7. Find the area of S if S is the part of the cone $x^2+y^2=16z^2$ between $z=0$ and $z=h$.
8. Parametrizing the cylinder $x^2+y^2=a^2$ by $x=a\cos v, y=a\sin v, z=u$, show that the area element is $dA=a\,du\,dv$
9. Find the area enclosed by the limacon $r=2+\cos\theta$.

10. Find the surface area of the paraboloid $z = h\left(1 - x^2 - y^2\right)$ between $z = 0$ and $z = h$.

11. Evaluate $\int\int_S (1 + x)\, dA$ where S is the part of the plane $4x + y + 3z = 12$ which is in the first octant.

12. Evaluate $\int\int_S (1 + x)\, dA$ where S is the part of the cylinder $x^2 + y^2 = 9$ between $z = 0$ and $z = h$.

13. Evaluate $\int\int_S (1 + x)\, dA$ where S is the hemisphere $x^2 + y^2 + z^2 = 4$ between $x = 0$ and $x = 2$.

14. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let $\mathbf{f}(\theta, \alpha) \equiv (\cos\theta(4 + \cos\alpha), -\sin\theta(4 + \cos\alpha), \sin\alpha)^T$. Find the area of $\mathbf{f}([0, 2\pi] \times [0, 2\pi])$.

15. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let $\mathbf{f}(\theta, \alpha) \equiv$

$$(\cos\theta(3 + 2\cos\alpha), -\sin\theta(3 + 2\cos\alpha), 2\sin\alpha)^T.$$

Also let $h(\mathbf{x}) = \cos\alpha$ where α is such that

$$\mathbf{x} = (\cos\theta(3 + 2\cos\alpha), -\sin\theta(3 + 2\cos\alpha), 2\sin\alpha)^T.$$

Find $\int_{\mathbf{f}([0,2\pi]\times[0,2\pi])} h\, dV$.

16. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, let $\mathbf{f}(\theta, \alpha) \equiv$

$$(\cos\theta(4 + 3\cos\alpha), -\sin\theta(4 + 3\cos\alpha), 3\sin\alpha)^T.$$

Also let $h(\mathbf{x}) = \cos^2\theta$ where θ is such that

$$\mathbf{x} = (\cos\theta(4 + 3\cos\alpha), -\sin\theta(4 + 3\cos\alpha), 3\sin\alpha)^T.$$

Find $\int_{\mathbf{f}([0,2\pi]\times[0,2\pi])} h\, dV$.

17. For $(\theta, \alpha) \in [0, 28] \times [0, 2\pi]$, let $\mathbf{f}(\theta, \alpha) \equiv$

$$(\cos\theta(4 + 2\cos\alpha), -\sin\theta(4 + 2\cos\alpha), 2\sin\alpha + \theta)^T.$$

Find a double integral which gives the area of $\mathbf{f}([0, 28] \times [0, 2\pi])$.

18. For $(\theta, \alpha) \in [0, 2\pi] \times [0, 2\pi]$, and β a fixed real number, define $\mathbf{f}(\theta, \alpha) \equiv$

$$\begin{pmatrix} \cos\theta(3 + 2\cos\alpha), -\cos\beta\sin\theta(3 + 2\cos\alpha) + \\ 2\sin\beta\sin\alpha, \sin\beta\sin\theta(3 + 2\cos\alpha) + 2\cos\beta\sin\alpha \end{pmatrix}^T.$$

Find a double integral which gives the area of $\mathbf{f}([0, 2\pi] \times [0, 2\pi])$.

19. In the case where $\mathbf{f} : U \to \mathbb{R}^3$, show that

$$\frac{\partial(x, y, z)}{\partial(u_1, u_2)} = |\mathbf{f}_{u_1} \times \mathbf{f}_{u_2}|.$$

Thus the area element is $|\mathbf{f}_{u_1} \times \mathbf{f}_{u_2}|\, du_1\, du_2$.

20. In spherical coordinates, $\phi = c, \rho \in [0, R]$ determines a cone. Find the area of this cone without doing any work involving Jacobians and such.

Calculus Of Vector Fields

34.0.1 Outcomes

1. Define and evaluate the divergence of a vector field in terms of Cartesian coordinates.
2. Define and evaluate the Curl of a vector field in Cartesian coordinates.
3. Discover vector identities involving the gradient, divergence, and curl.
4. Recall and verify the divergence theorem.
5. Apply the divergence theorem.

34.1 Divergence And Curl Of A Vector Field

Here the important concepts of divergence and curl are defined.

Definition 34.1.1 *Let* $\mathbf{f} : U \to \mathbb{R}^p$ *for* $U \subseteq \mathbb{R}^p$ *denote a vector field. A scalar valued function is called a scalar field. The function,* $\mathbf{f}$ *is called a* C^k *vector field if the function,* $\mathbf{f}$ *is a* C^k *function. For a* C^1 *vector field, as just described* $\nabla \cdot \mathbf{f}(\mathbf{x}) \equiv \operatorname{div} \mathbf{f}(\mathbf{x})$ *known as the divergence, is defined as*

$$\nabla \cdot \mathbf{f}(\mathbf{x}) \equiv \operatorname{div} \mathbf{f}(\mathbf{x}) \equiv \sum_{i=1}^{p} \frac{\partial f_i}{\partial x_i}(\mathbf{x}).$$

Using the repeated summation convention, this is often written as

$$f_{i,i}(\mathbf{x}) \equiv \partial_i f_i(\mathbf{x})$$

where the comma indicates a partial derivative is being taken with respect to the i^{th} *variable and* ∂_i *denotes differentiation with respect to the* i^{th} *variable. In words, the divergence is the sum of the* i^{th} *derivative of the* i^{th} *component function of* $\mathbf{f}$ *for all values of* i*. If* $p = 3$*, the curl of the vector field yields another vector field and it is defined as follows.*

$$(\operatorname{curl}(\mathbf{f})(\mathbf{x}))_i \equiv (\nabla \times \mathbf{f}(\mathbf{x}))_i \equiv \varepsilon_{ijk} \partial_j f_k(\mathbf{x})$$

where here ∂_j *means the partial derivative with respect to* x_j *and the subscript of* i *in* $(\operatorname{curl}(\mathbf{f})(\mathbf{x}))_i$ *means the* i^{th} *Cartesian component of the vector,* $\operatorname{curl}(\mathbf{f})(\mathbf{x})$*. Thus the curl is evaluated by expanding the following determinant along the top row.*

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ f_1(x,y,z) & f_2(x,y,z) & f_3(x,y,z) \end{vmatrix}.$$

Note the similarity with the cross product. Sometimes the curl is called rot. (Short for rotation not decay.) Also

$$\nabla^2 f \equiv \nabla \cdot (\nabla f).$$

This last symbol is important enough that it is given a name, the Laplacian.It is also denoted by Δ. *Thus* $\nabla^2 f = \Delta f$. *In addition for* $\mathbf{f}$ *a vector field, the symbol* $\mathbf{f} \cdot \nabla$ *is defined as a "differential operator" in the following way.*

$$\mathbf{f} \cdot \nabla (\mathbf{g}) \equiv f_1 (\mathbf{x}) \frac{\partial \mathbf{g} (\mathbf{x})}{\partial x_1} + f_2 (\mathbf{x}) \frac{\partial \mathbf{g} (\mathbf{x})}{\partial x_2} + \cdots + f_p (\mathbf{x}) \frac{\partial \mathbf{g} (\mathbf{x})}{\partial x_p}.$$

Thus $\mathbf{f} \cdot \nabla$ *takes vector fields and makes them into new vector fields.*

This definition is in terms of a given coordinate system but later coordinate free definitions of the curl and div are presented. For now, everything is defined in terms of a given Cartesian coordinate system. The divergence and curl have profound physical significance and this will be discussed later. For now it is important to understand their definition in terms of coordinates. Be sure you understand that for $\mathbf{f}$ a vector field, $\operatorname{div} \mathbf{f}$ is a scalar field meaning it is a scalar valued function of three variables. For a scalar field, f, ∇f is a vector field described earlier on Page 678. For $\mathbf{f}$ a vector field having values in $\mathbb{R}^3$, $\operatorname{curl} \mathbf{f}$ is another vector field.

Example 34.1.2 *Let* $\mathbf{f} (\mathbf{x}) = xy\mathbf{i} + (z - y)\,\mathbf{j} + (\sin (x) + z)\,\mathbf{k}$. *Find* $\operatorname{div} \mathbf{f}$ *and* $\operatorname{curl} \mathbf{f}$.

First the divergence of $\mathbf{f}$ is

$$\frac{\partial (xy)}{\partial x} + \frac{\partial (z - y)}{\partial y} + \frac{\partial (\sin (x) + z)}{\partial z} = y + (-1) + 1 = y.$$

Now $\operatorname{curl} \mathbf{f}$ is obtained by evaluating

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ xy & z - y & \sin (x) + z \end{vmatrix} =$$

$$\mathbf{i} \left(\frac{\partial}{\partial y} (\sin (x) + z) - \frac{\partial}{\partial z} (z - y) \right) - \mathbf{j} \left(\frac{\partial}{\partial x} (\sin (x) + z) - \frac{\partial}{\partial z} (xy) \right) +$$

$$\mathbf{k} \left(\frac{\partial}{\partial x} (z - y) - \frac{\partial}{\partial y} (xy) \right) = -\mathbf{i} - \cos (x)\,\mathbf{j} - x\mathbf{k}.$$

34.1.1 Vector Identities

There are many interesting identities which relate the gradient, divergence and curl.

Theorem 34.1.3 *Assuming* $\mathbf{f}, \mathbf{g}$ *are a* C^2 *vector fields whenever necessary, the following identities are valid.*

1. $\nabla \cdot (\nabla \times \mathbf{f}) = 0$
2. $\nabla \times \nabla \phi = \mathbf{0}$
3. $\nabla \times (\nabla \times \mathbf{f}) = \nabla (\nabla \cdot \mathbf{f}) - \nabla^2 \mathbf{f}$ *where* $\nabla^2 \mathbf{f}$ *is a vector field whose* i^{th} *component is* $\nabla^2 f_i$.
4. $\nabla \cdot (\mathbf{f} \times \mathbf{g}) = \mathbf{g} \cdot (\nabla \times \mathbf{f}) - \mathbf{f} \cdot (\nabla \times \mathbf{g})$

5. $\nabla \times (\mathbf{f} \times \mathbf{g}) = (\nabla \cdot \mathbf{g})\,\mathbf{f} - (\nabla \cdot \mathbf{f})\,\mathbf{g} + (\mathbf{g} \cdot \nabla)\,\mathbf{f} - (\mathbf{f} \cdot \nabla)\,\mathbf{g}$

Proof: These are all easy to establish if you use the repeated index summation convention and the reduction identities discussed on Page 481.

$$\begin{aligned}
\nabla \cdot (\nabla \times \mathbf{f}) &= \partial_i (\nabla \times \mathbf{f})_i \\
&= \partial_i (\varepsilon_{ijk} \partial_j f_k) \\
&= \varepsilon_{ijk} \partial_i (\partial_j f_k) \\
&= \varepsilon_{jik} \partial_j (\partial_i f_k) \\
&= -\varepsilon_{ijk} \partial_j (\partial_i f_k) \\
&= -\varepsilon_{ijk} \partial_i (\partial_j f_k) \\
&= -\nabla \cdot (\nabla \times \mathbf{f}).
\end{aligned}$$

This establishes the first formula. The second formula is done similarly. Now consider the third.

$$\begin{aligned}
(\nabla \times (\nabla \times \mathbf{f}))_i &= \varepsilon_{ijk} \partial_j (\nabla \times \mathbf{f})_k \\
&= \varepsilon_{ijk} \partial_j (\varepsilon_{krs} \partial_r f_s) \\
&= \overbrace{\varepsilon_{kij}}^{=\varepsilon_{ijk}} \varepsilon_{krs} \partial_j (\partial_r f_s) \\
&= (\delta_{ir}\delta_{js} - \delta_{is}\delta_{jr}) \partial_j (\partial_r f_s) \\
&= \partial_j (\partial_i f_j) - \partial_j (\partial_j f_i) \\
&= \partial_i (\partial_j f_j) - \partial_j (\partial_j f_i) \\
&= \left(\nabla (\nabla \cdot \mathbf{f}) - \nabla^2 \mathbf{f}\right)_i
\end{aligned}$$

This establishes the third identity.

Consider the fourth identity.

$$\begin{aligned}
\nabla \cdot (\mathbf{f} \times \mathbf{g}) &= \partial_i (\mathbf{f} \times \mathbf{g})_i \\
&= \partial_i \varepsilon_{ijk} f_j g_k \\
&= \varepsilon_{ijk} (\partial_i f_j) g_k + \varepsilon_{ijk} f_j (\partial_i g_k) \\
&= (\varepsilon_{kij} \partial_i f_j) g_k - (\varepsilon_{jik} \partial_i g_k) f_k \\
&= \nabla \times \mathbf{f} \cdot \mathbf{g} - \nabla \times \mathbf{g} \cdot \mathbf{f}.
\end{aligned}$$

This proves the fourth identity.

Consider the fifth.

$$\begin{aligned}
(\nabla \times (\mathbf{f} \times \mathbf{g}))_i &= \varepsilon_{ijk} \partial_j (\mathbf{f} \times \mathbf{g})_k \\
&= \varepsilon_{ijk} \partial_j \varepsilon_{krs} f_r g_s \\
&= \varepsilon_{kij} \varepsilon_{krs} \partial_j (f_r g_s) \\
&= (\delta_{ir}\delta_{js} - \delta_{is}\delta_{jr}) \partial_j (f_r g_s) \\
&= \partial_j (f_i g_j) - \partial_j (f_j g_i) \\
&= (\partial_j g_j) f_i + g_j \partial_j f_i - (\partial_j f_j) g_i - f_j (\partial_j g_i) \\
&= ((\nabla \cdot \mathbf{g})\,\mathbf{f} + (\mathbf{g} \cdot \nabla)(\mathbf{f}) - (\nabla \cdot \mathbf{f})\,\mathbf{g} - (\mathbf{f} \cdot \nabla)(\mathbf{g}))_i
\end{aligned}$$

and this establishes the fifth identity.

I think the important thing about the above is not that these identities can be proved and are valid as much as the method by which they were proved. The reduction identities on Page

481 were used to discover the identities. There is a difference between proving something someone tells you about and both discovering what should be proved and proving it. This notation and the reduction identity make the discovery of vector identities fairly routine and this is why these things are of great significance.

34.1.2 Vector Potentials

One of the above identities says $\nabla \cdot (\nabla \times \mathbf{f}) = 0$. Suppose now $\nabla \cdot \mathbf{g} = 0$. Does it follow that there exists $\mathbf{f}$ such that $\mathbf{g} = \nabla \times \mathbf{f}$? It turns out that this is usually the case and when such an $\mathbf{f}$ exists, it is called a vector potential. Here is one way to do it, assuming everything is defined so the following formulas make sense.

$$\mathbf{f}(x,y,z) = \left(\int_0^z g_2(x,y,t)\, dt, -\int_0^z g_1(x,y,t)\, dt + \int_0^x g_3(t,y,0)\, dt, 0 \right)^T. \tag{34.1}$$

In verifying this you need to use the following manipulation which will generally hold under reasonable conditions but which has not been carefully shown yet.

$$\frac{\partial}{\partial x} \int_a^b h(x,t)\, dt = \int_a^b \frac{\partial h}{\partial x}(x,t)\, dt. \tag{34.2}$$

The above formula seems plausible because the integral is a sort of a sum and the derivative of a sum is the sum of the derivatives. However, this sort of sloppy reasoning will get you into all sorts of trouble. The formula involves the interchange of two limit operations, the integral and the limit of a difference quotient. Such an interchange can only be accomplished through a theorem. The following gives the necessary result. This lemma is stated without proof.

Lemma 34.1.4 *Suppose h and $\frac{\partial h}{\partial x}$ are continuous on the rectangle $R = [c,d] \times [a,b]$. Then (34.2) holds.*

The second formula of Theorem 34.1.3 states $\nabla \times \nabla \phi = \mathbf{0}$. This suggests the following question: Suppose $\nabla \times \mathbf{f} = \mathbf{0}$, does it follow there exists ϕ, a scalar field such that $\nabla \phi = \mathbf{f}$? The answer to this is often yes and a theorem will be given and proved after the presentation of Stoke's theorem. This scalar field, ϕ, is called a scalar potential for $\mathbf{f}$.

34.1.3 The Weak Maximum Principle

There is also a fundamental result having great significance which involves ∇^2 called the maximum principle. This principle says that if $\nabla^2 u \geq 0$ on a bounded open set, U, then u achieves its maximum value on the boundary of U.

Theorem 34.1.5 *Let U be a bounded open set in $\mathbb{R}^n$ and suppose $u \in C^2(U) \cap C\left(\overline{U}\right)$ such that $\nabla^2 u \geq 0$ in U. Then letting $\partial U = \overline{U} \setminus U$, it follows that* $\max\left\{u(\mathbf{x}) : \mathbf{x} \in \overline{U}\right\} = \max\left\{u(\mathbf{x}) : x \in \partial U\right\}$.

Proof: If this is not so, there exists $\mathbf{x}_0 \in U$ such that $u(\mathbf{x}_0) > \max\{u(\mathbf{x}) : x \in \partial U\} \equiv M$. Since U is bounded, there exists $\varepsilon > 0$ such that

$$u(\mathbf{x}_0) > \max\left\{u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2 : \mathbf{x} \in \partial U\right\}.$$

Therefore, $u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2$ also has its maximum in U because for ε small enough,

$$u(\mathbf{x}_0) + \varepsilon |\mathbf{x}_0|^2 > u(\mathbf{x}_0) > \max\left\{u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2 : \mathbf{x} \in \partial U\right\}$$

for all $\mathbf{x} \in \partial U$.

Now let $\mathbf{x}_1$ be the point in U at which $u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2$ achieves its maximum. As an exercise you should show that $\nabla^2 (f+g) = \nabla^2 f + \nabla^2 g$ and therefore, $\nabla^2 \left(u(\mathbf{x}) + \varepsilon |\mathbf{x}|^2 \right) = \nabla^2 u(\mathbf{x}) + 2n\varepsilon$. (Why?) Therefore,

$$0 \geq \nabla^2 u(\mathbf{x}_1) + 2n\varepsilon \geq 2n\varepsilon,$$

a contradiction. This proves the theorem.

34.2 Exercises

1. Find $\operatorname{div} \mathbf{f}$ and $\operatorname{curl} \mathbf{f}$ where $\mathbf{f}$ is
   (a) $\left(xyz, x^2 + \ln(xy), \sin x^2 + z\right)^T$
   (b) $(\sin x, \sin y, \sin z)^T$
   (c) $(f(x), g(y), h(z))^T$
   (d) $(x-2, y-3, z-6)^T$
   (e) $\left(y^2, 2xy, \cos z\right)^T$
   (f) $(f(y,z), g(x,z), h(y,z))^T$
2. Prove formula 2 of Theorem 34.1.3.
3. Show that if u and v are C^2 functions, then $\operatorname{curl}(u\nabla v) = \nabla u \times \nabla v$.
4. Simplify the expression $\mathbf{f} \times (\nabla \times \mathbf{g}) + \mathbf{g} \times (\nabla \times \mathbf{f}) + (\mathbf{f} \cdot \nabla)\mathbf{g} + (\mathbf{g} \cdot \nabla)\mathbf{f}$.
5. Simplify $\nabla \times (\mathbf{v} \times \mathbf{r})$ where $\mathbf{r} = (x,y,z)^T = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ and $\mathbf{v}$ is a constant vector.
6. Discover a formula which simplifies $\nabla \cdot (v \nabla u)$.
7. Verify that $\nabla \cdot (u\nabla v) - \nabla \cdot (v \nabla u) = u\nabla^2 v - v\nabla^2 u$.
8. Verify that $\nabla^2 (uv) = v\nabla^2 u + 2(\nabla u \cdot \nabla v) + u\nabla^2 v$.
9. Functions, u, which satisfy $\nabla^2 u = 0$ are called harmonic functions. Show the following functions are harmonic where ever they are defined.
   (a) $2xy$
   (b) $x^2 - y^2$
   (c) $\sin x \cosh y$
   (d) $\ln\left(x^2 + y^2\right)$
   (e) $1/\sqrt{x^2+y^2+z^2}$
10. Verify the formula given in (34.1) is a vector potential for $\mathbf{g}$ assuming that $\operatorname{div} \mathbf{g} = 0$.
11. Show that if $\nabla^2 u_k = 0$ for each $k = 1, 2, \cdots, m$, and c_k is a constant, then $\nabla^2 \left(\sum_{k=1}^m c_k u_k \right) = 0$ also.
12. In Theorem 34.1.5 why is $\nabla^2 \left(\varepsilon |\mathbf{x}|^2 \right) = 2n\varepsilon$?

13. Using Theorem 34.1.5 prove the following: Let $f \in C(\partial U)$ (f is continuous on ∂U.) where U is a bounded open set. Then there exists at most one solution, $u \in C^2(U) \cap C(\overline{U})$ and $\nabla^2 u = 0$ in U with $u = f$ on ∂U. **Hint:** Suppose there are two solutions, u_i, $i = 1, 2$ and let $w = u_1 - u_2$. Then use the maximum principle.

14. Suppose $\mathbf{B}$ is a vector field and $\nabla \times \mathbf{A} = \mathbf{B}$. Thus $\mathbf{A}$ is a vector potential for $\mathbf{B}$. Show that $\mathbf{A} + \nabla\phi$ is also a vector potential for $\mathbf{B}$. Here ϕ is just a C^2 scalar field. Thus the vector potential is not unique.

34.3 The Divergence Theorem

The divergence theorem relates an integral over a set to one on the boundary of the set. It is also called Gauss's theorem.

Definition 34.3.1 *A subset, V of $\mathbb{R}^3$ is called cylindrical in the x direction if it is of the form*

$$V = \{(x, y, z) : \phi(y, z) \leq x \leq \psi(y, z) \text{ for } (y, z) \in D\}$$

where D is a subset of the yz plane. V is cylindrical in the z direction if

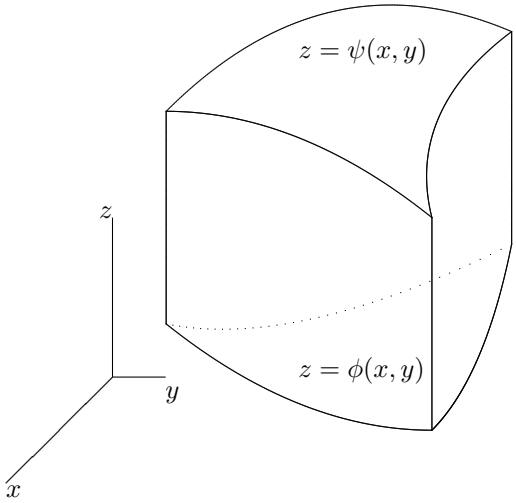
$$V = \{(x, y, z) : \phi(x, y) \leq z \leq \psi(x, y) \text{ for } (x, y) \in D\}$$

where D is a subset of the xy plane, and V is cylindrical in the y direction if

$$V = \{(x, y, z) : \phi(x, z) \leq y \leq \psi(x, z) \text{ for } (x, z) \in D\}$$

where D is a subset of the xz plane. If V is cylindrical in the z direction, denote by ∂V the boundary of V defined to be the points of the form $(x, y, \phi(x, y)), (x, y, \psi(x, y))$ for $(x, y) \in D$, along with points of the form (x, y, z) where $(x, y) \in \partial D$ and $\phi(x, y) \leq z \leq \psi(x, y)$. Points on ∂D are defined to be those for which every open ball contains points which are in D as well as points which are not in D. A similar definition holds for ∂V in the case that V is cylindrical in one of the other directions.

The following picture illustrates the above definition in the case of V cylindrical in the z direction.



Of course, many three dimensional sets are cylindrical in each of the coordinate directions. For example, a ball or a rectangle or a tetrahedron are all cylindrical in each direction. The following lemma allows the exchange of the volume integral of a partial derivative for an area integral in which the derivative is replaced with multiplication by an appropriate component of the unit exterior normal.

Lemma 34.3.2 *Suppose V is cylindrical in the z direction and that ϕ and ψ are the functions in the above definition. Assume ϕ and ψ are C^1 functions and suppose F is a C^1 function defined on V. Also, let $\mathbf{n} = (n_x, n_y, n_z)$ be the unit exterior normal to ∂V. Then*

$$\int\int\int_V \frac{\partial F}{\partial z}(x, y, z)\, dV = \int\int_{\partial V} F n_z\, dA.$$

Proof: From the fundamental theorem of calculus,

$$\begin{aligned} \int\int\int_V \frac{\partial F}{\partial z}(x, y, z)\, dV &= \int\int_D \int_{\phi(x,y)}^{\psi(x,y)} \frac{\partial F}{\partial z}(x, y, z)\, dz\, dx\, dy \\ &= \int\int_D \left[F(x, y, \psi(x, y)) - F(x, y, \phi(x, y))\right] dx\, dy \end{aligned} \tag{34.3}$$

Now the unit exterior normal on the top of V, the surface $(x, y, \psi(x, y))$ is

$$\frac{1}{\sqrt{\psi_x^2 + \psi_y^2 + 1}} \left(-\psi_x, -\psi_y, 1\right).$$

This follows from the observation that the top surface is the level surface, $z - \psi(x, y) = 0$ and so the gradient of this function of three variables is perpendicular to the level surface. It points in the correct direction because the z component is positive. Therefore, on the top surface,

$$n_z = \frac{1}{\sqrt{\psi_x^2 + \psi_y^2 + 1}}$$

Similarly, the unit normal to the surface on the bottom is

$$\frac{1}{\sqrt{\phi_x^2+\phi_y^2+1}}\left(\phi_x,\phi_y,-1\right)$$

and so on the bottom surface,

$$n_z=\frac{-1}{\sqrt{\phi_x^2+\phi_y^2+1}}$$

Note that here the z component is negative because since it is the outer normal it must point down. On the lateral surface, the one where $(x,y)\in\partial D$ and $z\in[\phi(x,y),\psi(x,y)]$, $n_z=0$.

The area element on the top surface is $dA=\sqrt{\psi_x^2+\psi_y^2+1}\,dx\,dy$ while the area element on the bottom surface is $\sqrt{\phi_x^2+\phi_y^2+1}\,dx\,dy$. Therefore, the last expression in (34.3) is of the form,

$$\int\int_D F(x,y,\psi(x,y))\overbrace{\frac{1}{\sqrt{\psi_x^2+\psi_y^2+1}}}^{n_z}\overbrace{\sqrt{\psi_x^2+\psi_y^2+1}\,dx\,dy}^{dA}+$$

$$\int\int_D F(x,y,\phi(x,y))\overbrace{\left(\frac{-1}{\sqrt{\phi_x^2+\phi_y^2+1}}\right)}^{n_z}\overbrace{\sqrt{\phi_x^2+\phi_y^2+1}\,dx\,dy}^{dA}$$

$$+\int\int_{\text{Lateral surface}} Fn_z\,dA,$$

the last term equaling zero because on the lateral surface, $n_z=0$. Therefore, this reduces to $\int\int_{\partial V}Fn_z\,dA$ as claimed.

The following corollary is entirely similar to the above.

Corollary 34.3.3 *If V is cylindrical in the y direction, then*

$$\int\int\int_V\frac{\partial F}{\partial y}\,dV=\int\int_{\partial V}Fn_y\,dA$$

and if V is cylindrical in the x direction, then

$$\int\int\int_V\frac{\partial F}{\partial x}\,dV=\int\int_{\partial V}Fn_x\,dA$$

With this corollary, here is a proof of the divergence theorem.

Theorem 34.3.4 *Let V be cylindrical in each of the coordinate directions and let $\mathbf{F}$ be a C^1 vector field defined on V. Then*

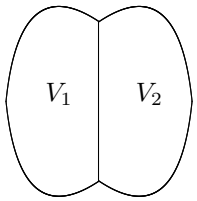
$$\int\int\int_V\nabla\cdot\mathbf{F}\,dV=\int\int_{\partial V}\mathbf{F}\cdot\mathbf{n}\,dA.$$

Proof: From the above lemma and corollary,

$$\begin{aligned}\int\int\int_V \nabla\cdot\mathbf{F}\,dV &= \int\int\int_V \frac{\partial F_1}{\partial x}+\frac{\partial F_2}{\partial y}+\frac{\partial F_3}{\partial y}\,dV\\ &= \int\int_{\partial V}\left(F_1n_x+F_2n_y+F_3n_z\right)\,dA\\ &= \int\int_{\partial V}\mathbf{F}\cdot\mathbf{n}\,dA.\end{aligned}$$

This proves the theorem.

The divergence theorem holds for much more general regions than this. Suppose for example you have a complicated region which is the union of finitely many disjoint regions of the sort just described which are cylindrical in each of the coordinate directions. Then the volume integral over the union of these would equal the sum of the integrals over the disjoint regions. If the boundaries of two of these regions intersect, then the area integrals will cancel out on the intersection because the unit exterior normals will point in opposite directions. Therefore, the sum of the integrals over the boundaries of these disjoint regions will reduce to an integral over the boundary of the union of these. Hence the divergence theorem will continue to hold. For example, consider the following picture. If the divergence theorem holds for each V_i in the following picture, then it holds for the union of these two.



General formulations of the divergence theorem involve Hausdorff measures and the Lebesgue integral, a better integral than the old fashioned Riemann integral which has been obsolete now for almost 100 years. When all is said and done, one finds that the conclusion of the divergence theorem is usually true and the theorem can be used with confidence.

34.3.1 Coordinate Free Concept Of Divergence

The divergence theorem also makes possible a coordinate free definition of the divergence.

Theorem 34.3.5 *Let $B(\mathbf{x},\delta)$ be the ball centered at $\mathbf{x}$ having radius δ and let $\mathbf{F}$ be a C^1 vector field. Then letting $v(B(\mathbf{x},\delta))$ denote the volume of $B(\mathbf{x},\delta)$ given by*

$$\int_{B(\mathbf{x},\delta)} dV,$$

it follows

$$\operatorname{div}\mathbf{F}(\mathbf{x}) = \lim_{\delta\to 0+}\frac{1}{v(B(\mathbf{x},\delta))}\int_{\partial B(\mathbf{x},\delta)}\mathbf{F}\cdot\mathbf{n}\,dA. \tag{34.4}$$

Proof: The divergence theorem holds for balls because they are cylindrical in every direction. Therefore,

$$\frac{1}{v(B(\mathbf{x},\delta))}\int_{\partial B(\mathbf{x},\delta)}\mathbf{F}\cdot\mathbf{n}\,dA = \frac{1}{v(B(\mathbf{x},\delta))}\int_{B(\mathbf{x},\delta)}\operatorname{div}\mathbf{F}(\mathbf{y})\,dV.$$

Therefore, since $\operatorname{div}\mathbf{F}(\mathbf{x})$ is a constant,

$$\begin{aligned}
&\left|\operatorname{div}\mathbf{F}(\mathbf{x}) - \frac{1}{v(B(\mathbf{x},\delta))}\int_{\partial B(\mathbf{x},\delta)} \mathbf{F}\cdot\mathbf{n}\,dA\right| \\
&= \left|\operatorname{div}\mathbf{F}(\mathbf{x}) - \frac{1}{v(B(\mathbf{x},\delta))}\int_{B(\mathbf{x},\delta)} \operatorname{div}\mathbf{F}(\mathbf{y})\,dV\right| \\
&= \left|\frac{1}{v(B(\mathbf{x},\delta))}\int_{B(\mathbf{x},\delta)} (\operatorname{div}\mathbf{F}(\mathbf{x}) - \operatorname{div}\mathbf{F}(\mathbf{y}))\,dV\right| \\
&\leq \frac{1}{v(B(\mathbf{x},\delta))}\int_{B(\mathbf{x},\delta)} |\operatorname{div}\mathbf{F}(\mathbf{x}) - \operatorname{div}\mathbf{F}(\mathbf{y})|\,dV \\
&\leq \frac{1}{v(B(\mathbf{x},\delta))}\int_{B(\mathbf{x},\delta)} \frac{\varepsilon}{2}\,dV < \varepsilon
\end{aligned}$$

whenever ε is small enough due to the continuity of $\operatorname{div}\mathbf{F}$. Since ε is arbitrary, this shows (34.4).

How is this definition independent of coordinates? It only involves geometrical notions of volume and dot product. This is why. Imagine rotating the coordinate axes, keeping all distances the same and expressing everything in terms of the new coordinates. The divergence would still have the same value because of this theorem.

34.4 Some Applications Of The Divergence Theorem

34.4.1 Hydrostatic Pressure

Imagine a fluid which does not move which is acted on by an acceleration, $\mathbf{g}$. Of course the acceleration is usually the acceleration of gravity. Also let the density of the fluid be ρ, a function of position. What can be said about the pressure, p, in the fluid? Let $B(\mathbf{x},\varepsilon)$ be a small ball centered at the point, $\mathbf{x}$. Then the force the fluid exerts on this ball would equal

$$-\int_{\partial B(\mathbf{x},\varepsilon)} p\mathbf{n}\,dA.$$

Here $\mathbf{n}$ is the unit exterior normal at a small piece of $\partial B(\mathbf{x},\varepsilon)$ having area dA. By the divergence theorem, (see Problem 1 on Page 784) this integral equals

$$-\int_{B(\mathbf{x},\varepsilon)} \nabla p\,dV.$$

Also the force acting on this small ball of fluid is

$$\int_{B(\mathbf{x},\varepsilon)} \rho\mathbf{g}\,dV.$$

Since it is given that the fluid does not move, the sum of these forces must equal zero. Thus

$$\int_{B(\mathbf{x},\varepsilon)} \rho\mathbf{g}\,dV = \int_{B(\mathbf{x},\varepsilon)} \nabla p\,dV.$$

Since this must hold for any ball in the fluid of any radius, it must be that

$$\nabla p = \rho \mathbf{g}. \tag{34.5}$$

It turns out that the pressure in a lake at depth z is equal to $62.5z$. This is easy to see from (34.5). In this case, $\mathbf{g} = g\mathbf{k}$ where $g = 32$ feet/sec^2. The weight of a cubic foot of water is 62.5 pounds. Therefore, the mass in slugs of this water is 62.5/32. Since it is a cubic foot, this is also the density of the water in slugs per cubic foot. Also, it is normally assumed that water is incompressible[1]. Therefore, this is the mass of water at any depth. Therefore,

$$\frac{\partial p}{\partial x}\mathbf{i}+\frac{\partial p}{\partial y}\mathbf{j}+\frac{\partial p}{\partial z}\mathbf{k} = \frac{62.5}{32} \times 32\mathbf{k}.$$

and so p does not depend on x and y and is only a function of z. It follows $p(0) = 0$, and $p'(z) = 62.5$. Therefore, $p(x, y, z) = 62.5z$. This establishes the claim. This is interesting but (34.5) is more interesting because it does not require ρ to be constant.

34.4.2 Archimedes Law Of Buoyancy

Archimedes principle states that when a solid body is immersed in a fluid the net force acting on the body by the fluid is directly up and equals the total weight of the fluid displaced.

Denote the set of points in three dimensions occupied by the body as V. Then for dA an increment of area on the surface of this body, the force acting on this increment of area would equal $-p\,dA\mathbf{n}$ where $\mathbf{n}$ is the exterior unit normal. Therefore, since the fluid does not move,

$$\int_{\partial V} -p\mathbf{n}\,dA = \int_V -\nabla p\,dV = \int_V \rho g\,dV\mathbf{k}$$

Which equals the total weight of the displaced fluid and you note the force is directed upward as claimed. Here ρ is the density and (34.5) is being used. There is an interesting point in the above explanation. Why does the second equation hold? Imagine that V were filled with fluid. Then the equation follows from (34.5) because in this equation $\mathbf{g} = -g\mathbf{k}$.

34.4.3 Equations Of Heat And Diffusion

Let $\mathbf{x}$ be a point in three dimensional space and let (x_1, x_2, x_3) be Cartesian coordinates of this point. Let there be a three dimensional body having density, $\rho = \rho(\mathbf{x}, t)$.

The heat flux, $\mathbf{J}$, in the body is defined as a vector which has the following property.

$$\text{Rate at which heat crosses } S = \int_S \mathbf{J}\cdot\mathbf{n}\,dA$$

where $\mathbf{n}$ is the unit normal in the desired direction. Thus if V is a three dimensional body,

$$\text{Rate at which heat leaves } V = \int_{\partial V} \mathbf{J}\cdot\mathbf{n}\,dA$$

where $\mathbf{n}$ is the unit exterior normal.

Fourier's law of heat conduction states that the heat flux, $\mathbf{J}$ satisfies $\mathbf{J} = -k\nabla(u)$ where u is the temperature and $k = k(u, \mathbf{x}, t)$ is called the coefficient of thermal conductivity. This changes depending on the material. It also can be shown by experiment to change

[1] There is no such thing as an incompressible fluid but this doesn't stop people from making this assumption.

with temperature. This equation for the heat flux states that the heat flows from hot places toward colder places in the direction of greatest rate of decrease in temperature. Let $c(\mathbf{x},t)$ denote the specific heat of the material in the body. This means the amount of heat within V is given by the formula $\int\int\int_V \rho(\mathbf{x},t)\,c(\mathbf{x},t)\,u(\mathbf{x},t)\;dV$. Suppose also there are sources for the heat within the material given by $f(\mathbf{x},u,t)$. If f is positive, the heat is increasing while if f is negative the heat is decreasing. For example such sources could result from a chemical reaction taking place. Then the divergence theorem can be used to verify the following equation for u. Such an equation is called a reaction diffusion equation.

$$\frac{\partial}{\partial t}\left(\rho(\mathbf{x},t)\,c(\mathbf{x},t)\,u(\mathbf{x},t)\right)=\nabla\cdot\left(k(u,\mathbf{x},t)\,\nabla u(\mathbf{x},t)\right)+f(\mathbf{x},u,t). \tag{34.6}$$

Take an arbitrary V for which the divergence theorem holds. Then the time rate of change of the heat in V is

$$\frac{d}{dt}\int_V \rho(\mathbf{x},t)\,c(\mathbf{x},t)\,u(\mathbf{x},t)\;dV=\int_V \frac{\partial\left(\rho(\mathbf{x},t)\,c(\mathbf{x},t)\,u(\mathbf{x},t)\right)}{\partial t}\,dV$$

where, as in the preceding example, this is a physical derivation so the consideration of hard mathematics is not necessary. Therefore, from the Fourier law of heat conduction, $\frac{d}{dt}\int_V \rho(\mathbf{x},t)\,c(\mathbf{x},t)\,u(\mathbf{x},t)\;dV=$

$$\int_V \frac{\partial\left(\rho(\mathbf{x},t)\,c(\mathbf{x},t)\,u(\mathbf{x},t)\right)}{\partial t}\,dV=\overbrace{\int_{\partial V}-\mathbf{J}\cdot\mathbf{n}\,dA}^{\text{rate at which heat enters}}+\int_V f(\mathbf{x},u,t)\;dV$$

$$\int_{\partial V} k\nabla(u)\cdot\mathbf{n}\,dA+\int_V f(\mathbf{x},u,t)\;dV=\int\int\int_V\left(\nabla\cdot\left(k\nabla(u)\right)+f\right)\,dV.$$

Since this holds for every sample volume, V it must be the case that the above reaction diffusion equation, (34.6) holds. Note that more interesting equations can be obtained by letting more of the quantities in the equation depend on temperature. However, the above is a fairly hard equation and people usually assume the coefficient of thermal conductivity depends only on $\mathbf{x}$ and that the reaction term, f depends only on $\mathbf{x}$ and t and that ρ and c are constant. Then it reduces to the much easier equation,

$$\frac{\partial}{\partial t}u(\mathbf{x},t)=\frac{1}{\rho c}\nabla\cdot\left(k(\mathbf{x})\,\nabla u(\mathbf{x},t)\right)+f(\mathbf{x},t). \tag{34.7}$$

This is often referred to as the heat equation. Sometimes there are modifications of this in which k is not just a scalar but a matrix to account for different heat flow properties in different directions. However, they are not much harder than the above. The major mathematical difficulties result from allowing k to depend on temperature.

It is known that the heat equation is not correct even if the thermal conductivity did not depend on u because it implies infinite speed of propagation of heat. However, this does not prevent people from using it.

34.4.4 Balance Of Mass

Let $\mathbf{y}$ be a point in three dimensional space and let (y_1,y_2,y_3) be Cartesian coordinates of this point. Let V be a region in three dimensional space and suppose a fluid having density, $\rho(\mathbf{y},t)$ and velocity, $\mathbf{v}(\mathbf{y},t)$ is flowing through this region. Then the mass of fluid leaving V per unit time is given by the area integral, $\int_{\partial V}\rho(\mathbf{y},t)\,\mathbf{v}(\mathbf{y},t)\cdot\mathbf{n}\,dA$ while the total mass of the fluid enclosed in V at a given time is $\int_V \rho(\mathbf{y},t)\;dV$. Also suppose mass originates at the

rate $f(\mathbf{y},t)$ per cubic unit per unit time within this fluid. Then the conclusion which can be drawn through the use of the divergence theorem is the following fundamental equation known as the mass balance equation.

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = f(\mathbf{y},t) \tag{34.8}$$

To see this is so, take an arbitrary V for which the divergence theorem holds. Then the time rate of change of the mass in V is

$$\frac{\partial}{\partial t}\int_V \rho(\mathbf{y},t)\, dV = \int_V \frac{\partial \rho(\mathbf{y},t)}{\partial t}\, dV$$

where the derivative was taken under the integral sign with respect to t. (This is a physical derivation and therefore, it is not necessary to fuss with the hard mathematics related to the change of limit operations. You should expect this to be true under fairly general conditions because the integral is a sort of sum and the derivative of a sum is the sum of the derivatives.) Therefore, the rate of change of mass, $\frac{\partial}{\partial t}\int_V \rho(\mathbf{y},t)\, dV$, equals

$$\begin{aligned}\int_V \frac{\partial \rho(\mathbf{y},t)}{\partial t}\, dV &= \overbrace{\int_{\partial V} \rho(\mathbf{y},t)\,\mathbf{v}(\mathbf{y},t)\cdot \mathbf{n}\, dA}^{\text{rate at which mass leaves}} + \int_V f(\mathbf{y},t)\, dV \\ &= \int_V \left(\nabla\cdot(\rho(\mathbf{y},t)\,\mathbf{v}(\mathbf{y},t)) + f(\mathbf{y},t)\right) dV.\end{aligned}$$

Since this holds for every sample volume, V it must be the case that the equation of continuity holds. Again, there are interesting mathematical questions here which can be explored but since it is a physical derivation, it is not necessary to dwell too much on them. If all the functions involved are continuous, it is certainly true but it is true under far more general conditions than that.

Also note this equation applies to many situations and f might depend on more than just $\mathbf{y}$ and t. In particular, f might depend also on temperature and the density, ρ. This would be the case for example if you were considering the mass of some chemical and f represented a chemical reaction. Mass balance is a general sort of equation valid in many contexts.

34.4.5 Balance Of Momentum

This example is a little more substantial than the above. It concerns the balance of momentum for a continuum. To see a full description of all the physics involved, you should consult a book on continuum mechanics. The situation is of a material in three dimensions and it deforms and moves about in three dimensions. This means this material is not a rigid body. Let B_0 denote an open set identifying a chunk of this material at time $t = 0$ and let B_t be an open set which identifies the same chunk of material at time $t > 0$.

Let $\mathbf{y}(t,\mathbf{x}) = (y_1(t,\mathbf{x}), y_2(t,\mathbf{x}), y_3(t,\mathbf{x}))$ denote the position with respect to Cartesian coordinates at time t of the point whose position at time $t = 0$ is $\mathbf{x} = (x_1,x_2,x_3)$. The coordinates, $\mathbf{x}$ are sometimes called the reference coordinates and sometimes the material coordinates and sometimes the Lagrangian coordinates. The coordinates, $\mathbf{y}$ are called the Eulerian coordinates or sometimes the spacial coordinates and the function, $(t,\mathbf{x}) \to \mathbf{y}(t,\mathbf{x})$ is called the motion. Thus

$$\mathbf{y}(0,\mathbf{x}) = \mathbf{x}. \tag{34.9}$$

The derivative,

$$D_2\mathbf{y}(t,\mathbf{x})$$

is called the deformation gradient. Recall the notation means you fix t and consider the function, $\mathbf{x} \to \mathbf{y}(t, \mathbf{x})$, taking its derivative. Since it is a linear transformation, it is represented by the usual matrix, whose ij^{th} entry is given by

$$F_{ij}(\mathbf{x}) = \frac{\partial y_i(t, \mathbf{x})}{\partial x_j}.$$

Let $\rho(t, \mathbf{y})$ denote the density of the material at time t at the point, $\mathbf{y}$ and let $\rho_0(\mathbf{x})$ denote the density of the material at the point, $\mathbf{x}$. Thus $\rho_0(\mathbf{x}) = \rho(0, \mathbf{x}) = \rho(0, y(0, \mathbf{x}))$. The first task is to consider the relationship between $\rho(t, \mathbf{y})$ and $\rho_0(\mathbf{x})$.

Lemma 34.4.1 *$\rho_0(\mathbf{x}) = \rho(t, \mathbf{y}(t, \mathbf{x})) \det(F)$ and in any reasonable physical motion, $\det(F) > 0$.*

Proof: Let V_0 represent a small chunk of material at $t = 0$ and let V_t represent the same chunk of material at time t. I will be a little sloppy and refer to V_0 as the small chunk of material at time $t = 0$ and V_t as the chunk of material at time t rather than an open set representing the chunk of material. Then by the change of variables formula for multiple integrals,

$$\int_{V_t} dV = \int_{V_0} |\det(F)| \, dV.$$

If $\det(F) = 0$ for some t the above formula shows that the chunk of material went from positive volume to zero volume and this is not physically possible. Therefore, it is impossible that $\det(F)$ can equal zero. However, at $t = 0$, $F = I$, the identity because of (34.9). Therefore, $\det(F) = 1$ at $t = 0$ and if it is assumed $t \to \det(F)$ is continuous it follows by the intermediate value theorem that $\det(F) > 0$ for all t. Of course it is not known for sure this function is continuous but the above shows why it is at least reasonable to expect $\det(F) > 0$.

Now using the change of variables formula,

$$\begin{aligned} \text{mass of } V_t &= \int_{V_t} \rho(t, \mathbf{y}) \, dV = \int_{V_0} \rho(t, \mathbf{y}(t, \mathbf{x})) \det(F) \, dV \\ &= \text{mass of } V_0 = \int_{V_0} \rho_0(\mathbf{x}) \, dV. \end{aligned}$$

Since V_0 is arbitrary, it follows $\rho_0(\mathbf{x}) = \rho(t, \mathbf{y}(t, \mathbf{x})) \det(F)$ as claimed. Note this shows that $\det(F)$ is a magnification factor for the density.

Now consider a small chunk of material, B_t at time t which corresponds to B_0 at time $t = 0$. The total linear momentum of this material at time t is

$$\int_{B_t} \rho(t, \mathbf{y}) \mathbf{v}(t, \mathbf{y}) \, dV$$

where $\mathbf{v}$ is the velocity. By Newton's second law, the time rate of change of this linear momentum should equal the total force acting on the chunk of material. In the following derivation, $dV(\mathbf{y})$ will indicate the integration is taking place with respect to the variable,

y. By Lemma 34.4.1 and the change of variables formula for multiple integrals

$$\begin{aligned}
\frac{d}{dt}\left(\int_{B_t}\rho(t,\mathbf{y})\,\mathbf{v}(t,\mathbf{y})\;dV(\mathbf{y})\right) &= \frac{d}{dt}\left(\int_{B_0}\rho(t,\mathbf{y}(t,\mathbf{x}))\,\mathbf{v}(t,\mathbf{y}(t,\mathbf{x}))\det(F)\;dV(\mathbf{x})\right)\\
&= \frac{d}{dt}\left(\int_{B_0}\rho_0(\mathbf{x})\,\mathbf{v}(t,\mathbf{y}(t,\mathbf{x}))\;dV(\mathbf{x})\right)\\
&= \int_{B_0}\rho_0(\mathbf{x})\left[\frac{\partial\mathbf{v}}{\partial t}+\frac{\partial\mathbf{v}}{\partial y_i}\frac{\partial y_i}{\partial t}\right]dV(\mathbf{x})\\
&= \int_{B_t}\rho(t,\mathbf{y})\det(F)\left[\frac{\partial\mathbf{v}}{\partial t}+\frac{\partial\mathbf{v}}{\partial y_i}\frac{\partial y_i}{\partial t}\right]\frac{1}{\det(F)}\,dV(\mathbf{y})\\
&= \int_{B_t}\rho(t,\mathbf{y})\left[\frac{\partial\mathbf{v}}{\partial t}+\frac{\partial\mathbf{v}}{\partial y_i}\frac{\partial y_i}{\partial t}\right]dV(\mathbf{y}).
\end{aligned}$$

Having taken the derivative of the total momentum, it is time to consider the total force acting on the chunk of material.

The force comes from two sources, a body force, **b** and a force which act on the boundary of the chunk of material called a traction force. Typically, the body force is something like gravity in which case, $\mathbf{b}=-g\rho\mathbf{k}$, assuming the Cartesian coordinate system has been chosen in the usual manner. The traction force is of the form

$$\int_{\partial B_t}\mathbf{s}(t,\mathbf{y},\mathbf{n})\;dA$$

where **n** is the unit exterior normal. Thus the traction force depends on position, time, and the orientation of the boundary of B_t. Cauchy showed the existence of a linear transformation, $T(t,\mathbf{y})$ such that $T(t,\mathbf{y})\,\mathbf{n}=\mathbf{s}(t,\mathbf{y},\mathbf{n})$. It follows there is a matrix, $T_{ij}(t,\mathbf{y})$ such that the i^{th} component of **s** is given by $\mathbf{s}_i(t,\mathbf{y},\mathbf{n})=T_{ij}(t,\mathbf{y})\,n_j$. Cauchy also showed this matrix is symmetric, $T_{ij}=T_{ji}$. It is called the Cauchy stress. Using Newton's second law to equate the time derivative of the total linear momentum with the applied forces and using the usual repeated index summation convention,

$$\int_{B_t}\rho(t,\mathbf{y})\left[\frac{\partial\mathbf{v}}{\partial t}+\frac{\partial\mathbf{v}}{\partial y_i}\frac{\partial y_i}{\partial t}\right]dV(\mathbf{y})=\int_{B_t}\mathbf{b}(t,\mathbf{y})\;dV(\mathbf{y})+\int_{\partial B_t}T_{ij}(t,\mathbf{y})\,n_j\,dA.$$

Here is where the divergence theorem is used. In the last integral, the multiplication by n_j is exchanged for the j^{th} partial derivative and an integral over B_t. Thus

$$\int_{B_t}\rho(t,\mathbf{y})\left[\frac{\partial\mathbf{v}}{\partial t}+\frac{\partial\mathbf{v}}{\partial y_i}\frac{\partial y_i}{\partial t}\right]dV(\mathbf{y})=\int_{B_t}\mathbf{b}(t,\mathbf{y})\;dV(\mathbf{y})+\int_{B_t}\frac{\partial\left(T_{ij}(t,\mathbf{y})\right)}{\partial y_j}\,dV(\mathbf{y}).$$

Since B_t was arbitrary, it follows

$$\begin{aligned}
\rho(t,\mathbf{y})\left[\frac{\partial\mathbf{v}}{\partial t}+\frac{\partial\mathbf{v}}{\partial y_i}\frac{\partial y_i}{\partial t}\right] &= \mathbf{b}(t,\mathbf{y})+\frac{\partial\left(T_{ij}(t,\mathbf{y})\right)}{\partial y_j}\\
&\equiv \mathbf{b}(t,\mathbf{y})+\operatorname{div}(T)
\end{aligned}$$

where here $\operatorname{div}T$ is a vector whose i^{th} component is given by

$$(\operatorname{div}T)_i=\frac{\partial T_{ij}}{\partial y_j}.$$

The term, $\frac{\partial\mathbf{v}}{\partial t}+\frac{\partial\mathbf{v}}{\partial y_i}\frac{\partial y_i}{\partial t}$, is the total derivative with respect to t of the velocity **v**. Thus you might see this written as

$$\rho\dot{\mathbf{v}}=\mathbf{b}+\operatorname{div}(T).$$

The above formulation of the balance of momentum involves the spatial coordinates, $\mathbf{y}$ but people also like to formulate momentum balance in terms of the material coordinates, $\mathbf{x}$. Of course this changes everything.

The momentum in terms of the material coordinates is

$$\int_{B_0} \rho_0(\mathbf{x})\,\mathbf{v}(t,\mathbf{x})\;dV$$

and so, since $\mathbf{x}$ does not depend on t,

$$\frac{d}{dt}\left(\int_{B_0} \rho_0(\mathbf{x})\,\mathbf{v}(t,\mathbf{x})\;dV\right) = \int_{B_0} \rho_0(\mathbf{x})\,\mathbf{v}_t(t,\mathbf{x})\;dV.$$

As indicated earlier, this is a physical derivation and so the mathematical questions related to interchange of limit operations are ignored. This must equal the total applied force. Thus

$$\int_{B_0} \rho_0(\mathbf{x})\,\mathbf{v}_t(t,\mathbf{x})\;dV = \int_{B_0} \mathbf{b}_0(t,\mathbf{x})\;dV + \int_{\partial B_t} T_{ij}n_j dA, \tag{34.10}$$

the first term on the right being the contribution of the body force given per unit volume in the material coordinates and the last term being the traction force discussed earlier. The task is to write this last integral as one over ∂B_0. For $\mathbf{y} \in \partial B_t$ there is a unit outer normal, $\mathbf{n}$. Here $\mathbf{y} = \mathbf{y}(t,\mathbf{x})$ for $\mathbf{x} \in \partial B_0$. Then define $\mathbf{N}$ to be the unit outer normal to B_0 at the point, $\mathbf{x}$. Near the point $\mathbf{y} \in \partial B_t$ the surface, ∂B_t is given parametrically in the form $\mathbf{y} = \mathbf{y}(s,t)$ for $(s,t) \in D \subseteq \mathbb{R}^2$ and it can be assumed the unit normal to ∂B_t near this point is

$$\mathbf{n} = \frac{\mathbf{y}_s(s,t) \times \mathbf{y}_t(s,t)}{|\mathbf{y}_s(s,t) \times \mathbf{y}_t(s,t)|}$$

with the area element given by $|\mathbf{y}_s(s,t) \times \mathbf{y}_t(s,t)|\;ds\,dt$. This is true for $\mathbf{y} \in P_t \subseteq \partial B_t$, a small piece of ∂B_t. Therefore, the last integral in (34.10) is the sum of integrals over small pieces of the form

$$\int_{P_t} T_{ij}n_j dA \tag{34.11}$$

where P_t is parametrized by $\mathbf{y}(s,t)$, $(s,t) \in D$. Thus the integral in (34.11) is of the form

$$\int_D T_{ij}(\mathbf{y}(s,t))\,(\mathbf{y}_s(s,t) \times \mathbf{y}_t(s,t))_j\;ds\,dt.$$

By the chain rule this equals

$$\int_D T_{ij}(\mathbf{y}(s,t))\left(\frac{\partial \mathbf{y}}{\partial x_\alpha}\frac{\partial x_\alpha}{\partial s} \times \frac{\partial \mathbf{y}}{\partial x_\beta}\frac{\partial x_\beta}{\partial t}\right)_j ds\,dt.$$

Remember $\mathbf{y} = \mathbf{y}(t,\mathbf{x})$ and it is always assumed the mapping $\mathbf{x} \to \mathbf{y}(t,\mathbf{x})$ is one to one and so, since on the surface ∂B_t near $\mathbf{y}$, the points are functions of (s,t), it follows $\mathbf{x}$ is also a function of (s,t). Now by the properties of the cross product, this last integral equals

$$\int_D T_{ij}(\mathbf{x}(s,t))\frac{\partial x_\alpha}{\partial s}\frac{\partial x_\beta}{\partial t}\left(\frac{\partial \mathbf{y}}{\partial x_\alpha} \times \frac{\partial \mathbf{y}}{\partial x_\beta}\right)_j ds\,dt \tag{34.12}$$

where here $\mathbf{x}(s,t)$ is the point of ∂B_0 which corresponds with $\mathbf{y}(s,t) \in \partial B_t$. Thus $T_{ij}(\mathbf{x}(s,t)) = T_{ij}(\mathbf{y}(s,t))$. (Perhaps this is a slight abuse of notation because T_{ij} is defined on ∂B_t, not on ∂B_0, but it avoids introducing extra symbols.) Next (34.12) equals

$$\int_D T_{ij}(\mathbf{x}(s,t))\frac{\partial x_\alpha}{\partial s}\frac{\partial x_\beta}{\partial t}\varepsilon_{jab}\frac{\partial y_a}{\partial x_\alpha}\frac{\partial y_b}{\partial x_\beta}\,ds\,dt$$

$$= \int_D T_{ij}\left(\mathbf{x}\left(s,t\right)\right) \frac{\partial x_\alpha}{\partial s}\frac{\partial x_\beta}{\partial t}\varepsilon_{cab}\delta_{jc}\frac{\partial y_a}{\partial x_\alpha}\frac{\partial y_b}{\partial x_\beta}\,ds\,dt$$

$$= \int_D T_{ij}\left(\mathbf{x}\left(s,t\right)\right) \frac{\partial x_\alpha}{\partial s}\frac{\partial x_\beta}{\partial t}\varepsilon_{cab}\overbrace{\frac{\partial y_c}{\partial x_p}\frac{\partial x_p}{\partial y_j}}^{=\delta_{jc}}\frac{\partial y_a}{\partial x_\alpha}\frac{\partial y_b}{\partial x_\beta}\,ds\,dt$$

$$= \int_D T_{ij}\left(\mathbf{x}\left(s,t\right)\right) \frac{\partial x_\alpha}{\partial s}\frac{\partial x_\beta}{\partial t}\frac{\partial x_p}{\partial y_j}\overbrace{\varepsilon_{cab}\frac{\partial y_c}{\partial x_p}\frac{\partial y_a}{\partial x_\alpha}\frac{\partial y_b}{\partial x_\beta}}^{=\varepsilon_{p\alpha\beta}\det(F)}\,ds\,dt$$

$$= \int_D \left(\det F\right) T_{ij}\left(\mathbf{x}\left(s,t\right)\right)\varepsilon_{p\alpha\beta}\frac{\partial x_\alpha}{\partial s}\frac{\partial x_\beta}{\partial t}\frac{\partial x_p}{\partial y_j}\,ds\,dt.$$

Now $\frac{\partial x_p}{\partial y_j} = F^{-1}_{pj}$ and also

$$\varepsilon_{p\alpha\beta}\frac{\partial x_\alpha}{\partial s}\frac{\partial x_\beta}{\partial t} = \left(\mathbf{x}_s \times \mathbf{x}_t\right)_p$$

so the result just obtained is of the form

$$\int_D \left(\det F\right) F^{-1}_{pj} T_{ij}\left(\mathbf{x}\left(s,t\right)\right)\left(\mathbf{x}_s \times \mathbf{x}_t\right)_p\,ds\,dt =$$

$$\int_D \left(\det F\right) T_{ij}\left(\mathbf{x}\left(s,t\right)\right)\left(F^{-T}\right)_{jp}\left(\mathbf{x}_s \times \mathbf{x}_t\right)_p\,ds\,dt.$$

This has transformed the integral over P_t to one over P_0, the part of ∂B_0 which corresponds with P_t. Thus the last integral is of the form

$$\int_{P_0} \det\left(F\right)\left(F^{-T}T\right)_{ip} N_p dA$$

Summing these up over the pieces of ∂B_t and ∂B_0 yields the last integral in (34.10) equals

$$\int_{\partial B_0} \det\left(F\right)\left(F^{-T}T\right)_{ip} N_p dA$$

and so the balance of momentum in terms of the material coordinates becomes

$$\int_{B_0} \rho_0\left(\mathbf{x}\right)\mathbf{v}_t\left(t,\mathbf{x}\right)\,dV = \int_{B_0} \mathbf{b}_0\left(t,\mathbf{x}\right)\,dV + \int_{\partial B_0} \det\left(F\right)\left(F^{-T}T\right)_{ip} N_p dA$$

The matrix, $\det\left(F\right)\left(F^{-T}T\right)_{ip}$ is called the Piola Kirchhoff stress, S. An application of the divergence theorem yields

$$\int_{B_0} \rho_0\left(\mathbf{x}\right)\mathbf{v}_t\left(t,\mathbf{x}\right)\,dV = \int_{B_0} \mathbf{b}_0\left(t,\mathbf{x}\right)\,dV + \int_{B_0} \frac{\partial\left(\det\left(F\right)\left(F^{-T}T\right)_{ip}\right)}{\partial x_p}\,dV.$$

Since B_0 is arbitrary, a balance law for momentum in terms of the material coordinates is obtained

$$\begin{aligned}\rho_0\left(\mathbf{x}\right)\mathbf{v}_t\left(t,\mathbf{x}\right) &= \mathbf{b}_0\left(t,\mathbf{x}\right) + \frac{\partial\left(\det\left(F\right)\left(F^{-T}T\right)_{ip}\right)}{\partial x_p} \\ &= \mathbf{b}_0\left(t,\mathbf{x}\right) + \operatorname{div}\left(\det\left(F\right)\left(F^{-T}T\right)\right) \\ &= \mathbf{b}_0\left(t,\mathbf{x}\right) + \operatorname{div} S. \end{aligned} \tag{34.13}$$

The main purpose of this presentation is to show how the divergence theorem is used in a significant way to obtain balance laws and to indicate a very interesting direction for further study. To continue, one needs to specify T or S as an appropriate function of things related to the motion, $\mathbf{y}$. Often the thing related to the motion is something called the strain and such relationships between the stress and the strain are known as constitutive laws. The proper formulation of constitutive laws involves more physical considerations such as frame indifference in which it is required the response of the system cannot depend on the manner in which the Cartesian coordinate system was chosen. There are also many other physical properties which can be included and which require a certain form for the constitutive equations. These considerations are outside the scope of this book and require a considerable amount of linear algebra.

There are also balance laws for energy which you may study later but these are more problematic than the balance laws for mass and momentum. However, the divergence theorem is used in these also.

34.4.6 The Wave Equation

As an example of how the balance law of momentum is used to obtain an important equation of mathematical physics, suppose $S = kF$ where k is a constant and F is the deformation gradient and let $\mathbf{u} \equiv \mathbf{y} - \mathbf{x}$. Thus $\mathbf{u}$ is the displacement. Then from (34.13) you can verify the following holds.

$$\rho_0(\mathbf{x})\,\mathbf{u}_{tt}(t,\mathbf{x}) = \mathbf{b}_0(t,\mathbf{x}) + k\Delta\mathbf{u}(t,\mathbf{x}) \tag{34.14}$$

In the case where ρ_0 is a constant and $\mathbf{b}_0 = 0$, this yields

$$\mathbf{u}_{tt} - c\Delta\mathbf{u} = \mathbf{0}.$$

The wave equation is $u_{tt} - c\Delta u = 0$ and so the above gives three wave equations, one for each component.

34.4.7 A Negative Observation

Many of the above applications of the divergence theorem are based on the assumption that matter is continuously distributed in a way that the above arguments are correct. In other words, a continuum. However, there is no such thing as a continuum. It has been known for some time now that matter is composed of atoms. It is not continuously distributed through some region of space as it is in the above. Apologists for this contradiction with reality sometimes say to consider enough of the material in question that it is reasonable to think of it as a continuum. This mystical reasoning is then violated as soon as they go from the integral form of the balance laws to the differential equations expressing the traditional formulation of these laws. See Problem 9 below, for example. However, these laws continue to be used and seem to lead to useful physical models which have value in predicting the behavior of physical systems. This is what justifies their use, not any fundamental truth.

34.4.8 Volumes Of Balls In $\mathbb{R}^n$ (For Those Who Know About The Gamma Function)

Recall, $B(\mathbf{x},r)$ denotes the set of all $\mathbf{y} \in \mathbb{R}^n$ such that $|\mathbf{y}-\mathbf{x}| < r$. By the change of variables formula for multiple integrals or simple geometric reasoning, all balls of radius r have the same volume. Furthermore, simple reasoning or change of variables formula will show that the volume of the ball of radius r equals $\alpha_n r^n$ where α_n will denote the volume of

the unit ball in $\mathbb{R}^n$. With the divergence theorem, it is now easy to give a simple relationship between the surface area of the ball of radius r and the volume. By the divergence theorem,

$$\int_{B(\mathbf{0},r)} \operatorname{div} \mathbf{x}\, dx = \int_{\partial B(\mathbf{0},r)} \mathbf{x}\cdot\frac{\mathbf{x}}{|\mathbf{x}|}dA$$

because the unit outward normal on $\partial B(\mathbf{0},r)$ is $\frac{\mathbf{x}}{|\mathbf{x}|}$. Therefore,

$$n\alpha_n r^n = rA(\partial B(\mathbf{0},r))$$

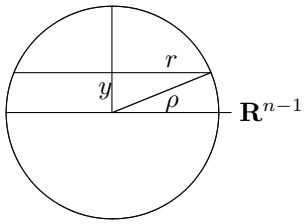
and so

$$A(\partial B(\mathbf{0},r)) = n\alpha_n r^{n-1}.$$

You recall the surface area of $S^2 \equiv \left\{\mathbf{x}\in\mathbb{R}^3 : |\mathbf{x}| = r\right\}$ is given by $4\pi r^2$ while the volume of the ball, $B(\mathbf{0},r)$ is $\frac{4}{3}\pi r^3$. This follows the above pattern. You just take the derivative with respect to the radius of the volume of the ball of radius r to get the area of the surface of this ball. Let ω_n denote the area of the sphere $S^{n-1} = \{\mathbf{x}\in\mathbb{R}^n : |\mathbf{x}| = 1\}$. I just showed that

$$\omega_n = n\alpha_n.$$

I want to find α_n now and also to get a relationship between ω_n and ω_{n-1}. Consider the following picture of the ball of radius ρ seen on the side.



Taking slices at height y as shown and using that these slices have $n-1$ dimensional area equal to $\alpha_{n-1}r^{n-1}$, it follows

$$\alpha_n\rho^n = 2\int_0^\rho \alpha_{n-1}\left(\rho^2 - y^2\right)^{(n-1)/2} dy$$

In the integral, change variables, letting $y = \rho\cos\theta$. Then

$$\alpha_n\rho^n = 2\rho^n\alpha_{n-1}\int_0^{\pi/2} \sin^n(\theta)\, d\theta.$$

It follows that

$$\alpha_n = 2\alpha_{n-1}\int_0^{\pi/2} \sin^n(\theta)\, d\theta. \tag{34.15}$$

Consequently,

$$\omega_n = \frac{2n\omega_{n-1}}{n-1}\int_0^{\pi/2} \sin^n(\theta)\, d\theta. \tag{34.16}$$

This is a little messier than I would like.

$$\begin{aligned}
\int_0^{\pi/2} \sin^n(\theta)\, d\theta &= -\cos\theta\sin^{n-1}\theta\big|_0^{\pi/2} + (n-1)\int_0^{\pi/2} \cos^2\theta\sin^{n-2}\theta \\
&= (n-1)\int_0^{\pi/2} \left(1-\sin^2\theta\right)\sin^{n-2}(\theta)\, d\theta \\
&= (n-1)\int_0^{\pi/2} \sin^{n-2}(\theta)\, d\theta - (n-1)\int_0^{\pi/2} \sin^n(\theta)\, d\theta
\end{aligned}$$

Hence

$$n\int_0^{\pi/2}\sin^n(\theta)\,d\theta=(n-1)\int_0^{\pi/2}\sin^{n-2}(\theta)\,d\theta \tag{34.17}$$

and so (34.16) is of the form

$$\omega_n=2\omega_{n-1}\int_0^{\pi/2}\sin^{n-2}(\theta)\,d\theta. \tag{34.18}$$

So what is α_n explicitly? Clearly $\alpha_1=2$ and $\alpha_2=\pi$.

Theorem 34.4.2 *$\alpha_n=\frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}$ where Γ denotes the gamma function, defined for $\alpha>0$ by*

$$\Gamma(\alpha)\equiv\int_0^\infty e^{-t}t^{\alpha-1}dt.$$

Proof: Recall that $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$. Now note the given formula holds if $n=1$ because

$$\Gamma\left(\frac{1}{2}+1\right)=\frac{1}{2}\Gamma\left(\frac{1}{2}\right)=\frac{\sqrt{\pi}}{2}.$$

(I leave it as an exercise for you to verify that $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$.) Thus

$$\alpha_1=2=\frac{\sqrt{\pi}}{\sqrt{\pi}/2}$$

satisfying the formula. Now suppose this formula holds for $k\leq n$. Then from the induction hypothesis, (34.18), (34.17), (34.15) and (34.16),

$$\begin{aligned}
\alpha_{n+1} &= 2\alpha_n\int_0^{\pi/2}\sin^{n+1}(\theta)\,d\theta\\
&= 2\alpha_n\frac{n}{n+1}\int_0^{\pi/2}\sin^{n-1}(\theta)\,d\theta\\
&= 2\alpha_n\frac{n}{n+1}\frac{\alpha_{n-1}}{2\alpha_{n-2}}\\
&= \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}\frac{n}{n+1}\pi^{1/2}\frac{\Gamma\left(\frac{n-2}{2}+1\right)}{\Gamma\left(\frac{n-1}{2}+1\right)}\\
&= \frac{\pi^{n/2}}{\Gamma\left(\frac{n-2}{2}+1\right)\left(\frac{n}{2}\right)}\frac{n}{n+1}\pi^{1/2}\frac{\Gamma\left(\frac{n-2}{2}+1\right)}{\Gamma\left(\frac{n-1}{2}+1\right)}\\
&= 2\pi^{(n+1)/2}\frac{1}{n+1}\frac{1}{\Gamma\left(\frac{n-1}{2}+1\right)}\\
&= \pi^{(n+1)/2}\frac{1}{\left(\frac{n+1}{2}\right)}\frac{1}{\Gamma\left(\frac{n-1}{2}+1\right)}\\
&= \pi^{(n+1)/2}\frac{1}{\left(\frac{n+1}{2}\right)\Gamma\left(\frac{n+1}{2}\right)}=\frac{\pi^{(n+1)/2}}{\Gamma\left(\frac{n+1}{2}+1\right)}.
\end{aligned}$$

This proves the theorem.

34.4.9 Electrostatics

Coloumb's law says that the electric field intensity at $\mathbf{x}$ of a charge q located at point, $\mathbf{x}_0$ is given by

$$\mathbf{E} = k\frac{q\left(\mathbf{x}-\mathbf{x}_0\right)}{\left|\mathbf{x}-\mathbf{x}_0\right|^3}$$

where the electric field intensity is defined to be the force experienced by a unit positive charge placed at the point, $\mathbf{x}$. Note that this is a vector and that its direction depends on the sign of q. It points away from $\mathbf{x}_0$ if q is positive and points toward $\mathbf{x}_0$ if q is negative. The constant, k is a physical constant like the gravitation constant. It has been computed through careful experiments similar to those used with the calculation of the gravitation constant.

The interesting thing about Coloumb's law is that $\mathbf{E}$ is the gradient of a function. In fact,

$$\mathbf{E} = \nabla\left(qk\frac{1}{\left|\mathbf{x}-\mathbf{x}_0\right|}\right).$$

The other thing which is significant about this is that in three dimensions and for $\mathbf{x} \neq \mathbf{x}_0$,

$$\nabla\cdot\nabla\left(qk\frac{1}{\left|\mathbf{x}-\mathbf{x}_0\right|}\right) = \nabla\cdot\mathbf{E} = 0. \tag{34.19}$$

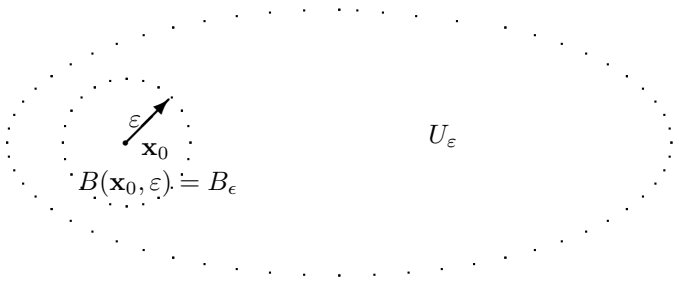
This is left as an exercise for you to verify.

These observations will be used to derive a very important formula for the integral,

$$\int_{\partial U}\mathbf{E}\cdot\mathbf{n}dS$$

where $\mathbf{E}$ is the electric field intensity due to a charge, q located at the point, $\mathbf{x}_0 \in U$, a bounded open set for which the divergence theorem holds.

Let U_ε denote the open set obtained by removing the open ball centered at $\mathbf{x}_0$ which has radius ε where ε is small enough that the following picture is a correct representation of the situation.



Then on the boundary of B_ε the unit outer normal to U_ε is $-\frac{\mathbf{x}-\mathbf{x}_0}{|\mathbf{x}-\mathbf{x}_0|}$. Therefore,

$$\begin{aligned}\int_{\partial B_\varepsilon}\mathbf{E}\cdot\mathbf{n}dS &= -\int_{\partial B_\varepsilon}k\frac{q\left(\mathbf{x}-\mathbf{x}_0\right)}{\left|\mathbf{x}-\mathbf{x}_0\right|^3}\cdot\frac{\mathbf{x}-\mathbf{x}_0}{\left|\mathbf{x}-\mathbf{x}_0\right|}dS\\ &= -kq\int_{\partial B_\varepsilon}\frac{1}{\left|\mathbf{x}-\mathbf{x}_0\right|^2}dS = \frac{-kq}{\varepsilon^2}\int_{\partial B_\varepsilon}dS\\ &= \frac{-kq}{\varepsilon^2}4\pi\varepsilon^2 = -4\pi kq.\end{aligned}$$

Therefore, from the divergence theorem and observation (34.19),

$$-4\pi kq + \int_{\partial U} \mathbf{E}\cdot\mathbf{n}dS = \int_{\partial U_\varepsilon} \mathbf{E}\cdot\mathbf{n}dS = \int_{U_\varepsilon} \nabla\cdot\mathbf{E}dV = 0.$$

It follows that

$$4\pi kq = \int_{\partial U} \mathbf{E}\cdot\mathbf{n}dS.$$

If there are several charges located inside U, say $q_1, q_2, \cdots, q_n$, then letting $\mathbf{E}_i$ denote the electric field intensity of the i^{th} charge and $\mathbf{E}$ denoting the total resulting electric field intensity due to all these charges,

$$\begin{aligned}\int_{\partial U} \mathbf{E}\cdot\mathbf{n}dS &= \sum_{i=1}^{n}\int_{\partial U} \mathbf{E}_i\cdot\mathbf{n}dS \\ &= \sum_{i=1}^{n} 4\pi kq_i = 4\pi k\sum_{i=1}^{n} q_i.\end{aligned}$$

This is known as Gauss's law and it is the fundamental result in electrostatics.

34.5 Exercises

1. To prove the divergence theorem, it was shown first that the spacial partial derivative in the volume integral could be exchanged for multiplication by an appropriate component of the exterior normal. This problem starts with the divergence theorem and goes the other direction. Assuming the divergence theorem, holds for a region, V, show that $\int_{\partial V}\mathbf{n}u\,dA = \int_V \nabla u\,dV$. Note this implies $\int_V \frac{\partial u}{\partial x}\,dV = \int_{\partial V} n_1 u\,dA$.

2. Let V be such that the divergence theorem holds. Show that $\int_V \nabla\cdot(u\nabla v)\,dV = \int_{\partial V} u\frac{\partial v}{\partial n}\,dA$ where $\mathbf{n}$ is the exterior normal and $\frac{\partial v}{\partial n}$ denotes the directional derivative of v in the direction $\mathbf{n}$.

3. Let V be such that the divergence theorem holds. Show that $\int\int\int_V \left(v\nabla^2 u - u\nabla^2 v\right)\,dV = \int\int_{\partial V}\left(v\frac{\partial u}{\partial n} - u\frac{\partial v}{\partial n}\right)\,dA$ where $\mathbf{n}$ is the exterior normal and $\frac{\partial u}{\partial n}$ is defined in Problem 2.

4. Let V be a ball and suppose $\nabla^2 u = f$ in V while $u = g$ on ∂V. Show there is at most one solution to this boundary value problem which is C^2 in V and continuous on V with its boundary. **Hint:** You might consider $w = u - v$ where u and v are solutions to the problem. Then use the result of Problem 2 and the identity

$$w\nabla^2 w = \nabla\cdot(w\nabla w) - \nabla w\cdot\nabla w$$

   to conclude $\nabla w = 0$. Then show this implies w must be a constant by considering $h(t) = w(t\mathbf{x} + (1-t)\mathbf{y})$ and showing h is a constant. Alternatively, you might consider the maximum principle.

5. Show that $\int_{\partial V}\nabla\times\mathbf{v}\cdot\mathbf{n}\,dA = 0$ where V is a region for which the divergence theorem holds and $\mathbf{v}$ is a C^2 vector field.

6. Let $\mathbf{F}(x,y,z) = (x,y,z)$ be a vector field in $\mathbb{R}^3$ and let V be a three dimensional shape and let $\mathbf{n} = (n_1, n_2, n_3)$. Show $\int_{\partial V}(xn_1 + yn_2 + zn_3)\,dA = 3\times$ volume of V.

7. Does the divergence theorem hold for higher dimensions? If so, explain why it does. How about two dimensions?

8. Let $\mathbf{F} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ and let V denote the tetrahedron formed by the planes, $x = 0, y = 0, z = 0$, and $\frac{1}{3}x + \frac{1}{3}y + \frac{1}{5}z = 1$. Verify the divergence theorem for this example.

9. Suppose $f : U \to \mathbb{R}$ is continuous where U is some open set and for all $B \subseteq U$ where B is a ball, $\int_B f(\mathbf{x})\, dV = 0$. Show this implies $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in U$.

10. Let U denote the box centered at $(0, 0, 0)$ with sides parallel to the coordinate planes which has width 4, length 2 and height 3. Find the flux integral $\int\int_{\partial U} \mathbf{F} \cdot \mathbf{n}\, dS$ where $\mathbf{F} = \langle x + 3, 2y, 3z \rangle$. **Hint:** If you like, you might want to use the divergence theorem.

11. Verify (34.14) from (34.13) and the assumption that $S = kF$.

12. Fick's law for diffusion states the flux of a diffusing species, $\mathbf{J}$ is proportional to the gradient of the concentration, c. Write this law getting the sign right for the constant of proportionality and derive an equation similar to the heat equation for the concentration, c. Typically, c is the concentration of some sort of pollutant or a chemical.

13. Show that if $u_k, k = 1, 2, \cdots, n$ each satisfies (34.7) then for any choice of constants, $c_1, \cdots, c_n$, so does
$$\sum_{k=1}^{n} c_k u_k.$$

14. Suppose $k(\mathbf{x}) = k$, a constant and $f = 0$. Then in one dimension, the heat equation is of the form $u_t = \alpha u_{xx}$. Show $u(x, t) = e^{-\alpha n^2 t} \sin(nx)$ satisfies the heat equation[2].

15. In a linear, viscous, incompressible fluid, the Cauchy stress is of the form
$$T_{ij}(t, \mathbf{y}) = \lambda \left(\frac{v_{i,j}(t, \mathbf{y}) + v_{j,i}(t, \mathbf{y})}{2} \right)$$
where the comma followed by an index indicates the partial derivative with respect to that variable and $\mathbf{v}$ is the velocity. Thus
$$v_{i,j} = \frac{\partial v_i}{\partial y_j}$$
Show, using the balance of mass equation that incompressible implies $\operatorname{div} \mathbf{v} = 0$. Next show the balance of momentum equation requires
$$\rho \dot{\mathbf{v}} - \frac{\lambda}{2} \Delta \mathbf{v} = \rho \left[\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial y_i} v_i \right] - \frac{\lambda}{2} \Delta \mathbf{v} = \mathbf{b}.$$
This is the famous Navier Stokes equation for incompressible viscous linear fluids. There are still open questions related to this equation, one of which is worth \$1,000,000 at this time.

[2] Fourier, an officer in Napoleon's army studied solutions to the heat equation back in 1813. He was interested in heat flow in cannons. He sought to find solutions by adding up infinitely many solutions of this form. Actually, it was a little more complicated because cannons are not one dimensional but it was the beginning of the study of Fourier series, a topic which fascinated mathematicians for the next 150 years and motivated the development of analysis.

Stokes And Green's Theorems

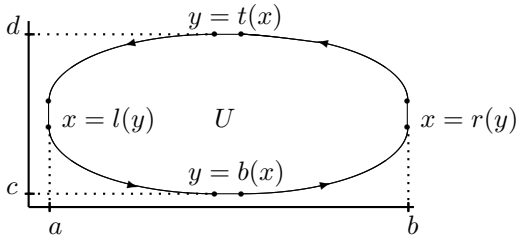
35.0.1 Outcomes

1. Recall and verify Green's theorem.
2. Apply Green's theorem to evaluate line integrals.
3. Apply Green's theorem to find the area of a region.
4. Explain what is meant by the curl of a vector field.
5. Evaluate the curl of a vector field.
6. Derive and apply formulas involving divergence, gradient and curl.
7. Recall and use Stoke's theorem.
8. Apply Stoke's theorem to calculate the circulation or work of a vector field around a simple closed curve.
9. Recall and apply the fundamental theorem for line integrals.
10. Determine whether a vector field is a gradient using the curl test.
11. Recover a function from its gradient when possible.

35.1 Green's Theorem

Green's theorem is an important theorem which relates line integrals to integrals over a surface in the plane. It can be used to establish the much more significant Stoke's theorem but is interesting for it's own sake. Historically, it was important in the development of complex analysis. I will first establish Green's theorem for regions of a particular sort and then show that the theorem holds for many other regions also. Suppose a region is of the form indicated in the following picture in which

$$\begin{aligned} U &= \{(x,y) : x \in (a,b) \text{ and } y \in (b(x), t(x))\} \\ &= \{(x,y) : y \in (c,d) \text{ and } x \in (l(y), r(y))\}. \end{aligned}$$



I will refer to such a region as being convex in both the x and y directions.

Lemma 35.1.1 *Let $\mathbf{F}(x,y) \equiv (P(x,y), Q(x,y))$ be a C^1 vector field defined near U where U is a region of the sort indicated in the above picture which is convex in both the x and y directions. Suppose also that the functions, r, l, t, and b in the above picture are all C^1 functions and denote by ∂U the boundary of U oriented such that the direction of motion is counter clockwise. (As you walk around U on ∂U, the points of U are on your left.) Then*

$$\int_{\partial U} P dx + Q dy \equiv$$

$$\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int\int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA. \tag{35.1}$$

Proof: First consider the right side of (35.1).

$$\int\int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA$$

$$\begin{aligned} &= \int_c^d \int_{l(y)}^{r(y)} \frac{\partial Q}{\partial x} dx dy - \int_a^b \int_{b(x)}^{t(x)} \frac{\partial P}{\partial y} dy dx \\ &= \int_c^d \left(Q(r(y), y) - Q(l(y), y) \right) dy + \int_a^b \left(P(x, b(x))) - P(x, t(x)) \right) dx. \end{aligned} \tag{35.2}$$

Now consider the left side of (35.1). Denote by V the vertical parts of ∂U and by H the horizontal parts.

$$\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} =$$

$$\begin{aligned} &= \int_{\partial U} (\langle 0, Q \rangle + \langle P, 0 \rangle) \cdot d\mathbf{R} \\ &= \int_c^d \langle 0, Q(r(s), s) \rangle \cdot \langle r'(s), 1 \rangle ds + \int_H \langle 0, Q(r(s), s) \rangle \cdot \langle \pm 1, 0 \rangle ds \\ &\quad - \int_c^d \langle 0, Q(l(s), s) \rangle \cdot \langle l'(s), 1 \rangle ds + \int_a^b \langle P(s, b(s)), 0 \rangle \cdot \langle 1, b'(s) \rangle ds \\ &\quad + \int_V \langle P(s, b(s)), 0 \rangle \cdot \langle 0, \pm 1 \rangle ds - \int_a^b \langle P(s, t(s)), 0 \rangle \cdot \langle 1, t'(s) \rangle ds \\ &= \int_c^d Q(r(s), s) ds - \int_c^d Q(l(s), s) ds + \int_a^b P(s, b(s)) ds - \int_a^b P(s, t(s)) ds \end{aligned}$$

which coincides with (35.2). This proves the lemma.

Corollary 35.1.2 *Let everything be the same as in Lemma 35.1.1 but only assume the functions r, l, t, and b are continuous and piecewise C^1 functions. Then the conclusion this lemma is still valid.*

Proof: The details are left for you. All you have to do is to break up the various line integrals into the sum of integrals over sub intervals on which the function of interest is C^1.

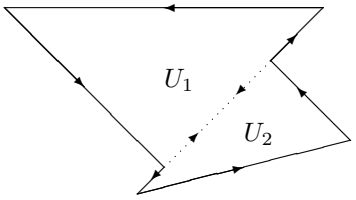
From this corollary, it follows (35.1) is valid for any triangle for example.

Now suppose (35.1) holds for $U_1, U_2, \cdots, U_m$ and the open sets, U_k have the property that no two have nonempty intersection and their boundaries intersect only in a finite number of piecewise smooth curves. Then (35.1) must hold for $U \equiv \cup_{i=1}^m U_i$, the union of these sets. This is because

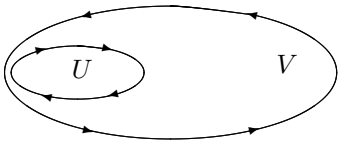
$$\int\int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA =$$

$$\begin{aligned} &= \sum_{k=1}^m \int\int_{U_k} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \\ &= \sum_{k=1}^m \int_{\partial U_k} \mathbf{F} \cdot d\mathbf{R} = \int_{\partial U} \mathbf{F} \cdot d\mathbf{R} \end{aligned}$$

because if $\Gamma = \partial U_k \cap \partial U_j$, then its orientation as a part of ∂U_k is opposite to its orientation as a part of ∂U_j and consequently the line integrals over Γ will cancel, points of Γ also not being in ∂U. As an illustration, consider the following picture for two such U_k.



Similarly, if $U \subseteq V$ and if also $\partial U \subseteq V$ and both U and V are open sets for which (35.1) holds, then the open set, $V \setminus (U \cup \partial U)$ consisting of what is left in V after deleting U along with its boundary also satisfies (35.1). Roughly speaking, you can drill holes in a region for which (35.1) holds and get another region for which this continues to hold provided (35.1) holds for the holes. To see why this is so, consider the following picture which typifies the situation just described.



Then

$$\int_{\partial V} \mathbf{F} \cdot d\mathbf{R} = \int\int_V \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA$$

$$\begin{aligned} &= \int\int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA + \int\int_{V\setminus U} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \\ &= \int_{\partial U} \mathbf{F}\cdot d\mathbf{R} + \int\int_{V\setminus U} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \end{aligned}$$

and so

$$\int\int_{V\setminus U} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = \int_{\partial V} \mathbf{F}\cdot d\mathbf{R} - \int_{\partial U} \mathbf{F}\cdot d\mathbf{R}$$

which equals

$$\int_{\partial(V\setminus U)} \mathbf{F}\cdot d\mathbf{R}$$

where ∂V is oriented as shown in the picture. (If you walk around the region, $V \setminus U$ with the area on the left, you get the indicated orientation for this curve.)

You can see that (35.1) is valid quite generally. This verifies the following theorem.

Theorem 35.1.3 *(Green's Theorem) Let U be an open set in the plane and let ∂U be piecewise smooth and let $\mathbf{F}(x,y) = \langle P(x,y), Q(x,y)\rangle$ be a C^1 vector field defined near U. Then it is often[1] the case that*

$$\int_{\partial U} \mathbf{F}\cdot d\mathbf{R} = \int\int_U \left(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y)\right) dA.$$

Proposition 35.1.4 *Let U be an open set in $\mathbb{R}^2$ for which Green's theorem holds. Then*

$$\text{Area of } U = \int_{\partial U} \mathbf{F}\cdot d\mathbf{R}$$

where $\mathbf{F}(x,y) = \frac{1}{2}(-y,x), (0,x),$ or $(-y,0)$.

Proof: This follows immediately from Green's theorem.

Example 35.1.5 *Use Proposition 35.1.4 to find the area of the ellipse*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} \le 1.$$

You can parameterize the boundary of this ellipse as

$$x = a\cos t,\ y = b\sin t,\ t \in [0, 2\pi].$$

Then from Use Proposition 35.1.4,

$$\begin{aligned} \text{Area equals} &= \frac{1}{2}\int_0^{2\pi} (-b\sin t, a\cos t)\cdot(-a\sin t, b\cos t)\, dt \\ &= \frac{1}{2}\int_0^{2\pi} (ab)\, dt = \pi ab. \end{aligned}$$

Example 35.1.6 *Find $\int_{\partial U} \mathbf{F}\cdot d\mathbf{R}$ where U is the set, $\{(x,y) : x^2 + 3y^2 \le 9\}$ and $\mathbf{F}(x,y) = (y,-x)$.*

[1] For a general version see the advanced calculus book by Apostol. The general versions involve the concept of a rectifiable Jordan curve.

One way to do this is to parameterize the boundary of U and then compute the line integral directly. It is easier to use Green's theorem. The desired line integral equals

$$\int\int_U ((-1)-1)\,dA = -2\int\int_U dA.$$

Now U is an ellipse having area equal to $3\sqrt{3}$ and so the answer is $-6\sqrt{3}$.

Example 35.1.7 *Find* $\int_{\partial U}\mathbf{F}\cdot d\mathbf{R}$ *where* U *is the set,* $\{(x,y): 2\le x\le 4, 0\le y\le 3\}$ *and* $\mathbf{F}(x,y)=\left(x\sin y, y^3\cos x\right)$.

From Green's theorem this line integral equals

$$\begin{aligned}&\int_2^4\int_0^3 \left(-y^3\sin x - x\cos y\right)dydx\\ =\;& \frac{81}{4}\cos 4 - 6\sin 3 - \frac{81}{4}\cos 2.\end{aligned}$$

This is much easier than computing the line integral because you don't have to break the boundary in pieces and consider each separately.

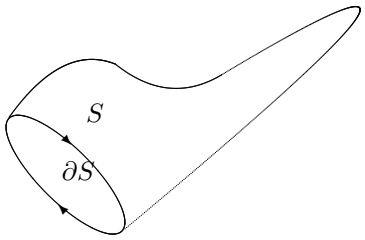
Example 35.1.8 *Find* $\int_{\partial U}\mathbf{F}\cdot d\mathbf{R}$ *where* U *is the set,* $\{(x,y): 2\le x\le 4, x\le y\le 3\}$ *and* $\mathbf{F}(x,y)=(x\sin y, y\sin x)$.

From Green's theorem this line integral equals

$$\begin{aligned}&\int_2^4\int_x^3 (y\cos x - x\cos y)\,dydx\\ =\;& -\frac{3}{2}\sin 4 - 6\sin 3 - 8\cos 4 - \frac{9}{2}\sin 2 + 4\cos 2.\end{aligned}$$

35.2 Stoke's Theorem From Green's Theorem

Stoke's theorem is a generalization of Green's theorem which relates the integral over a surface to the integral around the boundary of the surface. These terms are a little different from what occurs in $\mathbb{R}^2$. To describe this, consider a sock. The surface is the sock and its boundary will be the edge of the opening of the sock in which you place your foot. Another way to think of this is to imagine a region in $\mathbb{R}^2$ of the sort discussed above for Green's theorem. Suppose it is on a sheet of rubber and the sheet of rubber is stretched in three dimensions. The boundary of the resulting surface is the result of the stretching applied to the boundary of the original region in $\mathbb{R}^2$. Here is a picture describing the situation.



Recall the following definition of the curl of a vector field.

Definition 35.2.1 *Let*

$$\mathbf{F}(x,y,z) = (F_1(x,y,z), F_2(x,y,z), F_3(x,y,z))$$

be a C^1 vector field defined on an open set, V in $\mathbb{R}^3$. Then

$$\nabla \times \mathbf{F} \equiv \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{vmatrix}$$

$$\equiv \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right)\mathbf{i} + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right)\mathbf{j} + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right)\mathbf{k}.$$

This is also called $\operatorname{curl}(\mathbf{F})$ *and written as indicated,* $\nabla \times \mathbf{F}$.

The following lemma gives the fundamental identity which will be used in the proof of Stoke's theorem.

Lemma 35.2.2 *Let $\mathbf{R} : U \to V \subseteq \mathbb{R}^3$ where U is an open subset of $\mathbb{R}^2$ and V is an open subset of $\mathbb{R}^3$. Suppose $\mathbf{R}$ is C^2 and let $\mathbf{F}$ be a C^1 vector field defined in V.*

$$(\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times \mathbf{F})(\mathbf{R}(u,v)) = ((\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u)(u,v). \tag{35.3}$$

Proof: Start with the left side and let $x_i = R_i(u,v)$ for short.

$$\begin{aligned} (\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times \mathbf{F})(\mathbf{R}(u,v)) &= \varepsilon_{ijk} x_{ju} x_{kv} \varepsilon_{irs} \frac{\partial F_s}{\partial x_r} \\ &= (\delta_{jr}\delta_{ks} - \delta_{js}\delta_{kr}) x_{ju} x_{kv} \frac{\partial F_s}{\partial x_r} \\ &= x_{ju} x_{kv} \frac{\partial F_k}{\partial x_j} - x_{ju} x_{kv} \frac{\partial F_j}{\partial x_k} \\ &= \mathbf{R}_v \cdot \frac{\partial (\mathbf{F} \circ \mathbf{R})}{\partial u} - \mathbf{R}_u \cdot \frac{\partial (\mathbf{F} \circ \mathbf{R})}{\partial v} \end{aligned}$$

which proves (35.3).

The proof of Stoke's theorem given next follows [7]. First, it is convenient to give a definition.

Definition 35.2.3 *A vector valued function, $\mathbf{R} : U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ is said to be in $C^k(\overline{U}, \mathbb{R}^n)$ if it is the restriction to $\overline{U}$ of a vector valued function which is defined on $\mathbb{R}^m$ and is C^k. That is this function has continuous partial derivatives up to order k.*

Theorem 35.2.4 *(Stoke's Theorem) Let U be any region in $\mathbb{R}^2$ for which the conclusion of Green's theorem holds and let $\mathbf{R} \in C^2(\overline{U}, \mathbb{R}^3)$ be a one to one function satisfying $|(\mathbf{R}_u \times \mathbf{R}_v)(u,v)| \neq 0$ for all $(u,v) \in U$ and let S denote the surface,*

$$\begin{aligned} S &\equiv \{\mathbf{R}(u,v) : (u,v) \in U\}, \\ \partial S &\equiv \{\mathbf{R}(u,v) : (u,v) \in \partial U\} \end{aligned}$$

where the orientation on ∂S is consistent with the counter clockwise orientation on ∂U (U is on the left as you walk around ∂U). Then for $\mathbf{F}$ a C^1 vector field defined near S,

$$\int_{\partial S} \mathbf{F} \cdot d\mathbf{R} = \int\int_S \operatorname{curl}(\mathbf{F}) \cdot \mathbf{n} dS$$

where $\mathbf{n}$ is the normal to S defined by

$$\mathbf{n} \equiv \frac{\mathbf{R}_u \times \mathbf{R}_v}{|\mathbf{R}_u \times \mathbf{R}_v|}.$$

Proof: Letting C be an oriented part of ∂U having parametrization, $\mathbf{r}(t) \equiv (u(t), v(t))$ for $t \in [\alpha, \beta]$ and letting $\mathbf{R}(C)$ denote the oriented part of ∂S corresponding to C,

$$\int_{\mathbf{R}(C)} \mathbf{F} \cdot d\mathbf{R} =$$

$$\begin{aligned}
&= \int_\alpha^\beta \mathbf{F}(\mathbf{R}(u(t), v(t))) \cdot \langle \mathbf{R}_u u'(t) + \mathbf{R}_v v'(t) \rangle dt \\
&= \int_\alpha^\beta \mathbf{F}(\mathbf{R}(u(t), v(t))) \mathbf{R}_u(u(t), v(t)) u'(t)\, dt \\
&\quad + \int_\alpha^\beta \mathbf{F}(\mathbf{R}(u(t), v(t))) \mathbf{R}_v(u(t), v(t)) v'(t)\, dt \\
&= \int_C \langle (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u, (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v \rangle \cdot d\mathbf{r}.
\end{aligned}$$

Since this holds for each such piece of ∂U, it follows

$$\int_{\partial S} \mathbf{F} \cdot d\mathbf{R} = \int_{\partial U} \langle (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u, (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v \rangle \cdot d\mathbf{r}.$$

By the assumption that the conclusion of Green's theorem holds for U, this equals

$$\begin{aligned}
&\int\int_U [((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v)_u - ((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u)_v]\, dA \\
= &\int\int_U [(\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v + (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_{vu} - (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_{uv} - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u]\, dA \\
= &\int\int_U [(\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u]\, dA
\end{aligned}$$

the last step holding by equality of mixed partial derivatives, a result of the assumption that $\mathbf{R}$ is C^2. Now by Lemma 35.2.2, this equals

$$\begin{aligned}
&\int\int_U (\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times \mathbf{F})\, dA \\
= &\int\int_U \nabla \times \mathbf{F} \cdot (\mathbf{R}_u \times \mathbf{R}_v)\, dA \\
= &\int\int_S \nabla \times \mathbf{F} \cdot \mathbf{n} dS
\end{aligned}$$

because $dS = |(\mathbf{R}_u \times \mathbf{R}_v)|\, dA$ and $\mathbf{n} = \frac{(\mathbf{R}_u \times \mathbf{R}_v)}{|(\mathbf{R}_u \times \mathbf{R}_v)|}$. Thus

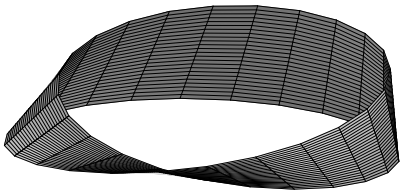
$$\begin{aligned}
(\mathbf{R}_u \times \mathbf{R}_v)\, dA &= \frac{(\mathbf{R}_u \times \mathbf{R}_v)}{|(\mathbf{R}_u \times \mathbf{R}_v)|} |(\mathbf{R}_u \times \mathbf{R}_v)|\, dA \\
&= \mathbf{n} dS.
\end{aligned}$$

This proves Stoke's theorem.

Note that there is no mention made in the final result that $\mathbf{R}$ is C^2. Therefore, it is not surprising that versions of this theorem are valid in which this assumption is not present. It is possible to obtain extremely general versions of Stoke's theorem if you use the Lebesgue integral.

35.2.1 Orientation

It turns out there are more general formulations of Stoke's theorem than what is presented above. However, it is always necessary for the surface, S to be orientable. This means it is possible to obtain a vector field for a unit normal to the surface which is a continuous function of position on S. An example of a surface which is not orientable is the famous Mobeus band, obtained by taking a long rectangular piece of paper and glueing the ends together after putting a twist in it. Here is a picture of one.



There is something quite interesting about this Mobeus band and this is that it can be written parametrically with a simple parameter domain. The picture above is a maple graph of the parametrically defined surface

$$\mathbf{R}(\theta, v) \equiv \begin{cases} x = 4\cos\theta + v\cos\frac{\theta}{2} \\ y = 4\sin\theta + v\cos\frac{\theta}{2} \\ z = v\sin\frac{\theta}{2} \end{cases}, \theta \in [0, 2\pi], v \in [-1, 1].$$

An obvious question is why the normal vector, $\mathbf{R}_{,\theta} \times \mathbf{R}_{,v} / |\mathbf{R}_{,\theta} \times \mathbf{R}_{,v}|$ is not a continuous function of position on S. You can see easily that it is a continuous function of both θ and v. However, the map, $\mathbf{R}$ is not one to one. In fact, $\mathbf{R}(0,0) = \mathbf{R}(2\pi, 0)$. Therefore, near this point on S, there are two different values for the above normal vector. In fact, a short computation will show this normal vector is

$$\frac{\left(4\sin\frac{1}{2}\theta\cos\theta - \frac{1}{2}v, 4\sin\frac{1}{2}\theta\sin\theta + \frac{1}{2}v, -8\cos^2\frac{1}{2}\theta\sin\frac{1}{2}\theta - 8\cos^3\frac{1}{2}\theta + 4\cos\frac{1}{2}\theta\right)}{\sqrt{16\sin^2\left(\frac{\theta}{2}\right) + \frac{v^2}{2} + 4\sin\left(\frac{\theta}{2}\right)v(\sin\theta - \cos\theta) + \left(-8\cos^2\frac{1}{2}\theta\sin\frac{1}{2}\theta - 8\cos^3\frac{1}{2}\theta + 4\cos\frac{1}{2}\theta\right)^2}}$$

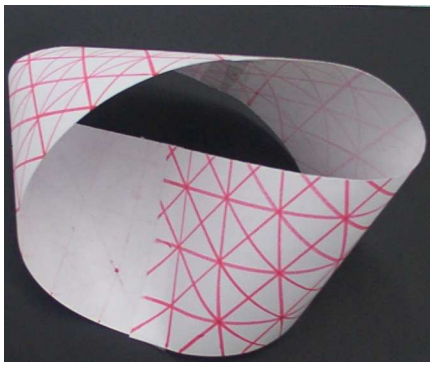
and you can verify that the denominator will not vanish. Letting $v = 0$ and $\theta = 0$ and 2π yields the two vectors,

$$(0, 0, -1), (0, 0, 1)$$

so there is a discontinuity. This is why I was careful to say in the statement of Stoke's theorem given above that $\mathbf{R}$ is one to one.

The Mobeus band has some usefulness. In old machine shops the equipment was run by a belt which was given a twist to spread the surface wear on the belt over twice the area.

The above explanation shows that $\mathbf{R}_{,\theta} \times \mathbf{R}_{,v} / |\mathbf{R}_{,\theta} \times \mathbf{R}_{,v}|$ fails to deliver an orientation for the Mobeus band. However, this does not answer the question whether there is some orientation for it other than this one. In fact there is none. You can see this by looking at the first of the two pictures below or by making one and tracing it with a pencil. There is only one side to the Mobeus band. An oriented surface must have two sides, one side identified by the given unit normal which varies continuously over the surface and the other side identified by the negative of this normal. The second picture below was taken by Dr. Ouyang when he was at meetings in Paris and saw it at a museum.




35.2.2 Conservative Vector Fields

Definition 35.2.5 *A vector field,* $\mathbf{F}$ *defined in a three dimensional region is said to be conservative*[2] *if for every piecewise smooth closed curve,* C*, it follows* $\int_C \mathbf{F}\cdot d\mathbf{R} = 0$.

Definition 35.2.6 *Let* $(\mathbf{x}, \mathbf{p}_1, \cdots, \mathbf{p}_n, \mathbf{y})$ *be an ordered list of points in* $\mathbb{R}^p$*. Let*

$$\mathbf{p}(\mathbf{x}, \mathbf{p}_1, \cdots, \mathbf{p}_n, \mathbf{y})$$

denote the piecewise smooth curve consisting of a straight line segment from $\mathbf{x}$ *to* $\mathbf{p}_1$ *and then the straight line segment from* $\mathbf{p}_1$ *to* $\mathbf{p}_2 \cdots$ *and finally the straight line segment from* $\mathbf{p}_n$ *to* $\mathbf{y}$*. This is called a polygonal curve. An open set in* $\mathbb{R}^p$*,* U*, is said to be a region if it has the property that for any two points,* $\mathbf{x}, \mathbf{y} \in U$*, there exists a polygonal curve joining the two points.*

Conservative vector fields are important because of the following theorem, sometimes called the fundamental theorem for line integrals.

Theorem 35.2.7 *Let* U *be a region in* $\mathbb{R}^p$ *and let* $\mathbf{F} : U \to \mathbb{R}^p$ *be a continuous vector field. Then* $\mathbf{F}$ *is conservative if and only if there exists a scalar valued function of* p *variables,* ϕ *such that* $\mathbf{F} = \nabla\phi$*. Furthermore, if* C *is an oriented curve which goes from* $\mathbf{x}$ *to* $\mathbf{y}$ *in* U*, then*

$$\int_C \mathbf{F}\cdot\, d\mathbf{R} = \phi(\mathbf{y}) - \phi(\mathbf{x}). \tag{35.4}$$

Thus the line integral is path independent in this case. This function, ϕ *is called a scalar potential for* $\mathbf{F}$*.*

Proof: To save space and fussing over things which are unimportant, denote by $\mathbf{p}(\mathbf{x}_0, \mathbf{x})$ a polygonal curve from $\mathbf{x}_0$ to $\mathbf{x}$. Thus the orientation is such that it goes from $\mathbf{x}_0$ to $\mathbf{x}$. The curve $\mathbf{p}(\mathbf{x}, \mathbf{x}_0)$ denotes the same set of points but in the opposite order. Suppose first $\mathbf{F}$ is conservative. Fix $\mathbf{x}_0 \in U$ and let

$$\phi(\mathbf{x}) \equiv \int_{\mathbf{p}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R}.$$

[2]There is no such thing as a liberal vector field.

This is well defined because if $\mathbf{q}(\mathbf{x}_0, \mathbf{x})$ is another polygonal curve joining $\mathbf{x}_0$ to $\mathbf{x}$, Then the curve obtained by following $\mathbf{p}(\mathbf{x}_0, \mathbf{x})$ from $\mathbf{x}_0$ to $\mathbf{x}$ and then from $\mathbf{x}$ to $\mathbf{x}_0$ along $\mathbf{q}(\mathbf{x}, \mathbf{x}_0)$ is a closed piecewise smooth curve and so by assumption, the line integral along this closed curve equals 0. However, this integral is just

$$\int_{\mathbf{p}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R} + \int_{\mathbf{q}(\mathbf{x},\mathbf{x}_0)} \mathbf{F}\cdot d\mathbf{R} = \int_{\mathbf{p}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R} - \int_{\mathbf{q}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R}$$

which shows

$$\int_{\mathbf{p}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R} = \int_{\mathbf{q}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R}$$

and that ϕ is well defined. For small t,

$$\begin{aligned}\frac{\phi(\mathbf{x} + t\mathbf{e}_i) - \phi(\mathbf{x})}{t} &= \frac{\int_{\mathbf{p}(\mathbf{x}_0,\mathbf{x}+t\mathbf{e}_i)} \mathbf{F}\cdot d\mathbf{R} - \int_{\mathbf{p}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R}}{t} \\ &= \frac{\int_{\mathbf{p}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R} + \int_{\mathbf{p}(\mathbf{x},\mathbf{x}+t\mathbf{e}_i)} \mathbf{F}\cdot d\mathbf{R} - \int_{\mathbf{p}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R}}{t}.\end{aligned}$$

Since U is open, for small t, the ball of radius $|t|$ centered at $\mathbf{x}$ is contained in U. Therefore, the line segment from $\mathbf{x}$ to $\mathbf{x} + t\mathbf{e}_i$ is also contained in U and so one can take $\mathbf{p}(\mathbf{x}, \mathbf{x} + t\mathbf{e}_i)(s) = \mathbf{x} + s(t\mathbf{e}_i)$ for $s \in [0,1]$. Therefore, the above difference quotient reduces to

$$\begin{aligned}\frac{1}{t}\int_0^1 \mathbf{F}(\mathbf{x} + s(t\mathbf{e}_i)) \cdot t\mathbf{e}_i \, ds &= \int_0^1 F_i(\mathbf{x} + s(t\mathbf{e}_i)) \, ds \\ &= F_i(\mathbf{x} + s_t(t\mathbf{e}_i))\end{aligned}$$

by the mean value theorem for integrals. Here s_t is some number between 0 and 1. By continuity of $\mathbf{F}$, this converges to $F_i(\mathbf{x})$ as $t \to 0$. Therefore, $\nabla\phi = \mathbf{F}$ as claimed.

Conversely, if $\nabla\phi = \mathbf{F}$, then if $\mathbf{R} : [a,b] \to \mathbb{R}^p$ is any C^1 curve joining $\mathbf{x}$ to $\mathbf{y}$,

$$\begin{aligned}\int_a^b \mathbf{F}(\mathbf{R}(t)) \cdot \mathbf{R}'(t) \, dt &= \int_a^b \nabla\phi(\mathbf{R}(t)) \cdot \mathbf{R}'(t) \, dt \\ &= \int_a^b \frac{d}{dt}(\phi(\mathbf{R}(t))) \, dt \\ &= \phi(\mathbf{R}(b)) - \phi(\mathbf{R}(a)) \\ &= \phi(\mathbf{y}) - \phi(\mathbf{x})\end{aligned}$$

and this verifies (35.4) in the case where the curve joining the two points is smooth. The general case follows immediately from this by using this result on each of the pieces of the piecewise smooth curve. For example if the curve goes from $\mathbf{x}$ to $\mathbf{p}$ and then from $\mathbf{p}$ to $\mathbf{y}$, the above would imply the integral over the curve from $\mathbf{x}$ to $\mathbf{p}$ is $\phi(\mathbf{p}) - \phi(\mathbf{x})$ while from $\mathbf{p}$ to $\mathbf{y}$ the integral would yield $\phi(\mathbf{y}) - \phi(\mathbf{p})$. Adding these gives $\phi(\mathbf{y}) - \phi(\mathbf{x})$. The formula (35.4) implies the line integral over any closed curve equals zero because the starting and ending points of such a curve are the same. This proves the theorem.

Example 35.2.8 *Let* $\mathbf{F}(x,y,z) = (\cos x - yz\sin(xz), \cos(xz), -yx\sin(xz))$. *Let* C *be a piecewise smooth curve which goes from* $(\pi, 1, 1)$ *to* $\left(\frac{\pi}{2}, 3, 2\right)$. *Find* $\int_C \mathbf{F}\cdot d\mathbf{R}$.

The specifics of the curve are not given so the problem is nonsense unless the vector field is conservative. Therefore, it is reasonable to look for the function, ϕ satisfying $\nabla\phi = \mathbf{F}$. Such a function satisfies

$$\phi_x = \cos x - y(\sin xz)\, z$$

and so, assuming ϕ exists,

$$\phi(x,y,z) = \sin x + y\cos(xz) + \psi(y,z).$$

I have to add in the most general thing possible, $\psi(y,z)$ to ensure possible solutions are not being thrown out. It wouldn't be good at this point to add in a constant since the answer could involve a function of either or both of the other variables. Now from what was just obtained,

$$\phi_y = \cos(xz) + \psi_y = \cos xz$$

and so it is possible to take $\psi_y = 0$. Consequently, ϕ, if it exists is of the form

$$\phi(x,y,z) = \sin x + y\cos(xz) + \psi(z).$$

Now differentiating this with respect to z gives

$$\phi_z = -yx\sin(xz) + \psi_z = -yx\sin(xz)$$

and this shows ψ does not depend on z either. Therefore, it suffices to take $\psi = 0$ and

$$\phi(x,y,z) = \sin(x) + y\cos(xz).$$

Therefore, the desired line integral equals

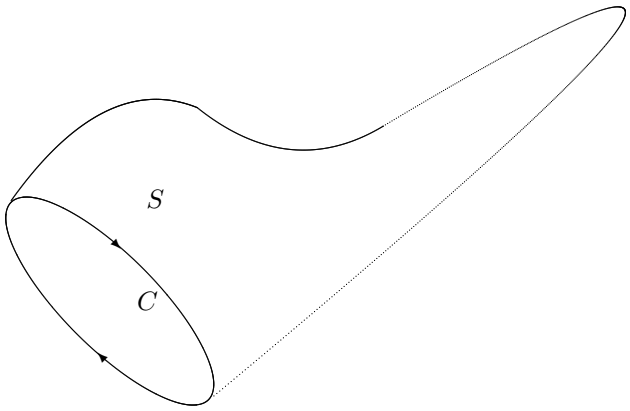
$$\sin\left(\frac{\pi}{2}\right) + 3\cos(\pi) - (\sin(\pi) + \cos(\pi)) = -1.$$

The above process for finding ϕ will not lead you astray in the case where there does not exist a scalar potential. As an example, consider the following.

Example 35.2.9 *Let* $\mathbf{F}(x,y,z) = \left(x, y^2x, z\right)$. *Find a scalar potential for* $\mathbf{F}$ *if it exists.*

If ϕ exists, then $\phi_x = x$ and so $\phi = \frac{x^2}{2} + \psi(y,z)$. Then $\phi_y = \psi_y(y,z) = xy^2$ but this is impossible because the left side depends only on y and z while the right side depends also on x. Therefore, this vector field is not conservative and there does not exist a scalar potential.

Definition 35.2.10 *A set of points in three dimensional space,* V *is simply connected if every piecewise smooth closed curve,* C *is the edge of a surface,* S *which is contained entirely within* V *in such a way that Stokes theorem holds for the surface,* S *and its edge,* C.



This is like a sock. The surface is the sock and the curve, C goes around the opening of the sock.

As an application of Stoke's theorem, here is a useful theorem which gives a way to check whether a vector field is conservative.

Theorem 35.2.11 *For a three dimensional simply connected open set, V and $\mathbf{F}$ a C^1 vector field defined in V, $\mathbf{F}$ is conservative if $\nabla \times \mathbf{F} = \mathbf{0}$ in V.*

Proof: If $\nabla \times \mathbf{F} = \mathbf{0}$ then taking an arbitrary closed curve, C, and letting S be a surface bounded by C which is contained in V, Stoke's theorem implies

$$0 = \int_S \nabla \times \mathbf{F} \cdot \mathbf{n}\, dA = \int_C \mathbf{F} \cdot d\mathbf{R}.$$

Thus $\mathbf{F}$ is conservative.

Example 35.2.12 *Determine whether the vector field,*

$$\left(4x^3 + 2\left(\cos\left(x^2+z^2\right)\right)x, 1, 2\left(\cos\left(x^2+z^2\right)\right)z\right)$$

is conservative.

Since this vector field is defined on all of $\mathbb{R}^3$, it only remains to take its curl and see if it is the zero vector.

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial_x & \partial_y & \partial_z \\ 4x^3 + 2\left(\cos\left(x^2+z^2\right)\right)x & 1 & 2\left(\cos\left(x^2+z^2\right)\right)z \end{vmatrix}.$$

This is obviously equal to zero. Therefore, the given vector field is conservative. Can you find a potential function for it? Let ϕ be the potential function. Then $\phi_z = 2\left(\cos\left(x^2+z^2\right)\right)z$ and so $\phi(x,y,z) = \sin\left(x^2+z^2\right) + g(x,y)$. Now taking the derivative of ϕ with respect to y, you see $g_y = 1$ so $g(x,y) = y + h(x)$. Hence $\phi(x,y,z) = y + g(x) + \sin\left(x^2+z^2\right)$. Taking the derivative with respect to x, you get $4x^3 + 2\left(\cos\left(x^2+z^2\right)\right)x = g'(x) + 2x\cos\left(x^2+z^2\right)$ and so it suffices to take $g(x) = x^4$. Hence $\phi(x,y,z) = y + x^4 + \sin\left(x^2+z^2\right)$.

35.2.3 Some Terminology

If $\mathbf{F} = (P,Q,R)$ is a vector field. Then the statement that $\mathbf{F}$ is conservative is the same as saying the differential form $Pdx + Qdy + Rdz$ is exact. Some people like to say things in terms of vector fields and some say it in terms of differential forms. In Example 35.2.12, the differential form $\left(4x^3 + 2\left(\cos\left(x^2+z^2\right)\right)x\right)dx + dy + \left(2\left(\cos\left(x^2+z^2\right)\right)z\right)dz$ is exact.

35.2.4 Maxwell's Equations And The Wave Equation

Many of the ideas presented above are useful in analyzing Maxwell's equations. These equations are derived in advanced physics courses. They are

$$\nabla \times \mathbf{E} + \frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} = \mathbf{0} \tag{35.5}$$

$$\nabla \cdot \mathbf{E} = 4\pi\rho \tag{35.6}$$

$$\nabla \times \mathbf{B} - \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} = \frac{4\pi}{c}\mathbf{f} \tag{35.7}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{35.8}$$

and it is assumed these hold on all of $\mathbb{R}^3$ to eliminate technical considerations having to do with whether something is simply connected.

In these equations, $\mathbf{E}$ is the electrostatic field and $\mathbf{B}$ is the magnetic field while ρ and $\mathbf{f}$ are sources. By (35.8) $\mathbf{B}$ has a vector potential, $\mathbf{A}_1$ such that $\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A}_1$. Now go to (35.5) and write

$$\nabla \times \mathbf{E} + \frac{1}{c} \nabla \times \frac{\partial \mathbf{A}_1}{\partial t} = \mathbf{0}$$

showing that

$$\nabla \times \left(\mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{A}_1}{\partial t} \right) = \mathbf{0}$$

It follows $\mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{A}_1}{\partial t}$ has a scalar potential, ψ_1 satisfying

$$\nabla \psi_1 = \mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{A}_1}{\partial t}. \tag{35.9}$$

Now suppose ϕ is a time dependent scalar field satisfying

$$\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = \frac{1}{c} \frac{\partial \psi_1}{\partial t} - \nabla \cdot \mathbf{A}_1. \tag{35.10}$$

Next define

$$\mathbf{A} \equiv \mathbf{A}_1 + \nabla \phi, \quad \psi \equiv \psi_1 + \frac{1}{c} \frac{\partial \phi}{\partial t}. \tag{35.11}$$

Therefore, in terms of the new variables, (35.10) becomes

$$\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = \frac{1}{c} \left(\frac{\partial \psi}{\partial t} - \frac{1}{c} \frac{\partial^2 \phi}{\partial t^2} \right) - \nabla \cdot A + \nabla^2 \phi$$

which yields

$$0 = \frac{\partial \psi}{\partial t} - c \nabla \cdot A. \tag{35.12}$$

Then it follows from Theorem 34.1.3 on Page 764 that $\mathbf{A}$ is also a vector potential for $\mathbf{B}$. That is

$$\nabla \times \mathbf{A} = \mathbf{B}. \tag{35.13}$$

From (35.9)

$$\nabla \left(\psi - \frac{1}{c} \frac{\partial \phi}{\partial t} \right) = \mathbf{E} + \frac{1}{c} \left(\frac{\partial A}{\partial t} - \nabla \frac{\partial \phi}{\partial t} \right)$$

and so

$$\nabla \psi = \mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{A}}{\partial t}. \tag{35.14}$$

Using (35.7) and (35.14),

$$\nabla \times (\nabla \times \mathbf{A}) - \frac{1}{c} \frac{\partial}{\partial t} \left(\nabla \psi - \frac{1}{c} \frac{\partial \mathbf{A}}{\partial t} \right) = \frac{4\pi}{c} \mathbf{f}. \tag{35.15}$$

Now from Theorem 34.1.3 on Page 764 this implies

$$\nabla (\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} - \nabla \left(\frac{1}{c} \frac{\partial \psi}{\partial t} \right) + \frac{1}{c^2} \frac{\partial^2 \mathbf{A}}{\partial t^2} = \frac{4\pi}{c} \mathbf{f}$$

and using (35.12), this gives

$$\frac{1}{c^2} \frac{\partial^2 \mathbf{A}}{\partial t^2} - \nabla^2 \mathbf{A} = \frac{4\pi}{c} \mathbf{f}. \tag{35.16}$$

Also from (35.14), (35.6), and (35.12),

$$\begin{aligned}\nabla^2\psi &= \nabla\cdot\mathbf{E}+\frac{1}{c}\frac{\partial}{\partial t}(\nabla\cdot\mathbf{A})\\ &= 4\pi\rho+\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2}\end{aligned}$$

and so

$$\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2}-\nabla^2\psi=-4\pi\rho. \tag{35.17}$$

This is very interesting. If a solution to the wave equations, (35.17), and (35.16) can be found along with a solution to (35.12), then letting the magnetic field be given by (35.13) and letting $\mathbf{E}$ be given by (35.14) the result is a solution to Maxwells equations. This is significant because wave equations are easier to think of than Maxwell's equations. Note the above argument also showed that it is always possible, by solving another wave equation, to get (35.12) to hold.

35.3 Exercises

1. Determine whether the vector field, $\left(2xy^3\sin z^4, 3x^2y^2\sin z^4+1, 4x^2y^3\left(\cos z^4\right)z^3+1\right)$ is conservative. If it is conservative, find a potential function.
2. Determine whether the vector field, $\left(2xy^3\sin z+y^2+z, 3x^2y^2\sin z+2xy, x^2y^3\cos z+x\right)$ is conservative. If it is conservative, find a potential function.
3. Determine whether the vector field, $\left(2xy^3\sin z+z, 3x^2y^2\sin z+2xy, x^2y^3\cos z+x\right)$ is conservative. If it is conservative, find a potential function.
4. Find scalar potentials for the following vector fields if it is possible to do so. If it is not possible to do so, explain why.
   (a) $\left(y^2, 2xy+\sin z, 2z+y\cos z\right)$
   (b) $\left(2z\left(\cos\left(x^2+y^2\right)\right)x, 2z\left(\cos\left(x^2+y^2\right)\right)y, \sin\left(x^2+y^2\right)+2z\right)$
   (c) $\left(f(x), g(y), h(z)\right)$
   (d) $\left(xy, z^2, y^3\right)$
   (e) $\left(z+2\frac{x}{x^2+y^2+1}, 2\frac{y}{x^2+y^2+1}, x+3z^2\right)$
5. If a vector field is not conservative on the set U, is it possible the same vector field could be conservative on some subset of U? Explain and give examples if it is possible. If it is not possible also explain why.
6. Prove that if a vector field, $\mathbf{F}$ has a scalar potential, then it has infinitely many scalar potentials.
7. Here is a vector field: $\mathbf{F}\equiv\left(2xy, x^2-5y^4, 3z^2\right)$. Find $\int_C\mathbf{F}\cdot d\mathbf{R}$ where C is a curve which goes from $(1,2,3)$ to $(4,-2,1)$.
8. Here is a vector field: $\mathbf{F}\equiv\left(2xy, x^2-5y^4, 3\left(\cos z^3\right)z^2\right)$. Find $\int_C\mathbf{F}\cdot d\mathbf{R}$ where C is a curve which goes from $(1,0,1)$ to $(-4,-2,1)$.
9. Find $\int_{\partial U}\mathbf{F}\cdot d\mathbf{R}$ where U is the set, $\{(x,y): 2\le x\le 4, 0\le y\le x\}$ and $\mathbf{F}(x,y)=(x\sin y, y\sin x)$.

10. Find $\int_{\partial U} \mathbf{F}\cdot d\mathbf{R}$ where U is the set, $\left\{(x,y) : 2 \leq x \leq 3, 0 \leq y \leq x^2\right\}$ and $\mathbf{F}(x,y) = (x\cos y, y+x)$.

11. Find $\int_{\partial U} \mathbf{F}\cdot d\mathbf{R}$ where U is the set, $\{(x,y) : 1 \leq x \leq 2, x \leq y \leq 3\}$ and $\mathbf{F}(x,y) = (x\sin y, y\sin x)$.

12. Find $\int_{\partial U} \mathbf{F}\cdot d\mathbf{R}$ where U is the set, $\left\{(x,y) : x^2+y^2 \leq 2\right\}$ and $\mathbf{F}(x,y) = \left(-y^3, x^3\right)$.

13. Show that for many open sets in $\mathbb{R}^2$, Area of $U = \int_{\partial U} x dy$, and Area of $U = \int_{\partial U} -y dx$ and Area of $U = \frac{1}{2}\int_{\partial U} -y dx + x dy$. **Hint:** Use Green's theorem.

14. Two smooth oriented surfaces, S_1 and S_2 intersect in a piecewise smooth oriented closed curve, C. Let $\mathbf{F}$ be a C^1 vector field defined on $\mathbb{R}^3$. Explain why $\int\int_{S_1} \operatorname{curl}(\mathbf{F})\cdot \mathbf{n}\, dS = \int\int_{S_2} \operatorname{curl}(\mathbf{F})\cdot \mathbf{n}\, dS$. Here $\mathbf{n}$ is the normal to the surface which corresponds to the given orientation of the curve, C.

15. Show that $\operatorname{curl}(\psi\nabla\phi) = \nabla\psi\times\nabla\phi$ and explain why $\int\int_S \nabla\psi\times\nabla\phi\cdot\mathbf{n}\, dS = \int_{\partial S}(\psi\nabla\phi)\cdot d\mathbf{r}$.

16. Find a simple formula for $\operatorname{div}(\nabla(u^\alpha))$ where $\alpha \in \mathbb{R}$.

17. Parametric equations for one arch of a cycloid are given by $x = a(t-\sin t)$ and $y = a(1-\cos t)$ where here $t \in [0, 2\pi]$. Sketch a rough graph of this arch of a cycloid and then find the area between this arch and the x axis. **Hint:** This is very easy using Green's theorem and the vector field, $\mathbf{F} = (-y, x)$.

18. Let $\mathbf{r}(t) = \left(\cos^3(t), \sin^3(t)\right)$ where $t \in [0, 2\pi]$. Sketch this curve and find the area enclosed by it using Green's theorem.

19. Consider the vector field, $\left(\frac{-y}{(x^2+y^2)}, \frac{x}{(x^2+y^2)}, 0\right) = \mathbf{F}$. Show that $\nabla\times\mathbf{F} = \mathbf{0}$ but that for the closed curve, whose parameterization is $\mathbf{R}(t) = (\cos t, \sin t, 0)$ for $t \in [0, 2\pi]$, $\int_C \mathbf{F}\cdot d\mathbf{R} \neq 0$. Therefore, the vector field is not conservative. Does this contradict Theorem 35.2.11? Explain.

20. Let $\mathbf{x}$ be a point of $\mathbb{R}^3$ and let $\mathbf{n}$ be a unit vector. Let D_r be the circular disk of radius r containing $\mathbf{x}$ which is perpendicular to $\mathbf{n}$. Placing the tail of $\mathbf{n}$ at $\mathbf{x}$ and viewing D_r from the point of $\mathbf{n}$, orient ∂D_r in the counter clockwise direction. Now suppose $\mathbf{F}$ is a vector field defined near $\mathbf{x}$. Show $\operatorname{curl}(\mathbf{F})\cdot\mathbf{n} = \lim_{r\to 0}\frac{1}{\pi r^2}\int_{\partial D_r}\mathbf{F}\cdot d\mathbf{R}$. This last integral is sometimes called the circulation density of $\mathbf{F}$. Explain how this shows that $\operatorname{curl}(\mathbf{F})\cdot\mathbf{n}$ measures the tendency for the vector field to "curl" around the point, the vector $\mathbf{n}$ at the point $\mathbf{x}$.

21. The cylinder $x^2+y^2 = 4$ is intersected with the plane $x+y+z = 2$. This yields a closed curve, C. Orient this curve in the counter clockwise direction when viewed from a point high on the z axis. Let $\mathbf{F} = \left(x^2y, z+y, x^2\right)$. Find $\int_C \mathbf{F}\cdot d\mathbf{R}$.

22. The cylinder $x^2+4y^2 = 4$ is intersected with the plane $x+3y+2z = 1$. This yields a closed curve, C. Orient this curve in the counter clockwise direction when viewed from a point high on the z axis. Let $\mathbf{F} = \left(y, z+y, x^2\right)$. Find $\int_C \mathbf{F}\cdot d\mathbf{R}$.

23. The cylinder $x^2+y^2 = 4$ is intersected with the plane $x+3y+2z = 1$. This yields a closed curve, C. Orient this curve in the clockwise direction when viewed from a point high on the z axis. Let $\mathbf{F} = (y, z+y, x)$. Find $\int_C \mathbf{F}\cdot d\mathbf{R}$.

24. Let $\mathbf{F} = \left(xz, z^2(y+\sin x), z^3y\right)$. Find the surface integral, $\int\int_S \operatorname{curl}(\mathbf{F})\cdot\mathbf{n}dA$ where S is the surface, $z = 4-\left(x^2+y^2\right)$, $z \geq 0$.

25. Let $\mathbf{F} = \left(xz, \left(y^3 + x\right), z^3 y\right)$. Find the surface integral, $\int\int_S \text{curl}\left(\mathbf{F}\right) \cdot \mathbf{n} dA$ where S is the surface, $z = 16 - \left(x^2 + y^2\right)$, $z \geq 0$.

26. The cylinder $z = y^2$ intersects the surface $z = 8 - x^2 - 4y^2$ in a curve, C which is oriented in the counter clockwise direction when viewed high on the z axis. Find $\int_C \mathbf{F} \cdot d\mathbf{R}$ if $\mathbf{F} = \left(\frac{z^2}{2}, xy, xz\right)$. **Hint:** This is not too hard if you show you can use Stokes theorem on a domain in the xy plane.

27. Suppose solutions have been found to (35.17), (35.16), and (35.12). Then define $\mathbf{E}$ and $\mathbf{B}$ using (35.14) and (35.13). Verify Maxwell's equations hold for $\mathbf{E}$ and $\mathbf{B}$.

28. Suppose now you have found solutions to (35.17) and (35.16), ψ_1 and A_1. Then go show again that if ϕ satisfies (35.10) and $\psi \equiv \psi_1 + \frac{1}{c}\frac{\partial \phi}{\partial t}$, while $\mathbf{A} \equiv \mathbf{A}_1 + \nabla \phi$, then (35.12) holds for $\mathbf{A}$ and ψ.

29. Why consider Maxwell's equations? Why not just consider (35.17), (35.16), and (35.12)?

30. Tell which open sets are simply connected.

    (a) The inside of a car radiator.

    (b) A donut.

    (c) The solid part of a cannon ball which contains a void on the interior.

    (d) The inside of a donut which has had a large bite taken out of it.

    (e) All of $\mathbb{R}^3$ except the z axis.

    (f) All of $\mathbb{R}^3$except the xy plane.

31. Let P be a polygon with vertices $\left(x_1, y_1\right), \left(x_2, y_2\right), \cdots, \left(x_n, y_n\right), \left(x_1, y_1\right)$ encountered as you move over the boundary of the polygon in the counter clockwise direction. Using Problem 13, find a nice formula for the area of the polygon in terms of the vertices.

Curvilinear Coordinates

36.0.1 Outcomes

1. Define the basis and dual basis for general curvilinear coordinates.

2. Recall and use the transformation equations relating quantities in different curvilinear coordinate systems.

3. Define and use the Christoffel symbols.

4. Recall and use the formula for divergence and gradient in curvilinear coordinate systems.

You have already seen examples of curvilinear coordinates in the case of cylindrical and spherical coordinates. You use other coordinates other than x, y, z the rectangular coordinates to identify points. A general situation is the following: Let $D \subseteq \mathbb{R}^n$ be an open set and let $\mathbf{M} : D \to \mathbb{R}^n$ satisfy

$$\mathbf{M} \text{ is } C^2, \tag{36.1}$$

$$\mathbf{M} \text{ is one to one.} \tag{36.2}$$

Letting $\mathbf{x} \in D$, we can write

$$\mathbf{M}(\mathbf{x}) = M^k(\mathbf{x})\,\mathbf{i}_k$$

where, $\mathbf{i}_k$ are the standard basis vectors for $\mathbb{R}^n$, $\mathbf{i}_k$ being the vector in $\mathbb{R}^n$ which has a one in the k^{th} coordinate and a 0 in every other spot. For a fixed $\mathbf{x} \in D$, we can consider the curves,

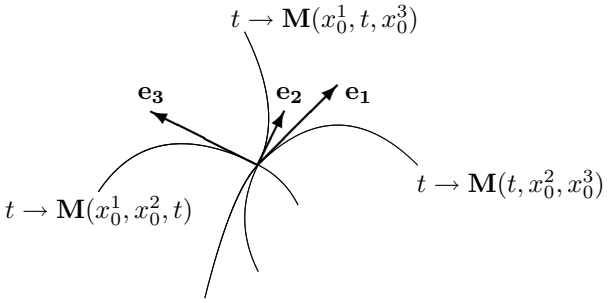
$$t \to \mathbf{M}(\mathbf{x} + t\mathbf{i}_k)$$

for $t \in I$, some open interval containing 0. Thus these curves are obtained by fixing all the variables except the k^{th} and then considering the curve which results. Then for the point $\mathbf{x}$,

$$\mathbf{e}_k \equiv \frac{\partial \mathbf{M}}{\partial x^k}(\mathbf{x}).$$

Denote this vector as $\mathbf{e}_k(\mathbf{x})$ to emphasize its dependence on $\mathbf{x}$. The following picture illus-

trates the situation in $\mathbb{R}^3$.



It is desired that $\{\mathbf{e}_k\}_{k=1}^n$ should be a basis. This holds if and only if

$$\det\left(\frac{\partial M^i}{\partial x^k}\right) \neq 0. \tag{36.3}$$

Let

$$y^i = M^i(\mathbf{x}) \;\; i = 1, \cdots, n \tag{36.4}$$

so that the y^i are the usual coordinates with respect to the usual basis vectors $\{\mathbf{i}_k\}_{k=1}^n$ of the point $\mathbf{M}(\mathbf{x})$. Letting $\mathbf{x} \equiv \left(x^1, \cdots, x^n\right)$, it follows from the inverse function theorem of advanced calculus that $\mathbf{M}(D)$ is open, and that (36.3), (36.1), and (36.2) imply the equations (36.4) define each x^i as a C^2 function of $\mathbf{y} \equiv \left(y^1, \cdots, y^n\right)^T$. Thus, abusing notation slightly, the equations (36.4) are equivalent to

$$x^i = x^i(\mathbf{y}), \;\; i = 1, \cdots, n$$

where x^i is a C^2 function. Thus

$$\nabla x^k(\mathbf{y}) = \frac{\partial x^k(\mathbf{y})}{\partial y^j}\mathbf{i}^j.$$

Then

$$\nabla x^k(\mathbf{y}) \cdot \mathbf{e}_j = \frac{\partial x^k}{\partial y^s}\mathbf{i}^s \cdot \frac{\partial y^r}{\partial x^j}\mathbf{i}_r = \frac{\partial x^k}{\partial y^s}\frac{\partial y^s}{\partial x^j} = \delta_j^k$$

by the chain rule. Therefore, the dual basis is given by

$$\mathbf{e}^k(\mathbf{x}) = \nabla x^k(\mathbf{y}). \tag{36.5}$$

Notice that it might be hard or even impossible to solve algebraically for x^i in terms of the y^j. Thus the straight forward approach to finding $\mathbf{e}^k$ by (36.5) might be impossible! Also, this approach leads to an expression in terms of the $\mathbf{y}$ coordinates rather than the desired $\mathbf{x}$ coordinates and so it is probably not a good idea to use it in the first place. It is expedient to use another method to obtain these vectors. The vectors, $\mathbf{e}^k(\mathbf{x})$ may always be found by raising the index using the inverse of the metric tensor as explained on Page 492 and the result is in terms of the curvilinear coordinates, $\mathbf{x}$. Consider the following example.

Example 36.0.1 *$D \equiv (0,\infty) \times (0,\pi) \times (0,2\pi)$ and*

$$\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 \sin\left(x^2\right)\cos\left(x^3\right) \\ x^1 \sin\left(x^2\right)\sin\left(x^3\right) \\ x^1 \cos\left(x^2\right) \end{pmatrix},$$

usually written as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \rho\sin(\phi)\cos(\theta) \\ \rho\sin(\phi)\sin(\theta) \\ \rho\cos(\phi) \end{pmatrix}$$

where (ρ,ϕ,θ) are the spherical coordinates. These coordinates are called x^1, x^2, and x^3 to preserve the notation just discussed.) Thus

$$\mathbf{e}_1(\mathbf{x}) = \sin\left(x^2\right)\cos\left(x^3\right)\mathbf{i}_1 + \sin\left(x^2\right)\sin\left(x^3\right)\mathbf{i}_2 + \cos\left(x^2\right)\mathbf{i}_3,$$

$$\mathbf{e}_2(\mathbf{x}) = x^1\cos\left(x^2\right)\cos\left(x^3\right)\mathbf{i}_1$$

$$+x^1\cos\left(x^2\right)\sin\left(x^3\right)\mathbf{i}_2 - x^1\sin\left(x^2\right)\mathbf{i}_3,$$

$$\mathbf{e}_3(\mathbf{x}) = -x^1\sin\left(x^2\right)\sin\left(x^3\right)\mathbf{i}_1 + x^1\sin\left(x^2\right)\cos\left(x^3\right)\mathbf{i}_2 + 0\mathbf{i}_3.$$

It follows the metric tensor is

$$G = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \left(x^1\right)^2 & 0 \\ 0 & 0 & \left(x^1\right)^2\sin^2\left(x^2\right) \end{pmatrix} = (g_{ij}) = (\mathbf{e}_i\cdot\mathbf{e}_j). \tag{36.6}$$

Therefore,

$$G^{-1} = \left(g^{ij}\right)$$

$$= \left(\mathbf{e}^i,\mathbf{e}^j\right) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \left(x^1\right)^{-2} & 0 \\ 0 & 0 & \left(x^1\right)^{-2}\sin^{-2}\left(x^2\right) \end{pmatrix}.$$

To obtain the dual basis, use Theorem 19.2.7 to write

$$\mathbf{e}^1 = g^{1j}\mathbf{e}_j = \mathbf{e}_1$$

$$\mathbf{e}^2 = g^{2j}\mathbf{e}_j = \left(x^1\right)^{-2}\mathbf{e}_2$$

$$\mathbf{e}^3 = g^{3j}\mathbf{e}_j = \left(x^1\right)^{-2}\sin^{-2}\left(x^2\right)\mathbf{e}_3.$$

It is natural to ask if there exists a transformation $\mathbf{M}$ such that

$$\frac{\partial\mathbf{M}}{\partial x^1} = \mathbf{i} = \mathbf{i}_1, \frac{\partial\mathbf{M}}{\partial x^2} = \mathbf{j} = \mathbf{i}_2, \frac{\partial\mathbf{M}}{\partial x^3} = \mathbf{k} = \mathbf{i}_3. \tag{36.7}$$

Let

$$\mathbf{M}\left(x^1,x^2,x^3\right) \equiv x^1\mathbf{i} + x^2\mathbf{j} + x^3\mathbf{k}.$$

Then (36.7) holds for this transformation.

Example 36.0.2 *Let*

$$\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} 3u+v \\ v-w \\ u-v^3 \end{pmatrix}$$

where the y^i are the rectangular coordinates of the point. Find $\mathbf{e}^i, \mathbf{e}_i, i = 1,2,3$, and find $(g_{ij})(\mathbf{x})$ and $\left(g^{ij}(\mathbf{x})\right)$.

First

$$\mathbf{e}_1 = \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 1 \\ 1 \\ -3v^2 \end{pmatrix}, \mathbf{e}_3 = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}.$$

Then the metric tensor is

$$\begin{pmatrix} 3 & 0 & 1 \\ 1 & 1 & -3v^2 \\ 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} 3 & 1 & 0 \\ 0 & 1 & -1 \\ 1 & -3v^2 & 0 \end{pmatrix} = \begin{pmatrix} 10 & 3-3v^2 & 0 \\ 3-3v^2 & 2+9v^4 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

Thus the inverse of the metric tensor is

$$\begin{pmatrix} \frac{1+9v^4}{1+81v^4+18v^2} & 3\frac{v^2-1}{1+81v^4+18v^2} & 3\frac{v^2-1}{1+81v^4+18v^2} \\ 3\frac{v^2-1}{1+81v^4+18v^2} & \frac{10}{1+81v^4+18v^2} & \frac{10}{1+81v^4+18v^2} \\ 3\frac{v^2-1}{1+81v^4+18v^2} & \frac{10}{1+81v^4+18v^2} & \frac{81v^4+18v^2+11}{1+81v^4+18v^2} \end{pmatrix}$$

and so the dual basis consists of the columns of

$$\begin{pmatrix} 3 & 1 & 0 \\ 0 & 1 & -1 \\ 1 & -3v^2 & 0 \end{pmatrix} \begin{pmatrix} \frac{1+9v^4}{1+81v^4+18v^2} & 3\frac{v^2-1}{1+81v^4+18v^2} & 3\frac{v^2-1}{1+81v^4+18v^2} \\ 3\frac{v^2-1}{1+81v^4+18v^2} & \frac{10}{1+81v^4+18v^2} & \frac{10}{1+81v^4+18v^2} \\ 3\frac{v^2-1}{1+81v^4+18v^2} & \frac{10}{1+81v^4+18v^2} & \frac{81v^4+18v^2+11}{1+81v^4+18v^2} \end{pmatrix} =$$

$$\begin{pmatrix} 3\frac{1+9v^4}{1+81v^4+18v^2} + 3\frac{v^2-1}{1+81v^4+18v^2} & 9\frac{v^2-1}{1+81v^4+18v^2} + \frac{10}{1+81v^4+18v^2} & 9\frac{v^2-1}{1+81v^4+18v^2} + \frac{10}{1+81v^4+18v^2} \\ 0 & 0 & \frac{10}{1+81v^4+18v^2} - \frac{81v^4+18v^2+11}{1+81v^4+18v^2} \\ \frac{1+9v^4}{1+81v^4+18v^2} - 9v^2\frac{v^2-1}{1+81v^4+18v^2} & 3\frac{v^2-1}{1+81v^4+18v^2} - 30\frac{v^2}{1+81v^4+18v^2} & 3\frac{v^2-1}{1+81v^4+18v^2} - 30\frac{v^2}{1+81v^4+18v^2} \end{pmatrix}$$

Clearly this is pretty horrible. That is because I picked a fairly arbitrary coordinate system.

36.1 Exercises

1. Let

$$\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 + 2x^2 \\ x^2 + x^3 \\ x^1 - 2x^2 \end{pmatrix}$$

where the y^i are the rectangular coordinates of the point. Find $\mathbf{e}^i, \mathbf{e}_i, i = 1, 2, 3$, and find $(g_{ij})(\mathbf{x})$ and $\left(g^{ij}(\mathbf{x})\right)$.

2. Let

$$\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 + x^2 \\ x^2 + x^3 \\ x^1 + 2x^2 \end{pmatrix}$$

where the y^i are the rectangular coordinates of the point. Find $\mathbf{e}^i, \mathbf{e}_i, i = 1, 2, 3$, and find $(g_{ij})(\mathbf{x})$ and $\left(g^{ij}(\mathbf{x})\right)$.

3. Let

$$\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 + 2x^2 \\ x^3 \\ 3x^1 + x^3 \end{pmatrix}$$

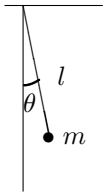
where the y^i are the rectangular coordinates of the point. Find $\mathbf{e}^i, \mathbf{e}_i, i = 1, 2, 3$, and find $(g_{ij})(\mathbf{x})$ and $\left(g^{ij}(\mathbf{x})\right)$.

4. If the above are too easy and you want lots of computations to do, let

$$\begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} x^1 + x^2 \\ x^2 + \sin\left(x^3\right) \\ x^1 + 2x^2 \end{pmatrix}$$

where the y^i are the rectangular coordinates of the point. Find $\mathbf{e}^i, \mathbf{e}_i, i = 1, 2, 3$, and find $\left(g_{ij}\right)(\mathbf{x})$ and $\left(g^{ij}\left(\mathbf{x}\right)\right)$.

5. Let $\mathbf{y} = \mathbf{y}\left(\mathbf{x}, t\right)$ where t signifies time and $\mathbf{x} \in U \subseteq \mathbb{R}^m$ for U an open set, while $\mathbf{y} \in \mathbb{R}^n$ and suppose $\mathbf{x}$ is a function of t. Physically, this corresponds to an object moving over a surface in $\mathbb{R}^n$ which may be changing as a function of t. The point $\mathbf{y} = \mathbf{y}\left(\mathbf{x}\left(t\right), t\right)$ is the point in $\mathbb{R}^n$ corresponding to t. For example, consider the pendulum



in which $n = 2, l$ is fixed and $y^1 = l\sin\theta, y^2 = l - l\cos\theta$. Thus, in this simple example, $m = 1$. If l were changing in a known way with respect to t, then this would be of the form $\mathbf{y} = \mathbf{y}\left(\mathbf{x}, t\right)$. The kinetic energy is defined as

$$T \equiv \frac{1}{2} m\dot{\mathbf{y}} \cdot \dot{\mathbf{y}} \tag{$*$}$$

where the dot on the top signifies differentiation with respect to t. Show

$$\frac{\partial T}{\partial \dot{x}^k} = m\dot{\mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial x^k}.$$

Hint: First show

$$\dot{\mathbf{y}} = \frac{\partial \mathbf{y}}{\partial x^j}\dot{x}^j + \frac{\partial \mathbf{y}}{\partial t} \tag{$**$}$$

and so

$$\frac{\partial \dot{\mathbf{y}}}{\partial \dot{x}^j} = \frac{\partial \mathbf{y}}{\partial x^j}.$$

6. ↑ Show

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{x}^k}\right) = m\ddot{\mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial x^k} + m\dot{\mathbf{y}} \cdot \frac{\partial^2 \mathbf{y}}{\partial x^k \partial x^r}\dot{x}^r + m\dot{\mathbf{y}} \cdot \frac{\partial^2 \mathbf{y}}{\partial t \partial x^k}.$$

7. ↑ Show

$$\frac{\partial T}{\partial x^k} = m\dot{\mathbf{y}} \cdot \left(\frac{\partial^2 \mathbf{y}}{\partial x^r \partial x^k}\dot{x}^r + \frac{\partial^2 \mathbf{y}}{\partial t \partial x^k}\right).$$

Hint: Use $*$ and $**$.

8. ↑ Now show from Newton's second law (mass times acceleration equals force) that for $\mathbf{F}$ the force,

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{x}^k}\right) - \frac{\partial T}{\partial x^k} = m\ddot{\mathbf{y}} \cdot \frac{\partial \mathbf{y}}{\partial x^k} = \mathbf{F} \cdot \frac{\partial \mathbf{y}}{\partial x^k}. \tag{$***$}$$

9. ↑ In the example of the simple pendulum above,

$$\mathbf{y} = \begin{pmatrix} l\sin\theta \\ l - l\cos\theta \end{pmatrix} = l\sin\theta\mathbf{i} + (l - l\cos\theta)\,\mathbf{j}.$$

Use $***$ to find a differential equation which describes the vibrations of the pendulum in terms of θ. First write the kinetic energy and then consider the force acting on the mass which is

$$-mg\mathbf{j}.$$

10. The above problem is fairly easy to do without the formalism developed. Now consider the case where $\mathbf{x} = (\rho, \theta, \phi)$, spherical coordinates, and write differential equations for ρ, θ, and ϕ to describe the motion of an object in terms of these coordinates given a force, $\mathbf{F}$.

11. Suppose the pendulum is not assumed to vibrate in a plane. Let it be suspended at the origin and consider spherical coordinates. Find differential equations for θ and ϕ.

12. If there are many masses, $m_\alpha, \alpha = 1, \cdots, R$, the kinetic energy is the sum of the kinetic energies of the individual masses. Thus,

$$T \equiv \frac{1}{2}\sum_{\alpha=1}^{R} m_\alpha \left|\dot{\mathbf{y}}_\alpha\right|^2.$$

Generalize the above problems to show that, assuming

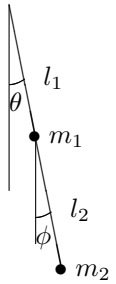
$$\mathbf{y}_\alpha = \mathbf{y}_\alpha(\mathbf{x}, t),$$

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{x}^k}\right) - \frac{\partial T}{\partial x^k} = \sum_{\alpha=1}^{R} \mathbf{F}_\alpha \cdot \frac{\partial \mathbf{y}_\alpha}{\partial x^k}$$

where $\mathbf{F}_\alpha$ is the force acting on m_α.

13. Discuss the equivalence of these formulae with Newton's second law, force equals mass times acceleration. What is gained from the above so called Lagrangian formalism?

14. The double pendulum has two masses instead of only one.

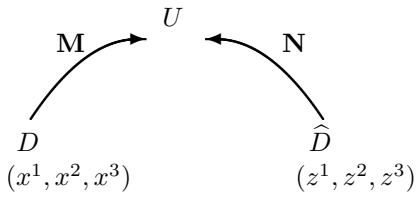


Write differential equations for θ and ϕ to describe the motion of the double pendulum.

36.2 Transformation Of Coordinates.

The scalars $\left\{x^i\right\}$ are called cuvilinear coordinates. Note they can be used to identify a point in $\mathbb{R}^n$ and $\mathbf{x} = \left(x^1, \cdots, x^n\right)$ is a point in $\mathbb{R}^n$. The basis vectors associated with this particular set of curvilinear coordinates at a point identified by $\mathbf{x}$ are denoted by $\mathbf{e}_i(\mathbf{x})$ and the dual basis vectors at this point are denoted by $\mathbf{e}^j(\mathbf{x})$. What if other curvilinear coordinates are used? How do you write $\mathbf{e}^k(\mathbf{x})$ in terms of the vectors, $\mathbf{e}^j(\mathbf{z})$ where $\mathbf{z}$ is some other type of curvilinear coordinates?

Consider the following picture in which U is an open set in $\mathbb{R}^n$, D, and $\widehat{D}$ are open sets in $\mathbb{R}^n$, and $\mathbf{M}, \mathbf{N}$ are C^2 mappings which are one to one from D and $\widehat{D}$ respectively. Suppose that a point in U is identified by the curvilinear coordinates $\mathbf{x}$ in D and $\mathbf{z}$ in $\widehat{D}$.



Thus $\mathbf{M}(\mathbf{x}) = \mathbf{N}(\mathbf{z})$. By the chain rule,

$$\mathbf{e}_i(\mathbf{z}) \equiv \frac{\partial \mathbf{N}}{\partial z^i} = \frac{\partial \mathbf{M}}{\partial x^j}\frac{\partial x^j}{\partial z^i} = \frac{\partial x^j}{\partial z^i}\mathbf{e}_j(\mathbf{x}). \tag{36.8}$$

Recall the covariant and contravariant coordinates defined in Chapter 19 which starts on Page 485. Thus,

$$\mathbf{v} = v_i(\mathbf{x})\,\mathbf{e}^i(\mathbf{x}) = v^i(\mathbf{x})\,\mathbf{e}_i(\mathbf{x}) = v_j(\mathbf{z})\,\mathbf{e}^j(\mathbf{z}) = v^j(\mathbf{z})\,\mathbf{e}_j(\mathbf{z}).$$

Then the following theorem tells how to transform various things defined above.

Theorem 36.2.1 *The following transformation rules hold for pairs of curvilinear coordinates.*

$$v_i(\mathbf{z}) = \frac{\partial x^j}{\partial z^i}v_j(\mathbf{x}),\ v^i(\mathbf{z}) = \frac{\partial z^i}{\partial x^j}v^j(\mathbf{x}), \tag{36.9}$$

$$\mathbf{e}_i(\mathbf{z}) = \frac{\partial x^j}{\partial z^i}\mathbf{e}_j(\mathbf{x}),\ \mathbf{e}^i(\mathbf{z}) = \frac{\partial z^i}{\partial x^j}\mathbf{e}^j(\mathbf{x}), \tag{36.10}$$

$$g_{ij}(\mathbf{z}) = \frac{\partial x^r}{\partial z^i}\frac{\partial x^s}{\partial z^j}g_{rs}(\mathbf{x}),\ g^{ij}(\mathbf{z}) = \frac{\partial z^i}{\partial x^r}\frac{\partial z^j}{\partial x^s}g^{rs}(\mathbf{x}). \tag{36.11}$$

Proof: The first part of (36.10) is shown in (36.8). Then, from (36.8) and Theorem 19.2.2 on Page 490

$$\begin{aligned}\mathbf{e}^i(\mathbf{z}) &= \mathbf{e}^i(\mathbf{z})\cdot\mathbf{e}_j(\mathbf{x})\,\mathbf{e}^j(\mathbf{x}) = \mathbf{e}^i(\mathbf{z})\cdot\frac{\partial z^k}{\partial x^j}\mathbf{e}_k(\mathbf{z})\,\mathbf{e}^j(\mathbf{x})\\ &= \delta^i_k\frac{\partial z^k}{\partial x^j}\mathbf{e}^j(\mathbf{x}) = \frac{\partial z^i}{\partial x^j}\mathbf{e}^j(\mathbf{x})\end{aligned}$$

and this proves the second part of (36.10). Now to show (36.9), use Theorem 19.2.2 on Page 490 again.

$$v_i(\mathbf{z}) = \mathbf{v}\cdot\mathbf{e}_i(\mathbf{z}) = \mathbf{v}\cdot\frac{\partial x^j}{\partial z_i}\mathbf{e}_j(\mathbf{x}) = \frac{\partial x^j}{\partial z_i}v_j(\mathbf{x})$$

and

$$v^i(\mathbf{z}) = \mathbf{v}\cdot\mathbf{e}^i(\mathbf{z}) = \mathbf{v}\cdot\frac{\partial z^i}{\partial x^j}\mathbf{e}^j(\mathbf{x}) = \frac{\partial z^i}{\partial x^j}v^j(\mathbf{x}).$$

To verify (36.11),

$$g_{ij}(\mathbf{z}) = \mathbf{e}_i(\mathbf{z}) \cdot \mathbf{e}_j(\mathbf{z}) = \mathbf{e}_r(\mathbf{x}) \frac{\partial x^r}{\partial z^i} \cdot \mathbf{e}_s(\mathbf{x}) \frac{\partial x^s}{\partial z^j} = g_{rs}(\mathbf{x}) \frac{\partial x^r}{\partial z^i} \frac{\partial x^s}{\partial z^j}.$$

This proves the theorem.

Denote by $\mathbf{y}$ the curvilinear coordinates with the property

$$\mathbf{e}^k(\mathbf{y}) = \mathbf{i}_k = \mathbf{e}_k(\mathbf{y}).$$

36.3 Differentiation And Christoffel Symbols

Let $\mathbf{F}: U \to \mathbb{R}^n$ be differentiable. $\mathbf{F}$ is a vector field and it is used to model force, velocity, acceleration, or any other vector quantity which may change from point to point in U. Then

$$\frac{\partial \mathbf{F}(\mathbf{x})}{\partial x^j}$$

is a vector and so there exist scalars, $F^i_{,j}(\mathbf{x})$ and $F_{i,j}(\mathbf{x})$ such that

$$\frac{\partial \mathbf{F}(\mathbf{x})}{\partial x^j} = F^i_{,j}(\mathbf{x}) \mathbf{e}_i(\mathbf{x}) = F_{i,j}(\mathbf{x}) \mathbf{e}^j(\mathbf{x}). \tag{36.12}$$

How do these scalars transform when the coordinates are changed?

Theorem 36.3.1 *If* $\mathbf{x}$ *and* $\mathbf{z}$ *are curvilinear coordinates,*

$$F^r_{,s}(\mathbf{x}) = F^i_{,j}(\mathbf{z}) \frac{\partial x^r}{\partial z^i} \frac{\partial z^j}{\partial x^s},\ F_{r,s}(\mathbf{x}) \frac{\partial x^r}{\partial z^i} \frac{\partial x^s}{\partial z^j} = F_{i,j}(\mathbf{z}). \tag{36.13}$$

Proof:

$$F^r_{,s}(\mathbf{x}) \mathbf{e}_r(\mathbf{x}) \equiv \frac{\partial \mathbf{F}(\mathbf{x})}{\partial x^s} = \frac{\partial \mathbf{F}(\mathbf{z})}{\partial z^j} \frac{\partial z^j}{\partial x^s} \equiv$$

$$F^i_{,j}(\mathbf{z}) \mathbf{e}_i(\mathbf{z}) \frac{\partial z^j}{\partial x^s} = F^i_{,j}(\mathbf{z}) \frac{\partial x^r}{\partial z^i} \frac{\partial z^j}{\partial x^s} \mathbf{e}_r(\mathbf{x})$$

which shows the first formula of (36.12). To show the other formula,

$$F_{i,j}(\mathbf{z}) \mathbf{e}^i(\mathbf{z}) \equiv \frac{\partial \mathbf{F}(\mathbf{z})}{\partial z^j} = \frac{\partial \mathbf{F}(\mathbf{x})}{\partial x^s} \frac{\partial x^s}{\partial z^j} \equiv$$

$$F_{r,s}(\mathbf{x}) \mathbf{e}^r(\mathbf{x}) \frac{\partial x^s}{\partial z^j} = F_{r,s}(\mathbf{x}) \frac{\partial x^r}{\partial z^i} \frac{\partial x^s}{\partial z^j} \mathbf{e}^i(\mathbf{z}),$$

and this shows the second formula for transforming these scalars.

Now $\mathbf{F}(\mathbf{x}) = F^i(\mathbf{x}) \mathbf{e}_i(\mathbf{x})$ and so by the product rule,

$$\frac{\partial \mathbf{F}}{\partial x^j} = \frac{\partial F^i}{\partial x^j} \mathbf{e}_i(\mathbf{x}) + F^i(\mathbf{x}) \frac{\partial \mathbf{e}_i(\mathbf{x})}{\partial x^j}.$$

Now $\frac{\partial \mathbf{e}_i(\mathbf{x})}{\partial x^j}$ is a vector and so there exist scalars, $\left\{ {k \atop ij} \right\}$ such that

$$\frac{\partial \mathbf{e}_i(\mathbf{x})}{\partial x^j} = \left\{ {k \atop ij} \right\} \mathbf{e}_k(\mathbf{x}).$$

Therefore,

$$\frac{\partial \mathbf{F}}{\partial x^j} = \frac{\partial F^k}{\partial x^j} \mathbf{e}_k(\mathbf{x}) + F^i(\mathbf{x}) \left\{ {k \atop ij} \right\} \mathbf{e}_k(\mathbf{x})$$

which shows

$$F^k_{,j}(\mathbf{x}) = \frac{\partial F^k}{\partial x^j} + F^i(\mathbf{x}) \left\{ \begin{matrix} k \\ ij \end{matrix} \right\}.$$

This is sometimes called the covariant derivative.

These scalars are called the Christoffel symbols of the second kind. The next theorem is devoted to properties of these Christoffel symbols. Before stating the theorem, recall that the mapping, $\mathbf{M}$ which defines the curvilinear coordinates is C^2. The reason for this is that it will be necessary to assert mixed partial derivatives are equal.

Theorem 36.3.2 *The Christoffel symbols of the second kind satisfy the following*

$$\frac{\partial \mathbf{e}_i(\mathbf{x})}{\partial x^j} = \left\{ \begin{matrix} k \\ ij \end{matrix} \right\} \mathbf{e}_k(\mathbf{x}), \tag{36.14}$$

$$\frac{\partial \mathbf{e}^i(\mathbf{x})}{\partial x^j} = -\left\{ \begin{matrix} i \\ kj \end{matrix} \right\} \mathbf{e}^k(\mathbf{x}), \tag{36.15}$$

$$\left\{ \begin{matrix} k \\ ij \end{matrix} \right\} = \left\{ \begin{matrix} k \\ ji \end{matrix} \right\}, \tag{36.16}$$

$$\left\{ \begin{matrix} m \\ ik \end{matrix} \right\} = \frac{g^{jm}}{2} \left[\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j} \right]. \tag{36.17}$$

Proof: Formula (36.14) is the definition of the Christoffel symbols. Consider (36.15) next. To do so, note

$$\mathbf{e}^i(\mathbf{x}) \cdot \mathbf{e}_k(\mathbf{x}) = \delta^i_k.$$

Then from the product rule,

$$\frac{\partial \mathbf{e}^i(\mathbf{x})}{\partial x^j} \cdot \mathbf{e}_k(\mathbf{x}) + \mathbf{e}^i(\mathbf{x}) \cdot \frac{\partial \mathbf{e}_k(\mathbf{x})}{\partial x^j} = 0.$$

Now from the definition,

$$\frac{\partial \mathbf{e}^i(\mathbf{x})}{\partial x^j} \cdot \mathbf{e}_k(\mathbf{x}) = -\mathbf{e}^i(\mathbf{x}) \cdot \left\{ \begin{matrix} r \\ kj \end{matrix} \right\} \mathbf{e}_r(\mathbf{x}) = -\left\{ \begin{matrix} i \\ kj \end{matrix} \right\}.$$

Therefore,

$$\frac{\partial \mathbf{e}^i(\mathbf{x})}{\partial x^j} = \left(\frac{\partial \mathbf{e}^i(\mathbf{x})}{\partial x^j} \cdot \mathbf{e}_k(\mathbf{x}) \right) \mathbf{e}^k(\mathbf{x}) = -\left\{ \begin{matrix} i \\ kj \end{matrix} \right\} \mathbf{e}^k(\mathbf{x}).$$

This verifies (36.15).

Letting $\frac{\partial \mathbf{M}(\mathbf{x})}{\partial x^j} = \mathbf{e}_j(\mathbf{x})$, it follows from equality of mixed partial derivatives,

$$\left\{ \begin{matrix} k \\ ij \end{matrix} \right\} \mathbf{e}_k(\mathbf{x}) = \frac{\partial \mathbf{e}_i}{\partial x^j} \equiv \frac{\partial^2 \mathbf{M}}{\partial x^j \partial x^i} = \frac{\partial^2 \mathbf{M}}{\partial x^i \partial x^j} = \frac{\partial \mathbf{e}_j}{\partial x^i} = \left\{ \begin{matrix} k \\ ji \end{matrix} \right\} \mathbf{e}_k(\mathbf{x}),$$

which shows (36.16).

It remains to show (36.17).

$$\frac{\partial g_{ij}}{\partial x^k} = \frac{\partial \mathbf{e}_i}{\partial x^k} \cdot \mathbf{e}_j + \mathbf{e}_i \cdot \frac{\partial \mathbf{e}_j}{\partial x^k} = \left\{ \begin{matrix} r \\ ik \end{matrix} \right\} \mathbf{e}_r \cdot \mathbf{e}_j + \mathbf{e}_i \cdot \mathbf{e}_r \left\{ \begin{matrix} r \\ jk \end{matrix} \right\}.$$

Therefore,

$$\frac{\partial g_{ij}}{\partial x^k} = \left\{ \begin{matrix} r \\ ik \end{matrix} \right\} g_{rj} + \left\{ \begin{matrix} r \\ jk \end{matrix} \right\} g_{ri}. \tag{36.18}$$

Switching i and k while remembering (36.16) yields

$$\frac{\partial g_{kj}}{\partial x^i} = \left\{ \begin{matrix} r \\ ik \end{matrix} \right\} g_{rj} + \left\{ \begin{matrix} r \\ ji \end{matrix} \right\} g_{rk}. \tag{36.19}$$

Now switching j and k in (36.18),

$$\frac{\partial g_{ik}}{\partial x^j} = \left\{ \begin{matrix} r \\ ij \end{matrix} \right\} g_{rk} + \left\{ \begin{matrix} r \\ jk \end{matrix} \right\} g_{ri}. \tag{36.20}$$

Adding (36.18) to (36.19) and subtracting (36.20) yields

$$\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j} = 2 \left\{ \begin{matrix} r \\ ik \end{matrix} \right\} g_{rj}.$$

Now multiplying both sides by g^{jm} and using the fact that $\left(g^{ij}\right)$ is the inverse matrix for $\left(g_{ij}\right)$,

$$2 \left\{ \begin{matrix} m \\ ik \end{matrix} \right\} = g^{jm} \left(\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^i} - \frac{\partial g_{ik}}{\partial x^j} \right)$$

which proves (36.17).

This is a very interesting formula because it shows the Christoffel symbols are completely determined by the metric tensor and its derivatives.

36.4 Gradients And Divergence

It is very important to express the gradient and the divergence in general coordinate systems. As before, $\mathbf{y}$ will denote the standard coordinates with respect to the usual basis vectors. Thus

$$\mathbf{N}\left(\mathbf{y}\right) \equiv y^k \mathbf{i}_k,\ \mathbf{e}_k\left(\mathbf{y}\right) = \mathbf{i}_k = \mathbf{e}^k\left(\mathbf{y}\right).$$

Let $\phi : U \to \mathbb{R}$ be a differentiable scalar function, sometimes called a "scalar field" in this subject. Write $\phi\left(\mathbf{x}\right)$ to denote the value of ϕ at the point whose coordinates are $\mathbf{x}$. In general, it is convenient to follow this practice for any field, vector or scalar. Thus $\mathbf{F}\left(\mathbf{x}\right)$ is the value of a vector field at the point of U determined by the coordinates $\mathbf{x}$. (Remember, vectors are those things which are determined by direction and magnitude.) In the standard coordinates, the gradient is known. It is given by the following formula.

$$\nabla\phi\left(\mathbf{y}\right) = \frac{\partial\phi\left(\mathbf{y}\right)}{\partial y^k}\mathbf{e}^k\left(\mathbf{y}\right).$$

Recall $\mathbf{e}^k\left(\mathbf{y}\right) = \mathbf{i}_k$, the vector whose Cartesian coordinates are all zero except for a 1 in the k^{th} position. Therefore, using the chain rule, if the coordinates of the point of U are given as $\mathbf{x}$,

$$\nabla\phi\left(\mathbf{x}\right) = \nabla\phi\left(\mathbf{y}\right)$$

$$= \frac{\partial\phi\left(\mathbf{x}\right)}{\partial x^r}\frac{\partial x^r}{\partial y^k}\frac{\partial y^k}{\partial x^s}\mathbf{e}^s\left(\mathbf{x}\right) = \frac{\partial\phi\left(\mathbf{x}\right)}{\partial x^r}\delta^r_s\mathbf{e}^s\left(\mathbf{x}\right) = \frac{\partial\phi\left(\mathbf{x}\right)}{\partial x^r}\mathbf{e}^r\left(\mathbf{x}\right).$$

This shows the covariant components of $\nabla\phi\left(\mathbf{x}\right)$ are

$$\left(\nabla\phi\left(\mathbf{x}\right)\right)_r = \frac{\partial\phi\left(\mathbf{x}\right)}{\partial x^r}. \tag{36.21}$$

To find the contravariant components, raise the index in the usual way. Thus

$$(\nabla\phi(\mathbf{x}))^r = g^{rk}(\mathbf{x})(\nabla\phi(\mathbf{x}))_k = g^{rk}(\mathbf{x})\frac{\partial\phi(\mathbf{x})}{\partial x^k}. \tag{36.22}$$

What about the divergence of a vector field? The divergence of a vector field, $\mathbf{F}$ defined on U is a scalar field, $\operatorname{div}(\mathbf{F})$ which from calculus is

$$\frac{\partial F^k}{\partial y^k}(\mathbf{y}) = F^k_{,k}(\mathbf{y})$$

in terms of the usual coordinates $\mathbf{y}$. The reason the above equation holds in this case is that $\mathbf{e}_k(\mathbf{y})$ is a constant and so the Christoffel symbols are zero. What is the expression for the divergence in an arbitrary coordinate system? From Theorem 36.3.1,

$$F^i_{,j}(\mathbf{y}) = F^r_{,s}(\mathbf{x})\frac{\partial x^s}{\partial y^j}\frac{\partial y^i}{\partial x^r}$$

$$= \left(\frac{\partial F^r(\mathbf{x})}{\partial x^s} + F^k(\mathbf{x})\begin{Bmatrix} r \\ ks \end{Bmatrix}(\mathbf{x})\right)\frac{\partial x^s}{\partial y^j}\frac{\partial y^i}{\partial x^r}.$$

Letting $j = i$ yields

$$\begin{aligned}
\operatorname{div}(\mathbf{F}) &= \left(\frac{\partial F^r(\mathbf{x})}{\partial x^s} + F^k(\mathbf{x})\begin{Bmatrix} r \\ ks \end{Bmatrix}(\mathbf{x})\right)\frac{\partial x^s}{\partial y^i}\frac{\partial y^i}{\partial x^r} \\
&= \left(\frac{\partial F^r(\mathbf{x})}{\partial x^s} + F^k(\mathbf{x})\begin{Bmatrix} r \\ ks \end{Bmatrix}(\mathbf{x})\right)\delta^s_r \\
&= \left(\frac{\partial F^r(\mathbf{x})}{\partial x^r} + F^k(\mathbf{x})\begin{Bmatrix} r \\ kr \end{Bmatrix}(\mathbf{x})\right).
\end{aligned} \tag{36.23}$$

The symbol, $\left\{ \begin{smallmatrix} r \\ kr \end{smallmatrix} \right\}$ is now simplified[1] using the description of it in Theorem 36.3.2. Thus, from this theorem,

$$\begin{Bmatrix} r \\ rk \end{Bmatrix} = \frac{g^{jr}}{2}\left[\frac{\partial g_{rj}}{\partial x^k} + \frac{\partial g_{kj}}{\partial x^r} - \frac{\partial g_{rk}}{\partial x^j}\right]$$

Now consider $\frac{g^{jr}}{2}$ times the last two terms in $[\cdot]$. Relabeling the indices r and j in the second term implies

$$\frac{g^{jr}}{2}\frac{\partial g_{kj}}{\partial x^r} - \frac{g^{jr}}{2}\frac{\partial g_{rk}}{\partial x^j} = \frac{g^{jr}}{2}\frac{\partial g_{kj}}{\partial x^r} - \frac{g^{rj}}{2}\frac{\partial g_{jk}}{\partial x^r} = 0.$$

Therefore,

$$\begin{Bmatrix} r \\ rk \end{Bmatrix} = \frac{g^{jr}}{2}\frac{\partial g_{rj}}{\partial x^k}. \tag{36.24}$$

Now recall $g \equiv \det(g_{ij}) = \det(G) > 0$ from Theorem 19.2.7 on Page 492. Also from the formula for the inverse of a matrix in which the inverse equals one divided by the determinant times the transpose of the cofactor matrix, and this theorem,

$$g^{jr} = A^{rj}(\det G)^{-1} = A^{jr}(\det G)^{-1}$$

where A^{rj} is the rj^{th} cofactor of the matrix (g_{ij}). Also recall that

$$g = \sum_{r=1}^{n} g_{rj}A^{rj} \text{ no sum on } j.$$

[1] This is called a contraction because there is a repeated index in the symbol.

Therefore, g is a function of the variables $\{g_{rj}\}$ and

$$\frac{\partial g}{\partial g_{rj}} = A^{rj}.$$

From (36.24),

$$\left\{ \begin{matrix} r \\ rk \end{matrix} \right\} = \frac{g^{jr}}{2}\frac{\partial g_{rj}}{\partial x^k} = \frac{1}{2g}\frac{\partial g_{rj}}{\partial x^k}A^{jr} = \frac{1}{2g}\frac{\partial g}{\partial g_{rj}}\frac{\partial g_{rj}}{\partial x^k} = \frac{1}{2g}\frac{\partial g}{\partial x^k}$$

and so from (36.23),

$$\operatorname{div}(\mathbf{F}) = \frac{\partial F^k(\mathbf{x})}{\partial x^k} +$$

$$+F^k(\mathbf{x})\frac{1}{2g(\mathbf{x})}\frac{\partial g(\mathbf{x})}{\partial x^k} = \frac{1}{\sqrt{g(\mathbf{x})}}\frac{\partial}{\partial x^i}\left(F^i(\mathbf{x})\sqrt{g(\mathbf{x})}\right). \tag{36.25}$$

This is the formula for the divergence of a vector field in general curvilinear coordinates.

The Laplacian of a scalar field is nothing more than the divergence of the gradient. In symbols,

$$\Delta\phi \equiv \nabla\cdot\nabla\phi$$

From (36.25) and (36.22) it follows

$$\Delta\phi(\mathbf{x}) = \frac{1}{\sqrt{g(\mathbf{x})}}\frac{\partial}{\partial x^i}\left(g^{ik}(\mathbf{x})\frac{\partial\phi(\mathbf{x})}{\partial x^k}\sqrt{g(\mathbf{x})}\right). \tag{36.26}$$

To summarize the conclusions of this section, here is a major theorem.

Theorem 36.4.1 *The following formulas hold for the gradient, divergence and Laplacian in general curvilinear coordinates.*

$$(\nabla\phi(\mathbf{x}))_r = \frac{\partial\phi(\mathbf{x})}{\partial x^r}, \tag{36.27}$$

$$(\nabla\phi(\mathbf{x}))^r = g^{rk}(\mathbf{x})\frac{\partial\phi(\mathbf{x})}{\partial x^k}, \tag{36.28}$$

$$\operatorname{div}(\mathbf{F}) = \frac{1}{\sqrt{g(\mathbf{x})}}\frac{\partial}{\partial x^i}\left(F^i(\mathbf{x})\sqrt{g(\mathbf{x})}\right), \tag{36.29}$$

$$\Delta\phi(\mathbf{x}) = \frac{1}{\sqrt{g(\mathbf{x})}}\frac{\partial}{\partial x^i}\left(g^{ik}(\mathbf{x})\frac{\partial\phi(\mathbf{x})}{\partial x^k}\sqrt{g(\mathbf{x})}\right). \tag{36.30}$$

36.5 Exercises

1. Let $y^1 = x^1 + 2x^2, y^2 = x^2 + 3x^3, y^3 = x^1 + x^3$. Let

   $$\mathbf{F}(\mathbf{x}) = x^1\mathbf{e}_1(\mathbf{x}) + x^2\mathbf{e}_2(\mathbf{x}) + \left(x^3\right)^2\mathbf{e}(\mathbf{x}).$$

   Find $\operatorname{div}(\mathbf{F})(\mathbf{x})$.

2. For the coordinates of the preceding problem, and ϕ a scalar field, find

   $$(\nabla\phi(\mathbf{x}))^3$$

   in terms of the partial derivatives of ϕ taken with respect to the variables x^i.

3. Let $y^1 = 7x^1 + 2x^2, y^2 = x^2 + 3x^3, y^3 = x^1 + x^3$. Let ϕ be a scalar field. Find $\nabla^2 \phi(\mathbf{x})$.

4. Derive $\nabla^2 u$ in cylindrical coordinates, r, θ, z, where u is a scalar field on $\mathbb{R}^3$.

$$x = r\cos\theta,\ y = r\sin\theta,\ z = z.$$

5. ↑ Find all solutions to $\nabla^2 u = 0$ which depend only on r where $r \equiv \sqrt{x^2 + y^2}$.

6. Let u be a scalar field on $\mathbb{R}^3$. Find all solutions to $\nabla^2 u = 0$ which depend only on

$$\rho \equiv \sqrt{x^2 + y^2 + z^2}.$$

7. Find all harmonic functions defined on $\mathbb{R}^n$ which depend only in ρ, the distance to the origin. **Hint:** Use the formula for the Laplacian in an appropriate coordinate system.

8. The temperature, u, in a solid satisfies $\nabla^2 u = 0$ after a long time. Suppose in a long pipe of inner radius 9 and outer radius 10 the exterior surface is held at 100° while the inner surface is held at 200° find the temperature in the solid part of the pipe.

9. Show

$$\left\{ \begin{matrix} l \\ ij \end{matrix} \right\} = \frac{\partial \mathbf{e}_i}{\partial x^j} \cdot \mathbf{e}^l.$$

Find the Christoffel symbols of the second kind for spherical coordinates in which $x^1 = \phi$, $x^2 = \theta$, and $x^3 = \rho$. Do the same for cylindrical coordinates letting $x^1 = r$, $x^2 = \theta$, $x^3 = z$.

10. Show velocity can be expressed as $\mathbf{v} = v_i(\mathbf{x})\,\mathbf{e}^i(\mathbf{x})$, where

$$v_i(\mathbf{x}) = \frac{\partial r_i}{\partial x^j}\frac{dx^j}{dt} - r_p(\mathbf{x})\left\{ \begin{matrix} p \\ ik \end{matrix} \right\}\frac{dx^k}{dt}$$

and $r_i(\mathbf{x})$ are the covariant components of the displacement vector,

$$\mathbf{r} = r_i(\mathbf{x})\,\mathbf{e}^i(\mathbf{x}).$$

11. ↑ Using problem 9 and 10, show the covariant components of velocity in spherical coordinates are

$$v_1 = \rho^2\frac{d\phi}{dt},\ v_2 = \rho^2\sin^2(\phi)\frac{d\theta}{dt},\ v_3 = \frac{d\rho}{dt}.$$

Hint: First observe that if $\mathbf{r}$ is the position vector from the origin, then $\mathbf{r} = \rho\mathbf{e}_3$ so $r_1 = 0 = r_2$, and $r_3 = \rho$. Now use 10.

36.6 Curl And Cross Products

I have not had occasion to use this material very much but for the sake of completeness, here it is. It involves the curl and cross product in terms of general curvilinear coordinates.It will always be assumed that for $\mathbf{x}$ a set of curvilinear coordinates,

$$\det\left(\frac{\partial y^i}{\partial x^j}\right) > 0 \tag{36.31}$$

Where the $\mathbf{y}_i$ are the usual coordinates in which $\mathbf{e}_k(\mathbf{y}) = \mathbf{i}_k$. This sort of fussy thing is necessary because of the antisymmetry of the cross product.

Theorem 36.6.1 *Let (36.31) hold. Then*

$$\det\left(\frac{\partial y^i}{\partial x^j}\right) = \sqrt{g(\mathbf{x})} \tag{36.32}$$

and

$$\det\left(\frac{\partial x^i}{\partial y^j}\right) = \frac{1}{\sqrt{g(\mathbf{x})}}. \tag{36.33}$$

Proof:

$$\mathbf{e}_i(\mathbf{x}) = \frac{\partial y^k}{\partial x^i}\mathbf{i}_k$$

and so

$$g_{ij}(\mathbf{x}) = \frac{\partial y^k}{\partial x^i}\mathbf{i}_k \cdot \frac{\partial y^l}{\partial x^j}\mathbf{i}_l = \frac{\partial y^k}{\partial x^i}\frac{\partial y^k}{\partial x^j}.$$

Therefore, $g = \det(g_{ij}(\mathbf{x})) = \left(\det\left(\frac{\partial y^k}{\partial x^i}\right)\right)^2$. By (36.31),$\sqrt{g} = \det\left(\frac{\partial y^k}{\partial x^i}\right)$ as claimed. Now

$$\frac{\partial y^k}{\partial x^i}\frac{\partial x^i}{\partial y^r} = \delta^k_r$$

and so

$$\det\left(\frac{\partial x^i}{\partial y^r}\right) = \frac{1}{\sqrt{g(\mathbf{x})}}.$$

This proves the theorem.

To get the curl and cross product in curvilinear coordinates, let ϵ^{ijk} be the usual permutation symbol. Thus,

$$\epsilon^{123} = 1$$

and when any two indices in ϵ^{ijk} are switched, the sign changes. Thus

$$\epsilon^{132} = -1, \epsilon^{312} = 1, \text{ etc.}$$

Now define

$$\varepsilon^{ijk}(\mathbf{x}) \equiv \epsilon^{ijk}\frac{1}{\sqrt{g(\mathbf{x})}}.$$

Then for $\mathbf{x}$ and $\mathbf{z}$ satisfying (36.31),

$$\varepsilon^{ijk}(\mathbf{x})\frac{\partial z^r}{\partial x^i}\frac{\partial z^s}{\partial x^j}\frac{\partial z^t}{\partial x^k} = \epsilon^{ijk}\det\left(\frac{\partial x^p}{\partial y^q}\right)\frac{\partial z^r}{\partial x^i}\frac{\partial z^s}{\partial x^j}\frac{\partial z^t}{\partial x^k}$$

$$= \epsilon^{rst}\det\left(\frac{\partial x^p}{\partial y^q}\right)\det\left(\frac{\partial z^i}{\partial x^k}\right) = \epsilon^{rst}\det(MN)$$

where N is the matrix whose pq^{th} entry is $\frac{\partial x^p}{\partial y^q}$ and M is the matrix whose ik^{th} entry is $\frac{\partial z^i}{\partial x^k}$. Therefore, from the definition of matrix multiplication and the chain rule, this equals

$$= \epsilon^{rst}\det\left(\frac{\partial z^i}{\partial y^p}\right) \equiv \varepsilon^{rst}(\mathbf{z})$$

from the above discussion.

Now $\varepsilon^{ijk}(\mathbf{y}) = \epsilon^{ijk}$ and for a vector field, $\mathbf{F}$,

$$\operatorname{curl}(\mathbf{F}) \equiv \varepsilon^{ijk}(\mathbf{y})F_{k,j}(\mathbf{y})\mathbf{e}_i(\mathbf{y}).$$

Therefore, since we know how everything transforms assuming (36.31), it is routine to write this in terms of $\mathbf{x}$.

$$\begin{aligned}\operatorname{curl}(\mathbf{F}) &= \varepsilon^{rst}(\mathbf{x})\frac{\partial y^i}{\partial x^r}\frac{\partial y^j}{\partial x^s}\frac{\partial y^k}{\partial x^t}F_{p,q}(\mathbf{x})\frac{\partial x^p}{\partial y^k}\frac{\partial x^q}{\partial y^j}\mathbf{e}_m(\mathbf{x})\frac{\partial x^m}{\partial y^i}\\ &= \varepsilon^{rst}(\mathbf{x})\delta_r^m\delta_s^q\delta_t^p F_{p,q}(\mathbf{x})\mathbf{e}_m(\mathbf{x}) = \varepsilon^{mqp}(\mathbf{x})F_{p,q}(\mathbf{x})\mathbf{e}_m(\mathbf{x}). \end{aligned} \tag{36.34}$$

More simplification is possible. Recalling the definition of $F_{p,q}(\mathbf{x})$,

$$\frac{\partial \mathbf{F}}{\partial x^q} \equiv F_{p,q}(\mathbf{x})\mathbf{e}^p(\mathbf{x}) = \frac{\partial}{\partial x^q}\left[F_p(\mathbf{x})\mathbf{e}^p(\mathbf{x})\right]$$

$$= \frac{\partial F_p(\mathbf{x})}{\partial x^q}\mathbf{e}^p(\mathbf{x}) + F_p(\mathbf{x})\frac{\partial \mathbf{e}^p}{\partial x^q} = \frac{\partial F_p(\mathbf{x})}{\partial x^q}\mathbf{e}^p(\mathbf{x}) - F_r(\mathbf{x})\begin{Bmatrix} r \\ pq \end{Bmatrix}\mathbf{e}^p(\mathbf{x})$$

by Theorem 36.3.2. Therefore,

$$F_{p,q}(\mathbf{x}) = \frac{\partial F_p(\mathbf{x})}{\partial x^q} - F_r(\mathbf{x})\begin{Bmatrix} r \\ pq \end{Bmatrix}$$

and so

$$\operatorname{curl}(\mathbf{F}) = \varepsilon^{mqp}(\mathbf{x})\frac{\partial F_p(\mathbf{x})}{\partial x^q}\mathbf{e}_m(\mathbf{x}) - \varepsilon^{mqp}(\mathbf{x})F_r(\mathbf{x})\begin{Bmatrix} r \\ pq \end{Bmatrix}\mathbf{e}_m(\mathbf{x}).$$

However, because $\left\{ {r \atop pq} \right\} = \left\{ {r \atop qp} \right\}$, the second term in this expression equals 0. To see this,

$$\varepsilon^{mqp}(\mathbf{x})\begin{Bmatrix} r \\ pq \end{Bmatrix} = \varepsilon^{mpq}(\mathbf{x})\begin{Bmatrix} r \\ qp \end{Bmatrix} = -\varepsilon^{mqp}(\mathbf{x})\begin{Bmatrix} r \\ pq \end{Bmatrix}.$$

Therefore, by (36.34),

$$\operatorname{curl}(\mathbf{F}) = \varepsilon^{mqp}(\mathbf{x})\frac{\partial F_p(\mathbf{x})}{\partial x^q}\mathbf{e}_m(\mathbf{x}). \tag{36.35}$$

What about the cross product of two vector fields? Let $\mathbf{F}$ and $\mathbf{G}$ be two vector fields. Then in terms of standard coordinates, $\mathbf{y}$,

$$\begin{aligned}\mathbf{F}\times\mathbf{G} &= \varepsilon^{ijk}(\mathbf{y})F_j(\mathbf{y})G_k(\mathbf{y})\mathbf{e}_i(\mathbf{y})\\ &= \varepsilon^{rst}(\mathbf{x})\frac{\partial y^i}{\partial x^r}\frac{\partial y^j}{\partial x^s}\frac{\partial y^k}{\partial x^t}F_p(\mathbf{x})\frac{\partial x^p}{\partial y^j}G_q(\mathbf{x})\frac{\partial x^q}{\partial y^k}\mathbf{e}_l(\mathbf{x})\frac{\partial x^l}{\partial y^i}\\ &= \varepsilon^{rst}(\mathbf{x})\delta_s^p\delta_t^q\delta_r^l F_p(\mathbf{x})G_q(\mathbf{x})\mathbf{e}_l(\mathbf{x}) = \varepsilon^{lpq}(\mathbf{x})F_p(\mathbf{x})G_q(\mathbf{x})\mathbf{e}_l(\mathbf{x}). \end{aligned} \tag{36.36}$$

The above is summarized in the following theorem.

Theorem 36.6.2 *Suppose* $\mathbf{x}$ *is a system of curvilinear coordinates in* $\mathbb{R}^3$ *such that*

$$\det\left(\frac{\partial y^i}{\partial x^j}\right) > 0.$$

Let

$$\varepsilon^{ijk}(\mathbf{x}) \equiv \epsilon^{ijk}\frac{1}{\sqrt{g(\mathbf{x})}}.$$

Then the following formulas for curl and cross product hold in this system of coordinates.

$$\operatorname{curl}(\mathbf{F}) = \varepsilon^{mqp}(\mathbf{x})\frac{\partial F_p(\mathbf{x})}{\partial x^q}\mathbf{e}_m(\mathbf{x}),$$

and

$$\mathbf{F}\times\mathbf{G} = \varepsilon^{lpq}(\mathbf{x})F_p(\mathbf{x})G_q(\mathbf{x})\mathbf{e}_l(\mathbf{x}).$$

The Theory Of The Riemann Integral*

The definition of the Riemann integral of a function of n variables uses the following definition.

Definition 37.0.3 *For $i = 1, \cdots, n$, let $\left\{\alpha_k^i\right\}_{k=-\infty}^{\infty}$ be points on $\mathbb{R}$ which satisfy*

$$\lim_{k\to\infty} \alpha_k^i = \infty,\ \lim_{k\to-\infty} \alpha_k^i = -\infty,\ \alpha_k^i < \alpha_{k+1}^i. \tag{37.1}$$

For such sequences, define a grid on $\mathbb{R}^n$ denoted by $\mathcal{G}$ or $\mathcal{F}$ as the collection of boxes of the form

$$Q = \prod_{i=1}^{n} \left[\alpha_{j_i}^i, \alpha_{j_i+1}^i\right]. \tag{37.2}$$

If $\mathcal{G}$ is a grid, $\mathcal{F}$ is called a refinement of $\mathcal{G}$ if every box of $\mathcal{G}$ is the union of boxes of $\mathcal{F}$.

Lemma 37.0.4 *If $\mathcal{G}$ and $\mathcal{F}$ are two grids, they have a common refinement, denoted here by $\mathcal{G} \vee \mathcal{F}$.*

Proof: Let $\left\{\alpha_k^i\right\}_{k=-\infty}^{\infty}$ be the sequences used to construct $\mathcal{G}$ and let $\left\{\beta_k^i\right\}_{k=-\infty}^{\infty}$ be the sequence used to construct $\mathcal{F}$. Now let $\left\{\gamma_k^i\right\}_{k=-\infty}^{\infty}$ denote the union of $\left\{\alpha_k^i\right\}_{k=-\infty}^{\infty}$ and $\left\{\beta_k^i\right\}_{k=-\infty}^{\infty}$. It is necessary to show that for each i these points can be arranged in order. To do so, let $\gamma_0^i \equiv \alpha_0^i$. Now if

$$\gamma_{-j}^i, \cdots, \gamma_0^i, \cdots, \gamma_j^i$$

have been chosen such that they are in order and all distinct, let γ_{j+1}^i be the first element of

$$\left\{\alpha_k^i\right\}_{k=-\infty}^{\infty} \cup \left\{\beta_k^i\right\}_{k=-\infty}^{\infty} \tag{37.3}$$

which is larger than γ_j^i and let $\gamma_{-(j+1)}^i$ be the last element of (37.3) which is strictly smaller than γ_{-j}^i. The assumption (31.8) insures such a first and last element exists. Now let the grid $\mathcal{G} \vee \mathcal{F}$ consist of boxes of the form

$$Q \equiv \prod_{i=1}^{n} \left[\gamma_{j_i}^i, \gamma_{j_i+1}^i\right].$$

The Riemann integral is only defined for functions, f which are bounded and are equal to zero out of some bounded set, D. In what follows f will always be such a function.

Definition 37.0.5 *Let f be a bounded function which equals zero off a bounded set, D, and let $\mathcal{G}$ be a grid. For $Q \in \mathcal{G}$, define*

$$M_Q(f) \equiv \sup\{f(\mathbf{x}) : \mathbf{x} \in Q\},\ m_Q(f) \equiv \inf\{f(\mathbf{x}) : \mathbf{x} \in Q\}. \tag{37.4}$$

Also define for Q a box, the volume of Q, denoted by $v(Q)$ by

$$v(Q) \equiv \prod_{i=1}^{n} (b_i - a_i),\ Q \equiv \prod_{i=1}^{n} [a_i, b_i].$$

Now define upper sums, $\mathcal{U}_{\mathcal{G}}(f)$ and lower sums, $\mathcal{L}_{\mathcal{G}}(f)$ with respect to the indicated grid, by the formulas

$$\mathcal{U}_{\mathcal{G}}(f) \equiv \sum_{Q \in \mathcal{G}} M_Q(f)\, v(Q),\ \mathcal{L}_{\mathcal{G}}(f) \equiv \sum_{Q \in \mathcal{G}} m_Q(f)\, v(Q).$$

A function of n variables is Riemann integrable when there is a unique number between all the upper and lower sums. This number is the value of the integral.

Note that in this definition, $M_Q(f) = m_Q(f) = 0$ for all but finitely many $Q \in \mathcal{G}$ so there are no convergence questions to be considered here.

Lemma 37.0.6 *If $\mathcal{F}$ is a refinement of $\mathcal{G}$ then*

$$\mathcal{U}_{\mathcal{G}}(f) \geq \mathcal{U}_{\mathcal{F}}(f),\ \mathcal{L}_{\mathcal{G}}(f) \leq \mathcal{L}_{\mathcal{F}}(f).$$

Also if $\mathcal{F}$ and $\mathcal{G}$ are two grids,

$$\mathcal{L}_{\mathcal{G}}(f) \leq \mathcal{U}_{\mathcal{F}}(f).$$

Proof: For $P \in \mathcal{G}$ let $\widehat{P}$ denote the set,

$$\{Q \in \mathcal{F} : Q \subseteq P\}.$$

Then $P = \cup\widehat{P}$ and

$$\mathcal{L}_{\mathcal{F}}(f) \equiv \sum_{Q \in \mathcal{F}} m_Q(f)\, v(Q) = \sum_{P \in \mathcal{G}} \sum_{Q \in \widehat{P}} m_Q(f)\, v(Q)$$

$$\geq \sum_{P \in \mathcal{G}} m_P(f) \sum_{Q \in \widehat{P}} v(Q) = \sum_{P \in \mathcal{G}} m_P(f)\, v(P) \equiv \mathcal{L}_{\mathcal{G}}(f).$$

Similarly, the other inequality for the upper sums is valid.

To verify the last assertion of the lemma, use Lemma 37.0.4 to write

$$\mathcal{L}_{\mathcal{G}}(f) \leq \mathcal{L}_{\mathcal{G}\vee\mathcal{F}}(f) \leq \mathcal{U}_{\mathcal{G}\vee\mathcal{F}}(f) \leq \mathcal{U}_{\mathcal{F}}(f).$$

This proves the lemma.

This lemma makes it possible to define the Riemann integral.

Definition 37.0.7 *Define an upper and a lower integral as follows.*

$$\overline{I}(f) \equiv \inf\{\mathcal{U}_{\mathcal{G}}(f) : \mathcal{G} \text{ is a grid}\},$$

$$\underline{I}(f) \equiv \sup\{\mathcal{L}_{\mathcal{G}}(f) : \mathcal{G} \text{ is a grid}\}.$$

Lemma 37.0.8 $\overline{I}(f) \geq \underline{I}(f)$.

Proof: From Lemma 37.0.6 it follows for any two grids $\mathcal{G}$ and $\mathcal{F}$,

$$\mathcal{L}_{\mathcal{G}}(f) \leq \mathcal{U}_{\mathcal{F}}(f).$$

Therefore, taking the supremum for all grids on the left in this inequality,

$$\underline{I}(f) \leq \mathcal{U}_{\mathcal{F}}(f)$$

for all grids $\mathcal{F}$. Taking the infimum in this inequality, yields the conclusion of the lemma.

Definition 37.0.9 *A bounded function, f which equals zero off a bounded set, D, is said to be Riemann integrable, written as $f \in \mathcal{R}(\mathbb{R}^n)$ exactly when $\underline{I}(f) = \overline{I}(f)$. In this case define*

$$\int f\,dV \equiv \int f\,dx = \overline{I}(f) = \underline{I}(f).$$

As in the case of integration of functions of one variable, one obtains the Riemann criterion which is stated as the following theorem.

Theorem 37.0.10 *(Riemann criterion) $f \in \mathcal{R}(\mathbb{R}^n)$ if and only if for all $\varepsilon > 0$ there exists a grid $\mathcal{G}$ such that*

$$\mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{G}}(f) < \varepsilon.$$

Proof: If $f \in \mathcal{R}(\mathbb{R}^n)$, then $\overline{I}(f) = \underline{I}(f)$ and so there exist grids $\mathcal{G}$ and $\mathcal{F}$ such that

$$\mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{F}}(f) \leq \overline{I}(f) + \frac{\varepsilon}{2} - \left(\underline{I}(f) - \frac{\varepsilon}{2}\right) = \varepsilon.$$

Then letting $\mathcal{H} = \mathcal{G} \vee \mathcal{F}$, Lemma 37.0.6 implies

$$\mathcal{U}_{\mathcal{H}}(f) - \mathcal{L}_{\mathcal{H}}(f) \leq \mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{F}}(f) < \varepsilon.$$

Conversely, if for all $\varepsilon > 0$ there exists $\mathcal{G}$ such that

$$\mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{G}}(f) < \varepsilon,$$

then

$$\overline{I}(f) - \underline{I}(f) \leq \mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{G}}(f) < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this proves the theorem.

37.1 Basic Properties

It is important to know that certain combinations of Riemann integrable functions are Riemann integrable. The following theorem will include all the important cases.

Theorem 37.1.1 *Let $f, g \in \mathcal{R}(\mathbb{R}^n)$ and let $\phi : K \to \mathbb{R}$ be continuous where K is a compact set in $\mathbb{R}^2$ containing $f(\mathbb{R}^n) \times g(\mathbb{R}^n)$. Also suppose that $\phi(0,0) = 0$. Then defining*

$$h(\mathbf{x}) \equiv \phi(f(\mathbf{x}), g(\mathbf{x})),$$

it follows that h is also in $\mathcal{R}(\mathbb{R}^n)$.

Proof: Let $\varepsilon > 0$ and let $\delta_1 > 0$ be such that if $(y_i, z_i), i = 1, 2$ are points in K, such that $|z_1 - z_2| \leq \delta_1$ and $|y_1 - y_2| \leq \delta_1$, then

$$|\phi(y_1, z_1) - \phi(y_2, z_2)| < \varepsilon.$$

Let $0 < \delta < \min(\delta_1, \varepsilon, 1)$. Let $\mathcal{G}$ be a grid with the property that for $Q \in \mathcal{G}$, the diameter of Q is less than δ and also for $k = f, g$,

$$\mathcal{U}_{\mathcal{G}}(k) - \mathcal{L}_{\mathcal{G}}(k) < \delta^2. \tag{37.5}$$

Then defining for $k = f, g$,

$$\mathcal{P}_k \equiv \{Q \in \mathcal{G} : M_Q(k) - m_Q(k) > \delta\},$$

it follows

$$\delta^2 > \sum_{Q \in \mathcal{G}} (M_Q(k) - m_Q(k)) v(Q) \geq$$

$$\sum_{\mathcal{P}_k} (M_Q(k) - m_Q(k)) v(Q) \geq \delta \sum_{\mathcal{P}_k} v(Q)$$

and so for $k = f, g$,

$$\varepsilon > \delta > \sum_{\mathcal{P}_k} v(Q). \tag{37.6}$$

Suppose for $k = f, g$,

$$M_Q(k) - m_Q(k) \leq \delta.$$

Then if $\mathbf{x}_1, \mathbf{x}_2 \in Q$,

$$|f(\mathbf{x}_1) - f(\mathbf{x}_2)| < \delta, \text{ and } |g(\mathbf{x}_1) - g(\mathbf{x}_2)| < \delta.$$

Therefore,

$$|h(\mathbf{x}_1) - h(\mathbf{x}_2)| \equiv |\phi(f(\mathbf{x}_1), g(\mathbf{x}_1)) - \phi(f(\mathbf{x}_2), g(\mathbf{x}_2))| < \varepsilon$$

and it follows that

$$|M_Q(h) - m_Q(h)| \leq \varepsilon.$$

Now let

$$\mathcal{S} \equiv \{Q \in \mathcal{G} : 0 < M_Q(k) - m_Q(k) \leq \delta,\ k = f, g\}.$$

Thus the union of the boxes in $\mathcal{S}$ is contained in some large box, R, which depends only on f and g and also, from the assumption that $\phi(0,0) = 0$, $M_Q(h) - m_Q(h) = 0$ unless $Q \subseteq R$. Then

$$\mathcal{U}_{\mathcal{G}}(h) - \mathcal{L}_{\mathcal{G}}(h) \leq \sum_{Q \in \mathcal{P}_f} (M_Q(h) - m_Q(h)) v(Q) +$$

$$\sum_{Q \in \mathcal{P}_g} (M_Q(h) - m_Q(h)) v(Q) + \sum_{Q \in \mathcal{S}} \delta v(Q).$$

Now since K is compact, it follows $\phi(K)$ is bounded and so there exists a constant, C, depending only on h and ϕ such that $M_Q(h) - m_Q(h) < C$. Therefore, the above inequality implies

$$\mathcal{U}_{\mathcal{G}}(h) - \mathcal{L}_{\mathcal{G}}(h) \leq C \sum_{Q \in \mathcal{P}_f} v(Q) + C \sum_{Q \in \mathcal{P}_g} v(Q) + \sum_{Q \in \mathcal{S}} \delta v(Q),$$

which by (37.6) implies

$$\mathcal{U}_{\mathcal{G}}(h) - \mathcal{L}_{\mathcal{G}}(h) \leq 2C\varepsilon + \delta v(R) \leq 2C\varepsilon + \varepsilon v(R).$$

Since ε is arbitrary, the Riemann criterion is satisfied and so $h \in \mathcal{R}(\mathbb{R}^n)$.

Corollary 37.1.2 *Let $f, g \in \mathcal{R}(\mathbb{R}^n)$ and let $a, b \in \mathbb{R}$. Then $af + bg$, fg, and $|f|$ are all in $\mathcal{R}(\mathbb{R}^n)$. Also,*

$$\int_{\mathbb{R}^n} (af + bg)\, dx = a \int_{\mathbb{R}^n} f\, dx + b \int_{\mathbb{R}^n} g\, dx, \tag{37.7}$$

and

$$\int |f|\, dx \geq \left| \int f\, dx \right|. \tag{37.8}$$

Proof: Each of the combinations of functions described above is Riemann integrable by Theorem 37.1.1. For example, to see $af + bg \in \mathcal{R}(\mathbb{R}^n)$ consider $\phi(y, z) \equiv ay + bz$. This is clearly a continuous function of (y, z) such that $\phi(0, 0) = 0$. To obtain $|f| \in \mathcal{R}(\mathbb{R}^n)$, let $\phi(y, z) \equiv |y|$. It remains to verify the formulas. To do so, let $\mathcal{G}$ be a grid with the property that for $k = f, g$, $|f|$ and $af + bg$,

$$\mathcal{U}_{\mathcal{G}}(k) - \mathcal{L}_{\mathcal{G}}(k) < \varepsilon. \tag{37.9}$$

Consider (37.7). For each $Q \in \mathcal{G}$ pick a point in Q, $\mathbf{x}_Q$. Then

$$\sum_{Q \in \mathcal{G}} k(\mathbf{x}_Q)\, v(Q) \in [\mathcal{L}_{\mathcal{G}}(k), \mathcal{U}_{\mathcal{G}}(k)]$$

and so

$$\left| \int k\, dx - \sum_{Q \in \mathcal{G}} k(\mathbf{x}_Q)\, v(Q) \right| < \varepsilon.$$

Consequently, since

$$\sum_{Q \in \mathcal{G}} (af + bg)(\mathbf{x}_Q)\, v(Q)$$

$$= a \sum_{Q \in \mathcal{G}} f(\mathbf{x}_Q)\, v(Q) + b \sum_{Q \in \mathcal{G}} g(\mathbf{x}_Q)\, v(Q),$$

it follows

$$\left| \int (af + bg)\, dx - a \int f\, dx - b \int g\, dx \right| \leq$$

$$\left| \int (af + bg)\, dx - \sum_{Q \in \mathcal{G}} (af + bg)(\mathbf{x}_Q)\, v(Q) \right| +$$

$$\left| a \sum_{Q \in \mathcal{G}} f(\mathbf{x}_Q)\, v(Q) - a \int f\, dx \right| + \left| b \sum_{Q \in \mathcal{G}} g(\mathbf{x}_Q)\, v(Q) - b \int g\, dx \right|$$

$$\leq \varepsilon + |a|\,\varepsilon + |b|\,\varepsilon.$$

Since ε is arbitrary, this establishes Formula (37.7) and shows the integral is linear.

It remains to establish the inequality (37.8). By (37.9), and the triangle inequality for sums,

$$\int |f|\ dx + \varepsilon \geq \sum_{Q\in\mathcal{G}} |f(\mathbf{x}_Q)|\, v(Q) \geq$$

$$\geq \left|\sum_{Q\in\mathcal{G}} f(\mathbf{x}_Q)\, v(Q)\right| \geq \left|\int f\,dx\right| - \varepsilon.$$

Then since ε is arbitrary, this establishes the desired inequality. This proves the corollary.

Which functions are in $\mathcal{R}(\mathbb{R}^n)$? Begin with step functions defined below.

Definition 37.1.3 *If*

$$Q \equiv \prod_{i=1}^{n} [a_i, b_i]$$

is a box, define $\operatorname{int}(Q)$ *as*

$$\operatorname{int}(Q) \equiv \prod_{i=1}^{n} (a_i, b_i).$$

f is called a step function if there is a grid, $\mathcal{G}$ such that f is constant on $\operatorname{int}(Q)$ *for each $Q \in \mathcal{G}$, f is bounded, and $f(\mathbf{x}) = 0$ for all $\mathbf{x}$ outside some bounded set.*

The next corollary states that step functions are in $\mathcal{R}(\mathbb{R}^n)$ and shows the expected formula for the integral is valid.

Corollary 37.1.4 *Let $\mathcal{G}$ be a grid and let f be a step function such that $f = f_Q$ on* $\operatorname{int}(Q)$ *for each $Q \in \mathcal{G}$. Then $f \in \mathcal{R}(\mathbb{R}^n)$ and*

$$\int f\,dx = \sum_{Q\in\mathcal{G}} f_Q v(Q).$$

Proof: Let Q be a box of $\mathcal{G}$,

$$Q \equiv \prod_{i=1}^{n} \left[\alpha^i_{j_i}, \alpha^i_{j_i+1}\right],$$

and suppose g is a bounded function, $|g(\mathbf{x})| \leq C$, and $g = 0$ off Q, and $g = 1$ on $\operatorname{int}(Q)$. Thus, g is the simplest sort of step function. Refine $\mathcal{G}$ by including the extra points,

$$\alpha^i_{j_i} + \eta \text{ and } \alpha^i_{j_i+1} - \eta$$

for each $i = 1, \cdots, n$. Here η is small enough that for each $i, \alpha^i_{j_i} + \eta < \alpha^i_{j_i+1} - \eta$. Also let L denote the largest of the lengths of the sides of Q. Let $\mathcal{F}$ be this refined grid and denote by Q_η the box

$$\prod_{i=1}^{n} \left[\alpha^i_{j_i} + \eta, \alpha^i_{j_i+1} - \eta\right].$$

Now define the box, B^k by

$$B^k \equiv \left[\alpha^1_{j_1}, \alpha^1_{j_1+1}\right] \times \cdots \times \left[\alpha^{k-1}_{j_{k-1}}, \alpha^{k-1}_{j_{k-1}+1}\right] \times$$

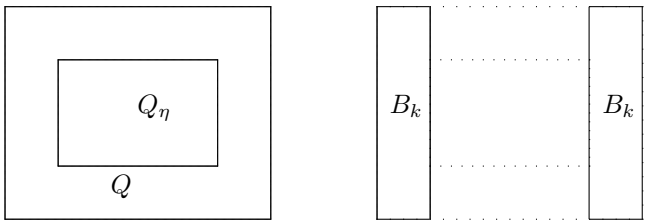
$$\left[\alpha_{j_k}^k, \alpha_{j_k}^k + \eta\right] \times \left[\alpha_{j_{k+1}}^{k+1}, \alpha_{j_{k+1}+1}^{k+1}\right] \times \cdots \times \left[\alpha_{j_n}^n, \alpha_{j_n+1}^n\right]$$

or

$$B^k \equiv \left[\alpha_{j_1}^1, \alpha_{j_1+1}^1\right] \times \cdots \times \left[\alpha_{j_{k-1}}^{k-1}, \alpha_{j_{k-1}+1}^{k-1}\right] \times$$

$$\left[\alpha_{j_k}^k - \eta, \alpha_{j_k}^k\right] \times \left[\alpha_{j_{k+1}}^{k+1}, \alpha_{j_{k+1}+1}^{k+1}\right] \times \cdots \times \left[\alpha_{j_n}^n, \alpha_{j_n+1}^n\right].$$

In words, replace the closed interval in the k^{th} slot used to define Q with a much thinner closed interval at one end or the other while leaving the other intervals used to define Q the same. This is illustrated in the following picture.



The important thing to notice, is that every point of Q is either in Q_η or one of the sets, B_k. Therefore,

$$\mathcal{L}_{\mathcal{F}}(g) \geq v(Q_\eta) - \sum_{k=1}^{n} 2Cv(B_k) \geq v(Q_\eta) - 4CL^{n-1}n\eta$$

$$= v(Q_\eta) - K\eta \tag{37.10}$$

where K is a constant which does not depend on η. Similarly,

$$\mathcal{U}_{\mathcal{F}}(g) \leq v(Q_\eta) + K\eta. \tag{37.11}$$

This implies $\mathcal{U}_{\mathcal{F}}(g) - \mathcal{L}_{\mathcal{F}}(g) < 2K\eta$ and since η is arbitrary, the Riemann criterion verifies that $g \in \mathcal{R}(\mathbb{R}^n)$. Formulas (37.10) and (37.11) also verify that

$$v(Q_\eta) \in [\mathcal{U}_{\mathcal{F}}(g) - K\eta, \mathcal{L}_{\mathcal{F}}(g) + K\eta]$$

$$\subseteq [\mathcal{L}_{\mathcal{F}}(g) - K\eta, \mathcal{U}_{\mathcal{F}}(g) + K\eta].$$

But also

$$\int g\,dx \in [\mathcal{L}_{\mathcal{F}}(g), \mathcal{U}_{\mathcal{F}}(g)] \subseteq [\mathcal{L}_{\mathcal{F}}(g) - K\eta, \mathcal{U}_{\mathcal{F}}(g) + K\eta]$$

and so

$$\left|\int g\,dx - v(Q_\eta)\right| \leq 4K\eta.$$

Now letting $\eta \to 0$, yields $\int g\,dx = v(Q)$.

Now let f be as described in the statement of the Corollary. Let f_Q be the value of f on $\text{int}(Q)$, and let g_Q be a function of the sort just considered which equals 1 on $\text{int}(Q)$. Then f is of the form

$$f = \sum_{Q \in \mathcal{G}} f_Q g_Q$$

with all but finitely many of the f_Q equal zero. Therefore, the above is really a finite sum and so by Corollary 37.1.2, $f \in \mathcal{R}(\mathbb{R}^n)$ and

$$\int f\,dx = \sum_{Q\in\mathcal{G}} f_Q \int g_Q\,dx = \sum_{Q\in\mathcal{G}} f_Q v(Q).$$

There is a good deal of sloppiness inherent in the above description of a step function due to the fact that the boxes may be different but match up on an edge. It is convenient to be able to consider a more precise sort of function and this is done next.

For Q a box of the form

$$Q = \prod_{i=1}^{k} [a_i, b_i],$$

define the half open box, Q' by

$$Q' = \prod_{i=1}^{k} (a_i, b_i].$$

The reason for considering these sets is that if $\mathcal{G}$ is a grid, the sets, Q' where $Q \in \mathcal{G}$ are disjoint. Defining a step function, ϕ as

$$\phi(\mathbf{x}) \equiv \sum_{Q\in\mathcal{G}} \phi_Q \mathcal{X}_{Q'}(\mathbf{x}),$$

the number, ϕ_Q is the value of ϕ on the set, Q'. As before, define

$$M_{Q'}(f) \equiv \sup\{f(\mathbf{x}) : \mathbf{x} \in Q'\},\ m_{Q'}(f) \equiv \inf\{f(\mathbf{x}) : \mathbf{x} \in Q'\}.$$

The next lemma will be convenient a little later.

Lemma 37.1.5 *Suppose f is a bounded function which equals zero off some bounded set. Then $f \in \mathcal{R}(\mathbb{R}^n)$ if and only if for all $\varepsilon > 0$ there exists a grid, $\mathcal{G}$ such that*

$$\sum_{Q\in\mathcal{G}} (M_{Q'}(f) - m_{Q'}(f)) v(Q) < \varepsilon. \tag{37.12}$$

Proof: Since $Q' \subseteq Q$,

$$M_{Q'}(f) - m_{Q'}(f) \le M_Q(f) - m_Q(f)$$

and therefore, the only if part of the equivalence is obvious.

Conversely, let $\mathcal{G}$ be a grid such that (37.12) holds with ε replaced with $\frac{\varepsilon}{2}$. It is necessary to show there is a grid such that (37.12) holds with no primes on the Q. Let $\mathcal{F}$ be a refinement of $\mathcal{G}$ obtained by adding the points $\alpha_k^i + \eta_k$ where $\eta_k \le \eta$ and is also chosen so small that for each $i = 1, \cdots, n$,

$$\alpha_k^i + \eta_k < \alpha_{k+1}^i.$$

Then for

$$Q \equiv \prod_{i=1}^{n} \left[\alpha_{k_i}^i, \alpha_{k_i+1}^i\right] \in \mathcal{G},$$

Let

$$\widehat{Q} \equiv \prod_{i=1}^{n} \left[\alpha_{k_i}^i + \eta_{k_i}, \alpha_{k_i+1}^i\right]$$

and denote by $\widehat{\mathcal{G}}$ the collection of these smaller boxes. For each set, Q in $\mathcal{G}$ there is the smaller set, $\widehat{Q}$ along with n boxes, $B_k, k = 1, \cdots, n$, one of whose sides is of length η_k and the remainder of whose sides are shorter than the diameter of Q such that the set, Q is the union of $\widehat{Q}$ and these sets, B_k. Now suppose f equals zero off the ball $B\left(\mathbf{0}, \frac{R}{2}\right)$. Then without loss of generality, you may assume the diameter of every box in $\mathcal{G}$ which has nonempty intersection with $B(\mathbf{0},R)$ is smaller than $\frac{R}{3}$. (If this is not so, simply refine $\mathcal{G}$ to make it so, such a refinement leaving (37.12) valid.) Suppose there are P sets of $\mathcal{G}$ contained in $B(\mathbf{0},R)$ and suppose that for all $\mathbf{x}$, $|f(\mathbf{x})| < C/2$. Then

$$\begin{aligned}\sum_{Q\in\mathcal{F}} \left(M_Q(f) - m_Q(f)\right) v(Q) &\leq \sum_{\widehat{Q}\in\widehat{\mathcal{G}}} \left(M_Q(f) - m_Q(f)\right) v(Q) \\ &+ \sum_{Q\in\mathcal{F}\setminus\widehat{\mathcal{G}}} \left(M_Q(f) - m_Q(f)\right) v(Q) \\ &\leq \varepsilon/2 + CPnR^{n-1}\eta < \varepsilon\end{aligned}$$

whenever η is small enough. Since ε is arbitrary, $f \in \mathcal{R}(\mathbb{R}^n)$ as claimed.

Definition 37.1.6 *A bounded set, E is a Jordan set in $\mathbb{R}^n$ or a contented set in $\mathbb{R}^n$ if $\mathcal{X}_E \in \mathcal{R}(\mathbb{R}^n)$. Also, for $\mathcal{G}$ a grid and E a set, denote by $\partial_{\mathcal{G}}(E)$ those boxes of $\mathcal{G}$ which have nonempty intersection with both E and $\mathbb{R}^n \setminus E$.*

The next theorem is a characterization of those sets which are Jordan sets.

Theorem 37.1.7 *A bounded set, E, is a Jordan set if and only if for every $\varepsilon > 0$ there exists a grid, $\mathcal{G}$, such that*

$$\sum_{Q\in\partial_{\mathcal{G}}(E)} v(Q) < \varepsilon.$$

Proof: If $Q \notin \partial_{\mathcal{G}}(E)$, then

$$M_Q(\mathcal{X}_E) - m_Q(\mathcal{X}_E) = 0$$

and if $Q \in \partial_{\mathcal{G}}(E)$, then

$$M_Q(\mathcal{X}_E) - m_Q(\mathcal{X}_E) = 1.$$

It follows that $\mathcal{U}_{\mathcal{G}}(\mathcal{X}_E) - \mathcal{L}_{\mathcal{G}}(\mathcal{X}_E) = \sum_{Q\in\partial_{\mathcal{G}}(E)} v(Q)$ and this implies the conclusion of the theorem.

Note that if E is a Jordan set and if $f \in \mathcal{R}(\mathbb{R}^n)$, then by Corollary 37.1.2, $\mathcal{X}_E f \in \mathcal{R}(\mathbb{R}^n)$.

Definition 37.1.8 *For E a Jordan set and $f\mathcal{X}_E \in \mathcal{R}(\mathbb{R}^n)$.*

$$\int_E f\,dV \equiv \int_{\mathbb{R}^n} \mathcal{X}_E f\,dV.$$

A bounded set, E, has Jordan content 0 or content 0 if for every $\varepsilon > 0$ there exists a grid, $\mathcal{G}$ such that

$$\sum_{Q\cap E\neq\emptyset} v(Q) < \varepsilon.$$

This symbol says to sum the volumes of all boxes from $\mathcal{G}$ which have nonempty intersection with E.

Note that any finite union of sets having Jordan content 0 also has Jordan content 0. (Why?)

Definition 37.1.9 *Let A be any subset of $\mathbb{R}^n$. Then ∂A denotes those points, $\mathbf{x}$ with the property that if U is any open set containing $\mathbf{x}$, then U contains points of A as well as points of A^C.*

Corollary 37.1.10 *If a bounded set, $E \subseteq \mathbb{R}^n$ is contented, then ∂E has content 0.*

Proof: Let $\varepsilon > 0$ be given and suppose E is contented. Then there exists a grid, $\mathcal{G}$ such that

$$\sum_{Q \in \partial_{\mathcal{G}}(E)} v(Q) < \frac{\varepsilon}{2n+1}. \tag{37.13}$$

Now refine $\mathcal{G}$ if necessary to get a new grid, $\mathcal{F}$ such that all boxes from $\mathcal{F}$ which have nonempty intersection with ∂E have sides no larger than δ where δ is the smallest of all the sides of all the Q in the above sum. Recall that $\partial_{\mathcal{G}}(E)$ consists of those boxes of $\mathcal{G}$ which have nonempty intersection with both E and $\mathbb{R}^n \setminus E$.

Let $\mathbf{x} \in \partial E$. Then since the dimension is n, there are at most 2^n boxes from $\mathcal{F}$ which contain $\mathbf{x}$. Furthermore, at least one of these boxes is in $\partial_{\mathcal{F}}(E)$ and is therefore a subset of a box from $\partial_{\mathcal{G}}(E)$. Here is why. If $\mathbf{x}$ is an interior point of some $Q \in \mathcal{F}$, then there are points of both E and E^C contained in Q and so $\mathbf{x} \in Q \in \partial_{\mathcal{F}}(E)$ and there are no other boxes from $\mathcal{F}$ which contain $\mathbf{x}$. If $\mathbf{x}$ is not an interior point of any $Q \in \mathcal{F}$, then the interior of the union of all the boxes from $\mathcal{F}$ which do contain $\mathbf{x}$ is an open set and therefore, must contain points of E and points from E^C. If $\mathbf{x} \in E$, then one of these boxes must contain points which are not in E since otherwise, $\mathbf{x}$ would fail to be in ∂E. Pick that box. It is in $\partial_{\mathcal{F}}(E)$ and contains $\mathbf{x}$. On the other hand, if $\mathbf{x} \notin E$, one of these boxes must contain points of E since otherwise, $\mathbf{x}$ would fail to be in ∂E. Pick that box. This shows that every set from $\mathcal{F}$ which contains a point of ∂E shares this point with a box of $\partial_{\mathcal{G}}(E)$. Let the boxes from $\partial_{\mathcal{G}}(E)$ be $\{P_1, \cdots, P_m\}$. Let $\mathcal{S}(P_i)$ denote those sets of $\mathcal{F}$ which contain a point of ∂E in common with P_i. Then if $Q \in \mathcal{S}(P_i)$, either $Q \subseteq P_i$ or it intersects P_i on one of its $2n$ faces. Therefore, the sum of the volumes of those boxes of $\mathcal{S}(P_i)$ which intersect P_i on a particular face of P_i is no larger than $v(P_i)$. Consequently,

$$\sum_{Q \in \mathcal{S}(P_i)} v(Q) \leq 2n v(P_i) + v(P_i)$$

and so for $Q \in \mathcal{F}$,

$$\sum_{Q \cap \partial E \neq \emptyset} v(Q) = \sum_{i=1}^{m} \sum_{Q \in \mathcal{S}(P_i)} v(Q) \leq \sum_{i=1}^{m} (2n+1)\, v(P_i) < \varepsilon$$

from (37.13). This proves the corollary.

Theorem 37.1.11 *If a bounded set, E, has Jordan content 0, then E is a Jordan set and if f is any bounded function defined on E, then $f\mathcal{X}_E \in \mathcal{R}(\mathbb{R}^n)$ and*

$$\int_E f\, dV = 0.$$

Proof: Let $\varepsilon > 0$. Then let $\mathcal{G}$ be a grid such that

$$\sum_{Q \cap E \neq \emptyset} v(Q) < \varepsilon.$$

Then every set of $\partial_{\mathcal{G}}(E)$ contains a point of E so

$$\sum_{Q \in \partial_{\mathcal{G}}(E)} v(Q) \leq \sum_{Q \cap E \neq \emptyset} v(Q) < \varepsilon$$

and since ε was arbitrary, this shows from Theorem 37.1.7 that E is a Jordan set. Now let M be a positive number larger than all values of f, let m be a negative number smaller than all values of f and let $\varepsilon > 0$ be given. Let $\mathcal{G}$ be a grid with

$$\sum_{Q\cap E\neq\emptyset} v(Q) < \frac{\varepsilon}{1+(M-m)}.$$

Then

$$\mathcal{U}_{\mathcal{G}}(f\mathcal{X}_E) \leq \sum_{Q\cap E\neq\emptyset} Mv(Q) \leq \frac{\varepsilon M}{1+(M-m)}$$

and

$$\mathcal{L}_{\mathcal{G}}(f\mathcal{X}_E) \geq \sum_{Q\cap E\neq\emptyset} mv(Q) \geq \frac{\varepsilon m}{1+(M-m)}$$

and so

$$\begin{aligned}\mathcal{U}_{\mathcal{G}}(f\mathcal{X}_E) - \mathcal{L}_{\mathcal{G}}(f\mathcal{X}_E) &\leq \sum_{Q\cap E\neq\emptyset} Mv(Q) - \sum_{Q\cap E\neq\emptyset} mv(Q)\\ &= (M-m)\sum_{Q\cap E\neq\emptyset} v(Q) < \frac{\varepsilon(m-N)}{1+(M-m)} < \varepsilon.\end{aligned}$$

This shows $f\mathcal{X}_E \in \mathcal{R}(\mathbb{R}^n)$. Now also,

$$m\varepsilon \leq \int f\mathcal{X}_E\, dV \leq M\varepsilon$$

and since ε is arbitrary, this shows

$$\int_E f\, dV \equiv \int f\mathcal{X}_E\, dV = 0$$

and proves the theorem.

Corollary 37.1.12 *If $f\mathcal{X}_{E_i} \in \mathcal{R}(\mathbb{R}^n)$ for $i = 1, 2, \cdots, r$ and for all $i \neq j$, $E_i \cap E_j$ is either the empty set or a set of Jordan content 0, then letting $F \equiv \cup_{i=1}^r E_i$, it follows $f\mathcal{X}_F \in \mathcal{R}(\mathbb{R}^n)$ and*

$$\int f\mathcal{X}_F\, dV \equiv \int_F f\, dV = \sum_{i=1}^r \int_{E_i} f\, dV.$$

Proof: By Corollary 37.1.2, this is true if $r = 1$. Suppose it is true for r. It will be shown that it is true for $r+1$. Let $F_r = \cup_{i=1}^r E_i$ and let F_{r+1} be defined similarly. By the induction hypothesis, $f\mathcal{X}_{F_r} \in \mathcal{R}(\mathbb{R}^n)$. Also, since F_r is a finite union of the E_i, it follows that $F_r \cap E_{r+1}$ is either empty or a set of Jordan content 0.

$$-f\mathcal{X}_{F_r\cap E_{r+1}} + f\mathcal{X}_{F_r} + f\mathcal{X}_{E_{r+1}} = f\mathcal{X}_{F_{r+1}}$$

and by Theorem 37.1.11 each function on the left is in $\mathcal{R}(\mathbb{R}^n)$ and the first one on the left has integral equal to zero. Therefore,

$$\int f\mathcal{X}_{F_{r+1}}\, dV = \int f\mathcal{X}_{F_r}\, dV + \int f\mathcal{X}_{E_{r+1}}\, dV$$

which by induction equals

$$\sum_{i=1}^r \int_{E_i} f\, dV + \int_{E_{r+1}} f\, dV = \sum_{i=1}^{r+1} \int_{E_i} f\, dV$$

and this proves the corollary.

What functions in addition to step functions are integrable? As in the case of integrals of functions of one variable, this is an important question. It turns out that the Riemann integrable functions are characterized by being continuous except on a very small set. To begin with it is necessary to define the oscillation of a function.

Definition 37.1.13 *Let f be a function defined on $\mathbb{R}^n$ and let*

$$\omega_{f,r}(\mathbf{x}) \equiv \sup\{|f(\mathbf{z}) - f(\mathbf{y})| : \mathbf{z}, \mathbf{y} \in B(\mathbf{x},r)\}.$$

This is called the oscillation of f on $B(\mathbf{x},r)$. Note that this function of r is decreasing in r. Define the oscillation of f as

$$\omega_f(\mathbf{x}) \equiv \lim_{r\to 0+} \omega_{f,r}(\mathbf{x}).$$

Note that as r decreases, the function, $\omega_{f,r}(\mathbf{x})$ decreases. It is also bounded below by 0 and so the limit must exist and equals $\inf\{\omega_{f,r}(\mathbf{x}) : r > 0\}$. (Why?) Then the following simple lemma whose proof follows directly from the definition of continuity gives the reason for this definition.

Lemma 37.1.14 *A function, f, is continuous at $\mathbf{x}$ if and only if $\omega_f(\mathbf{x}) = 0$.*

This concept of oscillation gives a way to define how discontinuous a function is at a point. The discussion will depend on the following fundamental lemma which gives the existence of something called the Lebesgue number.

Definition 37.1.15 *Let $\mathfrak{C}$ be a set whose elements are sets of $\mathbb{R}^n$ and let $K \subseteq \mathbb{R}^n$. The set, $\mathfrak{C}$ is called a cover of K if every point of K is contained in some set of $\mathfrak{C}$. If the elements of $\mathfrak{C}$ are open sets, it is called an open cover.*

Lemma 37.1.16 *Let K be sequentially compact and let $\mathfrak{C}$ be an open cover of K. Then there exists $r > 0$ such that whenever $\mathbf{x} \in K$, $B(\mathbf{x}, r)$ is contained in some set of $\mathfrak{C}$.*

Proof: Suppose this is not so. Then letting $r_n = 1/n$, there exists $\mathbf{x}_n \in K$ such that $B(\mathbf{x}_n, r_n)$ is not contained in any set of $\mathfrak{C}$. Since K is sequentially compact, there is a subsequence, $\mathbf{x}_{n_k}$ which converges to a point, $\mathbf{x} \in K$. But there exists $\delta > 0$ such that $B(\mathbf{x}, \delta) \subseteq U$ for some $U \in \mathfrak{C}$. Let k be so large that $1/k < \delta/2$ and $|\mathbf{x}_{n_k} - \mathbf{x}| < \delta/2$ also. Then if $\mathbf{z} \in B(\mathbf{x}_{n_k}, r_{n_k})$, it follows

$$|\mathbf{z} - \mathbf{x}| \leq |\mathbf{z} - \mathbf{x}_{n_k}| + |\mathbf{x}_{n_k} - \mathbf{x}| < \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

and so $B(\mathbf{x}_{n_k}, r_{n_k}) \subseteq U$ contrary to supposition. Therefore, the desired number exists after all.

Theorem 37.1.17 *Let f be a bounded function which equals zero off a bounded set and let W denote the set of points where f fails to be continuous. Then $f \in \mathcal{R}(\mathbb{R}^n)$ if W has content zero. That is, for all $\varepsilon > 0$ there exists a grid, $\mathcal{G}$ such that*

$$\sum_{Q\in\mathcal{G}_W} v(Q) < \varepsilon \tag{37.14}$$

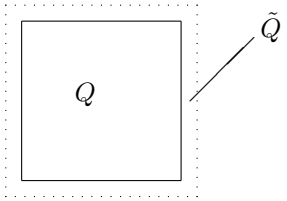
where

$$\mathcal{G}_W \equiv \{Q \in \mathcal{G} : Q \cap W \neq \emptyset\}.$$

Proof: Let $|f(\mathbf{x})| < C/2$ for all $\mathbf{x} \in \mathbb{R}^n$, let $\varepsilon > 0$ be given, and let $\mathcal{G}$ be a grid which satisfies (37.14). Since f equals zero off some bounded set, there exists R such that f equals zero off of $B\left(\mathbf{0}, \frac{R}{2}\right)$. Thus $W \subseteq B\left(\mathbf{0}, \frac{R}{2}\right)$. Also note that if $\mathcal{G}$ is a grid for which (37.14) holds, then this inequality continues to hold if $\mathcal{G}$ is replaced with a refined grid. Therefore, you may assume the diameter of every box in $\mathcal{G}$ which intersects $B(\mathbf{0}, R)$ is less than $\frac{R}{3}$. Since W is bounded, $\mathcal{G}_W$ contains only finitely many boxes. Letting

$$Q \equiv \prod_{i=1}^{n} [a_i, b_i]$$

be one of these boxes, enlarge the box slightly as indicated in the following picture.



The enlarged box is an open set of the form,

$$\widetilde{Q} \equiv \prod_{i=1}^{n} (a_i - \eta_i, b_i + \eta_i)$$

where η_i is chosen small enough that if

$$\prod_{i=1}^{n} (\ b_i + \eta_i - (a_i - \eta_i)) \equiv v\left(\widetilde{Q}\right),$$

then

$$\sum_{Q \in \mathcal{G}_W} v\left(\widetilde{Q}\right) < \varepsilon.$$

For each $\mathbf{x} \in \mathbb{R}^n$, let $r_{\mathbf{x}}$ be such that

$$\omega_{f, r_{\mathbf{x}}}(\mathbf{x}) < \varepsilon + \omega_f(\mathbf{x}). \tag{37.15}$$

Now let $\mathfrak{C}$ denote all intersections of the form $\widetilde{Q} \cap B(\mathbf{x}, r_{\mathbf{x}})$ such that $\mathbf{x} \in \overline{B(\mathbf{0}, R)}$ so that $\mathfrak{C}$ is an open cover of the compact set, $\overline{B(\mathbf{0}, R)}$. Let δ be a Lebesgue number for this open cover of $\overline{B(\mathbf{0}, R)}$ and let $\mathcal{F}$ be a refinement of $\mathcal{G}$ such that every box in $\mathcal{F}$ has diameter less than δ. Now let $\mathcal{F}_1$ consist of those boxes of $\mathcal{F}$ which have nonempty intersection with $B(\mathbf{0}, R/2)$. Thus all boxes of $\mathcal{F}_1$ are contained in $B(\mathbf{0}, R)$ and each one is contained in some set of $\mathfrak{C}$. Now let $\mathfrak{C}_W$ be those open sets of $\mathfrak{C}$, $\widetilde{Q} \cap B(\mathbf{x}, r_{\mathbf{x}})$, for which $\mathbf{x} \in W$ and let $\mathcal{F}_W$ be those sets of $\mathcal{F}_1$ which are subsets of some set of $\mathfrak{C}_W$. Then

$$\mathcal{U}_{\mathcal{F}}(f) - \mathcal{L}_{\mathcal{F}}(f) = \sum_{Q \in \mathcal{F}_W} (M_Q(f) - m_Q(f))\, v(Q)$$

$$+ \sum_{Q \in \mathcal{F}_1 \setminus \mathcal{F}_W} (M_Q(f) - m_Q(f))\, v(Q).$$

If $Q \in \mathcal{F}_1 \setminus \mathcal{F}_W$, then Q must be a subset of some set of $\mathfrak{C} \setminus \mathfrak{C}_W$ since it is not in any set of $\mathfrak{C}_W$. Therefore, from (37.15) and the observation that $\mathbf{x} \notin W$,

$$M_Q(f) - m_Q(f) \leq \varepsilon.$$

Therefore,

$$\begin{aligned}\mathcal{U}_{\mathcal{F}}(f)-\mathcal{L}_{\mathcal{F}}(f) &\leq \sum_{Q\in\mathcal{F}_W} Cv(Q)+\sum_{Q\in\mathcal{F}_1\setminus\mathcal{F}_W}\varepsilon v(Q)\\ &\leq C\varepsilon+\varepsilon(2R)^n.\end{aligned}$$

Since ε is arbitrary, this proves the theorem.[1]

From Theorem 37.1.7 you get a pretty good idea of what constitutes a contented set. These sets are essentially those which have thin boundaries. Most sets you are likely to think of will fall in this category. However, it is good to give specific examples of sets which are contented.

Theorem 37.1.18 *Suppose E is a bounded contented set in $\mathbb{R}^n$ and $f,g:E\to\mathbb{R}$ are two functions satisfying $f(\mathbf{x})\geq g(\mathbf{x})$ for all $\mathbf{x}\in E$ and $f\mathcal{X}_E$ and $g\mathcal{X}_E$ are both in $\mathcal{R}(\mathbb{R}^n)$. Now define*

$$P\equiv\{(\mathbf{x},x_{n+1}):\mathbf{x}\in E \text{ and } g(\mathbf{x})\leq x_{n+1}\leq f(\mathbf{x})\}.$$

Then P is a contented set in $\mathbb{R}^{n+1}$.

Proof: Let $\mathcal{G}$ be a grid such that for $k=f,g$,

$$\mathcal{U}_{\mathcal{G}}(k)-\mathcal{L}_{\mathcal{G}}(k)<\varepsilon/2. \tag{37.16}$$

Let the boxes of $\mathcal{G}$ which have nonempty intersection with E be $\{Q_1,\cdots,Q_m\}$ and let $\{a_i\}_{i=-\infty}^{\infty}$ be a sequence on $\mathbb{R}$, $a_i<a_{i+1}$ for all i, which includes

$$M_{Q_j}(f\mathcal{X}_E),M_{Q_j}(g\mathcal{X}_E),m_{Q_j}(f\mathcal{X}_E),m_{Q_j}(g\mathcal{X}_E)$$

for all $j=1,\cdots,m$. Now define a grid on $\mathbb{R}^{n+1}$ as follows.

$$\mathcal{G}'\equiv\{Q\times[a_i,a_{i+1}]:Q\in\mathcal{G},\,i\in\mathbb{Z}\}$$

In words, this grid consists of all possible boxes of the form $Q\times[a_i,a_{i+1}]$ where $Q\in\mathcal{G}$ and a_i is a term of the sequence just described. It is necessary to verify that $\mathcal{X}_P\in\mathcal{R}\left(\mathbb{R}^{n+1}\right)$. This is done by showing that $\mathcal{U}_{\mathcal{G}'}(\mathcal{X}_P)-\mathcal{L}_{\mathcal{G}'}(\mathcal{X}_P)<\varepsilon$ and then noting that $\varepsilon>0$ was arbitrary. For $\mathcal{G}'$ just described, denote by Q' a box in $\mathcal{G}'$. Thus $Q'=Q\times[a_i,a_{i+1}]$ for some i.

$$\begin{aligned}\mathcal{U}_{\mathcal{G}'}(\mathcal{X}_P)-\mathcal{L}_{\mathcal{G}'}(\mathcal{X}_P) &\equiv \sum_{Q'\in\mathcal{G}'}\left(M_{Q'}(\mathcal{X}_P)-m_{Q'}(\mathcal{X}_P)\right)v_{n+1}(Q')\\ &= \sum_{i=-\infty}^{\infty}\sum_{j=1}^{m}\left(M_{Q_j'}(\mathcal{X}_P)-m_{Q_j'}(\mathcal{X}_P)\right)v_n(Q_j)(a_{i+1}-a_i)\end{aligned}$$

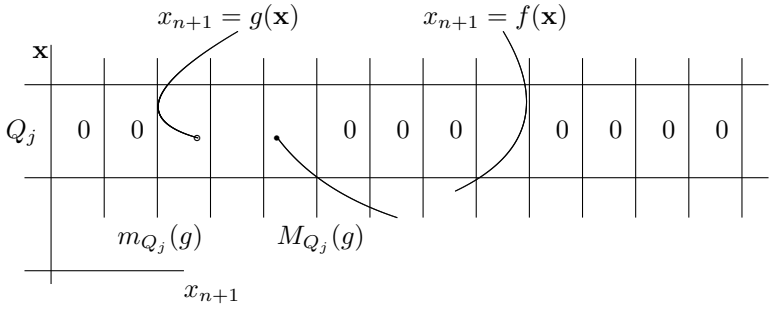
and all sums are bounded because the functions, f and g are given to be bounded. Therefore, there are no limit considerations needed here. Thus

$$\mathcal{U}_{\mathcal{G}'}(\mathcal{X}_P)-\mathcal{L}_{\mathcal{G}'}(\mathcal{X}_P)=$$

$$\sum_{j=1}^{m}v_n(Q_j)\sum_{i=-\infty}^{\infty}\left(M_{Q_j\times[a_i,a_{i+1}]}(\mathcal{X}_P)-m_{Q_j\times[a_i,a_{i+1}]}(\mathcal{X}_P)\right)(a_{i+1}-a_i).$$

[1] In fact one cannot do any better. It can be shown that if a function is Riemann integrable, then it must be the case that for all $\varepsilon>0$, (37.14) is satisfied for some grid, $\mathcal{G}$. This along with what was just shown is known as Lebesgue's theorem after Lebesgue who discovered it in the early years of the twentieth century. Actually, he also invented a far superior integral which has been the integral of serious mathematicians since that time.

Consider the inside sum with the aid of the following picture.



In this picture, the little rectangles represent the boxes $Q_j \times [a_i, a_{i+1}]$ for fixed j. The part of P having $\mathbf{x}$ contained in Q_j is between the two surfaces, $x_{n+1} = g(\mathbf{x})$ and $x_{n+1} = f(\mathbf{x})$ and there is a zero placed in those boxes for which $M_{Q_j \times [a_i,a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i,a_{i+1}]}(\mathcal{X}_P) = 0$. You see, $\mathcal{X}_P$ has either the value of 1 or the value of 0 depending on whether $(\mathbf{x}, y)$ is contained in P. For the boxes shown with 0 in them, either all of the box is contained in P or none of the box is contained in P. Either way, $M_{Q_j \times [a_i,a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i,a_{i+1}]}(\mathcal{X}_P) = 0$ on these boxes. However, on the boxes intersected by the surfaces, the value of $M_{Q_j \times [a_i,a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i,a_{i+1}]}(\mathcal{X}_P)$ is 1 because there are points in this box which are not in P as well as points which are in P. Because of the construction of $\mathcal{G}'$ which included all values of $M_{Q_j}(f\mathcal{X}_E), M_{Q_j}(g\mathcal{X}_E), m_{Q_j}(f\mathcal{X}_E), m_{Q_j}(g\mathcal{X}_E)$ for all $j = 1, \cdots, m$,

$$\sum_{i=-\infty}^{\infty} \left(M_{Q_j \times [a_i,a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i,a_{i+1}]}(\mathcal{X}_P)\right)(a_{i+1} - a_i) \leq$$

$$\sum_{\{i: m_{Q_j}(g) \leq a_i < M_{Q_j}(g)\}} 1\,(a_{i+1} - a_i) + \sum_{\{i: m_{Q_j}(f) \leq a_i < M_{Q_j}(f)\}} 1\,(a_{i+1} - a_i)$$

$$= \left(M_{Q_j}(g) - m_{Q_j}(g)\right) + \left(M_{Q_j}(f) - m_{Q_j}(f)\right).$$

(Note the inequality.) Therefore, by (37.16),

$$\begin{aligned} \mathcal{U}_{\mathcal{G}'}(\mathcal{X}_P) - \mathcal{L}_{\mathcal{G}'}(\mathcal{X}_P) &\leq \sum_{j=1}^{m} v_n(Q_j)\left[\left(M_{Q_j}(g) - m_{Q_j}(g)\right) + \left(M_{Q_j}(f) - m_{Q_j}(f)\right)\right] \\ &= \mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{G}}(f) + \mathcal{U}_{\mathcal{G}}(g) - \mathcal{L}_{\mathcal{G}}(g) < \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, this proves the theorem.

Corollary 37.1.19 *Suppose f and g are continuous functions defined on E, a contented set in $\mathbb{R}^n$ and that $g(\mathbf{x}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in E$. Then*

$$P \equiv \{(\mathbf{x}, x_{n+1}) : \mathbf{x} \in E \text{ and } g(\mathbf{x}) \leq x_{n+1} \leq f(\mathbf{x})\}$$

is a contented set in $\mathbb{R}^n$.

Proof: Extend f and g to equal 0 off E. The set of discontinuities of f and g is contained in ∂E and Corollary 37.1.10 on Page 828 implies this is a set of content 0. Therefore, from Theorem 37.1.17, for $k = f, g$, it follows that $k\mathcal{X}_E$ is in $\mathcal{R}(\mathbb{R}^n)$ because the set of discontinuities is contained in ∂E. The conclusion now follows from Theorem 37.1.18. This proves the corollary.

As an example of how this can be applied, it is obvious a closed interval is a contented set in $\mathbb{R}$. Therefore, if f, g are two continuous functions with $f(x) \geq g(x)$ for $x \in [a,b]$, it follows from the above theorem or its corollary that the set,

$$P_1 \equiv \{(x,y) : g(x) \leq y \leq f(x)\}$$

is a contented set in $\mathbb{R}^2$. Now using the theorem and corollary again, suppose $f_1(x,y) \geq g_1(x,y)$ for $(x,y) \in P_1$ and f, g are continuous. Then the set

$$P_2 \equiv \{(x,y,z) : g_1(x,y) \leq z \leq f_1(x,y)\}$$

is a contented set in $\mathbb{R}^3$. Clearly you can continue this way obtaining examples of contented sets.

Note that as a special case of Corollary 37.1.4 on Page 824, it follows that every box is a contented set.

37.2 Iterated Integrals

To evaluate an n dimensional Riemann integral, one uses iterated integrals. Formally, an iterated integral is defined as follows. For f a function defined on $\mathbb{R}^{n+m}$,

$$\mathbf{y} \to f(\mathbf{x}, \mathbf{y})$$

is a function of $\mathbf{y}$ for each $\mathbf{x} \in \mathbb{R}^{n+m}$. Therefore, it might be possible to integrate this function of $\mathbf{y}$ and write

$$\int_{\mathbb{R}^m} f(\mathbf{x}, \mathbf{y})\, dV_y.$$

Now the result is clearly a function of $\mathbf{x}$ and so, it might be possible to integrate this and write

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} f(\mathbf{x}, \mathbf{y})\, dV_y\, dV_x.$$

This symbol is called an iterated integral, because it involves the iteration of two lower dimensional integrations. Under what conditions are the two iterated integrals equal to the integral

$$\int_{\mathbb{R}^{n+m}} f(\mathbf{z})\, dV?$$

Definition 37.2.1 *Let $\mathcal{G}$ be a grid on $\mathbb{R}^{n+m}$ defined by the $n+m$ sequences,*

$$\left\{\alpha_k^i\right\}_{k=-\infty}^{\infty} \quad i = 1, \cdots, n+m.$$

Let $\mathcal{G}_n$ be the grid on $\mathbb{R}^n$ obtained by considering only the first n of these sequences and let $\mathcal{G}_m$ be the grid on $\mathbb{R}^m$ obtained by considering only the last m of the sequences. Thus a typical box in $\mathcal{G}_m$ would be

$$\prod_{i=n+1}^{n+m} \left[\alpha_{k_i}^i, \alpha_{k_i+1}^i\right], \; k_i \geq n+1$$

and a box in $\mathcal{G}_n$ would be of the form

$$\prod_{i=1}^{n} \left[\alpha_{k_i}^i, \alpha_{k_i+1}^i\right], \; k_i \leq n.$$

Lemma 37.2.2 *Let $\mathcal{G}$, $\mathcal{G}_n$, and $\mathcal{G}_m$ be the grids defined above. Then*

$$\mathcal{G} = \{R \times P : R \in \mathcal{G}_n \text{ and } P \in \mathcal{G}_m\}.$$

Proof: If $Q \in \mathcal{G}$, then Q is clearly of this form. On the other hand, if $R \times P$ is one of the sets described above, then from the above description of R and P, it follows $R \times P$ is one of the sets of $\mathcal{G}$.

Now let $\mathcal{G}$ be a grid on $\mathbb{R}^{n+m}$ and suppose

$$\phi(\mathbf{z}) = \sum_{Q \in \mathcal{G}} \phi_Q \mathcal{X}_{Q'}(\mathbf{z}) \tag{37.17}$$

where ϕ_Q equals zero for all but finitely many Q. Thus ϕ is a step function. Recall that for

$$Q = \prod_{i=1}^{n+m} [a_i, b_i), \; Q' \equiv \prod_{i=1}^{n+m} (a_i, b_i]$$

Letting $(\mathbf{x}, \mathbf{y}) = \mathbf{z}$, Lemma 37.2.2 implies

$$\phi(\mathbf{z}) = \phi(\mathbf{x}, \mathbf{y}) = \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \mathcal{X}_{R' \times P'}(\mathbf{x}, \mathbf{y})$$

$$= \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \mathcal{X}_{R'}(\mathbf{x}) \mathcal{X}_{P'}(\mathbf{y}). \tag{37.18}$$

For a function of two variables, h, denote by $h(\cdot, \mathbf{y})$ the function, $\mathbf{x} \to h(\mathbf{x}, \mathbf{y})$ and $h(\mathbf{x}, \cdot)$ the function $\mathbf{y} \to h(\mathbf{x}, \mathbf{y})$. The following lemma is a preliminary version of Fubini's theorem.

Lemma 37.2.3 *Let ϕ be a step function as described in (37.17). Then*

$$\phi(\mathbf{x}, \cdot) \in \mathcal{R}(\mathbb{R}^m), \tag{37.19}$$

$$\int_{\mathbb{R}^m} \phi(\cdot, \mathbf{y}) \, dV_y \in \mathcal{R}(\mathbb{R}^n), \tag{37.20}$$

and

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \phi(\mathbf{x}, \mathbf{y}) \, dV_y \, dV_x = \int_{\mathbb{R}^{n+m}} \phi(\mathbf{z}) \, dV. \tag{37.21}$$

Proof: To verify (37.19), note that $\phi(\mathbf{x}, \cdot)$ is the step function

$$\phi(\mathbf{x}, \mathbf{y}) = \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \mathcal{X}_{P'}(\mathbf{y}).$$

Where $\mathbf{x} \in R'$. By Corollary 37.1.4, this verifies (37.19). From the description in (37.18) and this corollary,

$$\int_{\mathbb{R}^m} \phi(\mathbf{x}, \mathbf{y}) \, dV_y = \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \mathcal{X}_{R'}(\mathbf{x}) v(P)$$

$$= \sum_{R \in \mathcal{G}_n} \left(\sum_{P \in \mathcal{G}_m} \phi_{R \times P} v(P) \right) \mathcal{X}_{R'}(\mathbf{x}), \tag{37.22}$$

another step function. Therefore, Corollary 37.1.4 applies again to verify (37.20). Finally, (37.22) implies

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^m}\phi(\mathbf{x},\mathbf{y})\,dV_y\,dV_x=\sum_{R\in\mathcal{G}_n}\sum_{P\in\mathcal{G}_m}\phi_{R\times P}v(P)\,v(R)$$

$$=\sum_{Q\in\mathcal{G}}\phi_Q v(Q)=\int_{\mathbb{R}^{n+m}}\phi(\mathbf{z})\,dV.$$

and this proves the lemma.

From (37.22),

$$M_{R_1'}\left(\int_{\mathbb{R}^m}\phi(\cdot,\mathbf{y})\,dV_y\right)\equiv\sup\left\{\sum_{R\in\mathcal{G}_n}\left(\sum_{P\in\mathcal{G}_m}\phi_{R\times P}v(P)\right)\mathcal{X}_{R'}(\mathbf{x}):\mathbf{x}\in R_1'\right\}$$

$$=\sum_{P\in\mathcal{G}_m}\phi_{R_1\times P}v(P).\tag{37.23}$$

Similarly,

$$m_{R_1'}\left(\int_{\mathbb{R}^m}\phi(\cdot,\mathbf{y})\,dV_y\right)\equiv\inf\left\{\sum_{R\in\mathcal{G}_n}\left(\sum_{P\in\mathcal{G}_m}\phi_{R\times P}v(P)\right)\mathcal{X}_{R'}(\mathbf{x}):\mathbf{x}\in R_1'\right\}$$

$$=\sum_{P\in\mathcal{G}_m}\phi_{R_1\times P}v(P).\tag{37.24}$$

Theorem 37.2.4 *(Fubini) Let $f\in\mathcal{R}(\mathbb{R}^{n+m})$ and suppose also that $f(\mathbf{x},\cdot)\in\mathcal{R}(\mathbb{R}^m)$ for each $\mathbf{x}$. Then*

$$\int_{\mathbb{R}^m}f(\cdot,\mathbf{y})\,dV_y\in\mathcal{R}(\mathbb{R}^n)\tag{37.25}$$

and

$$\int_{\mathbb{R}^{n+m}}f(\mathbf{z})\,dV=\int_{\mathbb{R}^n}\int_{\mathbb{R}^m}f(\mathbf{x},\mathbf{y})\,dV_y\,dV_x.\tag{37.26}$$

Proof: Let $\mathcal{G}$ be a grid such that $\mathcal{U}_{\mathcal{G}}(f)-\mathcal{L}_{\mathcal{G}}(f)<\varepsilon$ and let $\mathcal{G}_n$ and $\mathcal{G}_m$ be as defined above. Let

$$\phi(\mathbf{z})\equiv\sum_{Q\in\mathcal{G}}M_{Q'}(f)\,\mathcal{X}_{Q'}(\mathbf{z}),\ \psi(\mathbf{z})\equiv\sum_{Q\in\mathcal{G}}m_{Q'}(f)\,\mathcal{X}_{Q'}(\mathbf{z}).$$

By Corollary 37.1.4, and the observation that $M_{Q'}(f)\le M_Q(f)$ and $m_{Q'}(f)\ge m_Q(f)$,

$$\mathcal{U}_{\mathcal{G}}(f)\ge\int\phi\,dV,\ \mathcal{L}_{\mathcal{G}}(f)\le\int\psi\,dV.$$

Also $f(\mathbf{z})\in(\psi(\mathbf{z}),\phi(\mathbf{z}))$ for all $\mathbf{z}$. Thus from (37.23),

$$M_{R'}\left(\int_{\mathbb{R}^m}f(\cdot,\mathbf{y})\,dV_y\right)\le M_{R'}\left(\int_{\mathbb{R}^m}\phi(\cdot,\mathbf{y})\,dV_y\right)=\sum_{P\in\mathcal{G}_m}M_{R'\times P'}(f)\,v(P)$$

and from (37.24),

$$m_{R'}\left(\int_{\mathbb{R}^m}f(\cdot,\mathbf{y})\,dV_y\right)\ge m_{R'}\left(\int_{\mathbb{R}^m}\psi(\cdot,\mathbf{y})\,dV_y\right)=\sum_{P\in\mathcal{G}_m}m_{R'\times P'}(f)\,v(P).$$

Therefore,

$$\sum_{R\in\mathcal{G}_n}\left[M_{R'}\left(\int_{\mathbb{R}^m} f(\cdot,\mathbf{y})\,dV_y\right)-m_{R'}\left(\int_{\mathbb{R}^m} f(\cdot,\mathbf{y})\,dV_y\right)\right]v(R)\leq$$

$$\sum_{R\in\mathcal{G}_n}\sum_{P\in\mathcal{G}_m}\left[M_{R'\times P'}(f)-m_{R'\times P'}(f)\right]v(P)\,v(R)\leq\mathcal{U}_{\mathcal{G}}(f)-\mathcal{L}_{\mathcal{G}}(f)<\varepsilon.$$

This shows, from Lemma 37.1.5, that $\int_{\mathbb{R}^m} f(\cdot,\mathbf{y})\,dV_y\in\mathcal{R}(\mathbb{R}^n)$. It remains to verify (37.26). First note

$$\int_{\mathbb{R}^{n+m}} f(\mathbf{z})\,dV\in[\mathcal{L}_{\mathcal{G}}(f),\mathcal{U}_{\mathcal{G}}(f)\,].$$

Next, by Lemma 37.2.3,

$$\mathcal{L}_{\mathcal{G}}(f)\leq\int_{\mathbb{R}^{n+m}}\psi\,dV=\int_{\mathbb{R}^n}\int_{\mathbb{R}^m}\psi\,dV_y\,dV_x\leq\int_{\mathbb{R}^n}\int_{\mathbb{R}^m} f(\mathbf{x},\mathbf{y})\,dV_y\,dV_x$$

$$\leq\int_{\mathbb{R}^n}\int_{\mathbb{R}^m}\phi(\mathbf{x},\mathbf{y})\,dV_y\,dV_x=\int_{\mathbb{R}^{n+m}}\phi\,dV\leq\mathcal{U}_{\mathcal{G}}(f).$$

Therefore,

$$\left|\int_{\mathbb{R}^n}\int_{\mathbb{R}^m} f(\mathbf{x},\mathbf{y})\,dV_y\,dV_x-\int_{\mathbb{R}^{n+m}} f(\mathbf{z})\,dV\right|\leq\varepsilon$$

and since $\varepsilon>0$ is arbitrary, this proves Fubini's theorem[2].

Corollary 37.2.5 *Suppose E is a bounded contented set in $\mathbb{R}^n$ and let ϕ,ψ be continuous functions defined on E such that $\phi(\mathbf{x})\geq\psi(\mathbf{x})$. Also suppose f is a continuous bounded function defined on the set,*

$$P\equiv\{(\mathbf{x},y):\psi(\mathbf{x})\leq y\leq\phi(\mathbf{x})\},$$

It follows $f\mathcal{X}_P\in\mathcal{R}(\mathbb{R}^{n+1})$ and

$$\int_P f\,dV=\int_E\int_{\psi(\mathbf{x})}^{\phi(\mathbf{x})} f(\mathbf{x},y)\,dy\,dV_x.$$

Proof: Since f is continuous, there is no problem in writing $f(\mathbf{x},\cdot)\in\mathcal{R}(\mathbb{R}^1)$. Also, $f\mathcal{X}_P\in\mathcal{R}(\mathbb{R}^{n+1})$ because P is contented thanks to Corollary 37.1.19. Therefore, by Fubini's theorem

$$\begin{aligned}\int_P f\,dV&=\int_{\mathbb{R}^n}\int_{\mathbb{R}} f\mathcal{X}_P\,dy\,dV_x\\&=\int_E\int_{\psi(\mathbf{x})}^{\phi(\mathbf{x})} f(\mathbf{x},y)\,dy\,dV_x\end{aligned}$$

proving the corollary.

Other versions of this corollary are immediate and should be obvious whenever encountered.

[2]Actually, Fubini's theorem usually refers to a much more profound result in the theory of Lebesgue integration.

37.2.1 Some Observations

Some of the above material is very technical. This is because it gives complete answers to the fundamental questions on existence of the integral and related theoretical considerations. It was realized early in the twentieth century that these difficulties occur because from the point of view of mathematics, this is not the right way to define an integral! Better results are obtained much more easily using the Lebesgue integral. Many of the technicalities related to Jordan content disappear almost magically when the right integral is used. However, the Lebesgue integral is much more abstract than the Riemann integral and it is not traditional to consider it in a beginning calculus course. If you are interested in the fundamental properties of the integral and the theory behind it, you should abandon the Riemann integral which is an antiquated relic and begin to study the integral of the last century. One of the best introductions to it is in [23]. Another very good source is [12]. This advanced calculus text does everything in terms of the Lebesgue integral and never bothers to struggle with the inferior Riemann integral. A more general treatment is found in [18], [19], [24], and [20]. There is also a still more general integral called the generalized Riemann integral. A recent book on this subject is [5].

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 first proved by Gauss in 1797. The proof given here follows Rudin [23]. See also Hardy [15] for a similar proof, more discussion and references. The best proof is found in the theory of complex analysis. Recall De Moivre's theorem from trigonometry which is listed here for convenience.

Theorem A.0.6 *Let $r>0$ be given. Then if n is a positive integer,*

$$[r(\cos t+i\sin t)]^n = r^n(\cos nt+i\sin nt).$$

Recall that this theorem is the basis for proving the following corollary from trigonometry, also listed here for convenience.

Corollary A.0.7 *Let z be a non zero complex number and let k be a positive integer. Then there are always exactly k k^{th} roots of z in $\mathbb{C}$.*

Lemma A.0.8 *Let $a_k\in\mathbb{C}$ for $k=1,\cdots,n$ and let $p(z)\equiv\sum_{k=1}^n a_k z^k$. Then p is continuous.*

Proof:

$$|az^n-aw^n|\le|a||z-w|\left|z^{n-1}+z^{n-2}w+\cdots+w^{n-1}\right|.$$

Then for $|z-w|<1$, the triangle inequality implies $|w|<1+|z|$ and so if $|z-w|<1$,

$$|az^n-aw^n|\le|a||z-w|\,n(1+|z|)^n.$$

If $\varepsilon>0$ is given, let

$$\delta<\min\left(1,\frac{\varepsilon}{|a|\,n(1+|z|)^n}\right).$$

It follows from the above inequality that for $|z-w|<\delta$, $|az^n-aw^n|<\varepsilon$. The function of the lemma is just the sum of functions of this sort and so it follows that it is also continuous.

Theorem A.0.9 *(Fundamental theorem of Algebra) Let $p(z)$ be a nonconstant polynomial. Then there exists $z\in\mathbb{C}$ such that $p(z)=0$.*

Proof: Suppose not. Then

$$p(z) = \sum_{k=0}^{n} a_k z^k$$

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

$$\lim_{|z|\to\infty} |p(z)| = \infty. \tag{1.1}$$

Now let

$$\lambda \equiv \inf \{|p(z)| : z \in \mathbb{C}\}.$$

By (1.1), there exists an $R > 0$ such that if $|z| > R$, it follows that $|p(z)| > \lambda + 1$. Therefore,

$$\lambda \equiv \inf \{|p(z)| : z \in \mathbb{C}\} = \inf \{|p(z)| : |z| \leq R\}.$$

The set $\{z : |z| \leq R\}$ is a closed and bounded set and so this infimum is achieved at some point w with $|w| \leq R$. A contradiction is obtained if $|p(w)| = 0$ so assume $|p(w)| > 0$. 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 $|q(z)| \geq 1$ 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

$$w \neq 0, \theta = \frac{-\left|w^k\right| |c_k|}{w^k c_k}$$

and if $w = 0$, $\theta = 1$ will work. Now 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

[1] **Apostol, T. M.,** *Calculus second edition,* Wiley, 1967.

[2] **Apostol T.** *Calculus Volume II Second edition,* Wiley *1969.*

[3] **Apostol, T. M.,** *Mathematical Analysis,* Addison Wesley Publishing Co., 1974.

[4] **Baker, Roger**, *Linear Algebra*, Rinton Press 2001.

[5] **Bartle R.G.,** *A Modern Theory of Integration,* Grad. Studies in Math., Amer. Math. Society, Providence, RI, 2000.

[6] **Chahal J. S.** , *Historical Perspective of Mathematics* 2000 B.C. - 2000 A.D.

[7] **Davis H. and Snider A.,** *Vector Analysis* Wm. C. Brown 1995.

[8] **D'Angelo, J. and West D**. *Mathematical Thinking Problem Solving and Proofs*, Prentice Hall 1997.

[9] **Edwards C.H.** *Advanced Calculus of several Variables,* Dover 1994.

[10] **Euclid**, *The Thirteen Books of the Elements*, Dover, 1956.

[11] **Fitzpatrick P. M.,** *Advanced Calculus a course in Mathematical Analysis,* PWS Publishing Company 1996.

[12] **Fleming W**., *Functions of Several Variables,* Springer Verlag 1976.

[13] **Greenberg, M**. Advanced Engineering Mathematics, Second edition, Prentice Hall, 1998

[14] **Gurtin M.** *An introduction to continuum mechanics,* Academic press 1981.

[15] **Hardy G.,** *A Course Of Pure Mathematics, Tenth edition,* Cambridge University Press 1992.

[16] **Horn R. and Johnson C.** *matrix Analysis,* Cambridge University Press, 1985.

[17] **Karlin S. and Taylor H.** *A First Course in Stochastic Processes,* Academic Press, 1975.

[18] **Kuttler K. L.**, *Basic Analysis*, Rinton

[19] **Kuttler K.L.,** *Modern Analysis* CRC Press 1998.

[20] **Lang S.** *Real and Functional analysis* third edition Springer Verlag 1993. Press, 2001.

[21] **Nobel B. and Daniel J.** *Applied Linear Algebra,* Prentice Hall, 1977.

[22] **Rose, David, A.,** The College Math Journal, vol. 22, No.2 March 1991.

[23] **Rudin, W.,** *Principles of mathematical analysis,* McGraw Hill third edition 1976

[24] **Rudin W.,** *Real and Complex Analysis,* third edition, McGraw-Hill, 1987.

[25] **Salas S. and Hille E.,** *Calculus One and Several Variables,* Wiley 1990.

[26] **Sears and Zemansky,** *University Physics, Third edition, Addison Wesley 1963.*

[27] **Tierney John,** *Calculus and Analytic Geometry,* fourth edition, Allyn and Bacon, Boston, 1969.

[28] **Yosida K.,** *Functional Analysis, Springer Verlag, 1978.*

Index