Math 302 Lecture Notes

Kenneth Kuttler

October 6, 2006

Contents

Introduction

These are the lecture notes for my section of Math 302. They are pretty much in the order of the syllabus for the course. You don't need to read the starred sections and chapters and subsections. These are there to provide depth in the subject. To quote from the mission statement of BYU, " Depth comes when students realize the effect of rigorous, coherent, and progressively more sophisticated study. Depth helps students distinguish between what is fundamental and what is only peripheral; it requires focus, provides intense concentration. ..." To see clearly what is peripheral you need to read the fundamental and difficult concepts, most of which are presented in the starred sections. These are not always easy to read and I have indicated the most difficult with a picture of a dragon. Some are not much harder than what is presented in the course. A good example is the one which defines the derivative. If you don't learn this material, you will have trouble understanding many fundamental topics. Some which come to mind are basic continuum mechanics (The deformation gradient is a derivative.) and Newton's method for solving nonlinear systems of equations.(The entire method involves looking at the derivative and its inverse.) If you don't want to learn anything more than what you will be tested on, then you can omit these sections. This is up to you. It is your choice.

A word about notation might help. Most of the linear algebra works in any field. Examples are the rational numbers, the integers modulo a prime number, the complex numbers, or the real numbers. Therefore, I will often write $\mathbb{F}$ to denote this field. If you don't like this, just put in $\mathbb{R}$ and you will be fine. This is the main one of interest. However, I at least want you to realize that everything holds for the complex numbers in addition to the reals. In many applications this is essential so it does not hurt to begin to realize this. Also, I will write vectors in terms of bold letters. Thus $\mathbf{u}$ will denote a vector. Sometimes people write something like $\vec{u}$ to indicate a vector. However, the bold face is the usual notation so I am using this in these notes. On the board, I will likely write the other notation. The norm or length of a vector is often written as $||\mathbf{u}||$. I will usually write it as $|\mathbf{u}|$. This is standard notation also although most books use the double bar notation. The notation I am using emphasizes that the norm is just like the absolute value which is an important connection to make. It also seems less cluttered. You need to understand that either notation means the same thing.

For a more substantial treatment of certain topics, there is a complete calculus book on my web page. There are significant generalizations which unify all the notions of volume into one beautiful theory. I have not pursued this topic in these notes but it is in the calculus book. There are other things also, especially all the one variable theory if you need a review.

Part I

Vectors, Vector Products, Lines

Outcomes

Vectors in Two and Three Dimensions

A. Evaluate the distance between two points in 3-space.

B. Define vector and identify examples of vectors.

C. 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 combinations of unit coordinate vectors

D. Carry out the vector operations:

   (a) addition

   (b) scalar multiplication

   (c) magnitude (or norm or length)

   (d) normalize a vector (find the vector of unit length in the direction of a given vector)

E. Represent the operations of vector addition, scalar multiplication and norm geometrically.

F. Recall, apply and verify the basic properties of vector addition, scalar multiplication and norm.

G. Model and solve application problems using vectors.

Reading: Multivariable Calculus 1.1, Linear Algebra 1.1
Outcome Mapping:

A. 1,2,4

B. A1,A2

C. 8,9,11,13,14

D. 9,11,12,13

E. 8,10

F. 17,A3,A4

G. A5

Vector Products

A. Evaluate a dot product from the angle formula or the coordinate formula.

B. Interpret the dot product geometrically.

C. Evaluate the following using the dot product:

   i. the angle between two vectors.

ii. the magnitude of a vector.
iii. the projection of a vector onto another vector.
iv. the component of a vector in the direction of another vector.
v. the work done by a constant force on an object.

D. Evaluate a cross product from the angle formula or the coordinate formula.

E. Interpret the cross product geometrically.

F. Evaluate the following using the cross product:

i. the area of a parallelogram.
ii. the area or a triangle.
iii. physical quantities such as moment of force and angular velocity.

G. Find the volume of a parallelepiped using the scalar triple product.

H. Recall, apply and derive the algebraic properties of the dot and cross products.

Reading: Multivariable Calculus 1.2-3, Linear Algebra 1.2
Outcome Mapping:

A. 1,2bd,3,7

B. 3

C. 2egi

D. 2kmp,7dgh

E. 4

F. 5,15,B5

G. 6,B6

H. 8,17,B1,B2,B3,B4

Lines in Space

A. Represent a line in 3-space by a vector parameterization, a set of scalar parametric equations or using symmetric form.

B. Find a parameterization of a line given information about

(a) a point of the line and the direction of the line or
(b) two points contained in the line.
(c) the direction cosines of the line.

C. Determine the direction of a line given its parameterization.

D. Find the angle between two lines.

E. Determine a point of intersection between a line and a surface.

Reading: Multivariable Calculus 1.5, Linear Algebra 1.3
Outcome Mapping:

A. 3,4

B. 3,4

C. 1

D. 2

E. 11,14

Vectors And Points In $\mathbb{R}^n$ 5 Sept.

2.1 $\mathbb{R}^n$ Ordered $n-$ tuples

The notation, $\mathbb{R}^n$ refers to the collection of ordered lists of n real numbers. More precisely, consider the following definition.

Definition 2.1.1 *Define*

$$\mathbb{R}^n \equiv \{(x_1, \cdots, x_n) : x_j \in \mathbb{R} \text{ for } j = 1, \cdots, n\}.$$

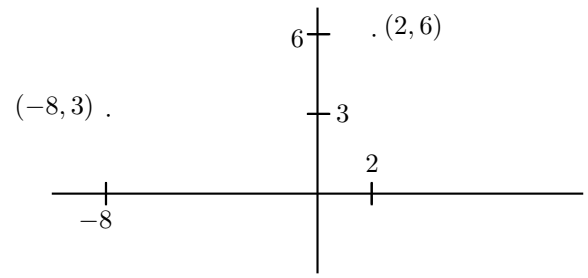
$(x_1, \cdots, x_n) = (y_1, \cdots, y_n)$ *if and only if for all* $j = 1, \cdots, n$, $x_j = y_j$. *When* $(x_1, \cdots, x_n) \in \mathbb{R}^n$, *it is conventional to denote* $(x_1, \cdots, x_n)$ *by the single bold face letter,* $\mathbf{x}$. *The numbers,* x_j *are called the* ***coordinates****. 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* ***coordinate axis****, the* x_i *axis for short. 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.

2.2 Vectors And Algebra In $\mathbb{R}^n$

There are two algebraic operations done with points of $\mathbb{R}^n$. One is addition and the other is multiplication by numbers, called scalars.

Definition 2.2.1 *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*

$$a\mathbf{x} = a(x_1, \cdots, x_n) \equiv (ax_1, \cdots, ax_n). \tag{2.1}$$

*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*

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1, \cdots, x_n) + (y_1, \cdots, y_n) \\ &\equiv (x_1 + y_1, \cdots, x_n + y_n) \end{aligned} \tag{2.2}$$

[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.

An element of $\mathbb{R}^n, \mathbf{x} \equiv (x_1, \cdots, x_n)$ *is often called a* ***vector.*** *The above definition is known as* ***vector addition.***

With this definition, the algebraic properties satisfy the conclusions of the following theorem.

Theorem 2.2.2 *For* $\mathbf{v}, \mathbf{w}$ *vectors in* $\mathbb{R}^n$ *and* α, β *scalars, (real numbers), the following hold.*

$$\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}, \tag{2.3}$$

the commutative law of addition,

$$(\mathbf{v} + \mathbf{w}) + \mathbf{z} = \mathbf{v} + (\mathbf{w} + \mathbf{z}), \tag{2.4}$$

the associative law for addition,

$$\mathbf{v} + \mathbf{0} = \mathbf{v}, \tag{2.5}$$

the existence of an additive identity,

$$\mathbf{v} + (-\mathbf{v}) = \mathbf{0}, \tag{2.6}$$

the existence of an additive inverse, Also

$$\alpha(\mathbf{v} + \mathbf{w}) = \alpha\mathbf{v} + \alpha\mathbf{w}, \tag{2.7}$$

$$(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}, \tag{2.8}$$

$$\alpha(\beta\mathbf{v}) = \alpha\beta(\mathbf{v}), \tag{2.9}$$

$$1\mathbf{v} = \mathbf{v}. \tag{2.10}$$

In the above $\mathbf{0} = (0, \cdots, 0)$.

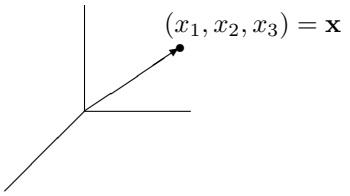
You should verify these properties all hold. For example, consider 2.7

$$\begin{aligned}
\alpha(\mathbf{v} + \mathbf{w}) &= \alpha(v_1 + w_1, \cdots, v_n + w_n) \\
&= (\alpha(v_1 + w_1), \cdots, \alpha(v_n + w_n)) \\
&= (\alpha v_1 + \alpha w_1, \cdots, \alpha v_n + \alpha w_n) \\
&= (\alpha v_1, \cdots, \alpha v_n) + (\alpha w_1, \cdots, \alpha w_n) \\
&= \alpha\mathbf{v} + \alpha\mathbf{w}.
\end{aligned}$$

As usual subtraction is defined as $\mathbf{x} - \mathbf{y} \equiv \mathbf{x} + (-\mathbf{y})$.

2.3 Geometric Meaning Of Vectors

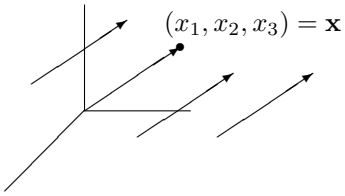
Definition 2.3.1 *Let* $\mathbf{x} = (x_1, \cdots, x_n)$ *be the coordinates of a point in* $\mathbb{R}^n$. *Imagine an arrow with its tail at* $\mathbf{0} = (0, \cdots, 0)$ *and its point at* $\mathbf{x}$ *as shown in the following picture in the case of* $\mathbb{R}^3$.



Then this arrow is called the ***position vector*** *of the point,* $\mathbf{x}$.

Thus every point determines such a vector and conversely, every such vector (arrow) which has its tail at **0** determines a point of $\mathbb{R}^n$, namely the point of $\mathbb{R}^n$ which coincides with the point of the vector.

Imagine taking the above position vector and moving it around, always keeping it pointing in the same direction as shown in the following picture.



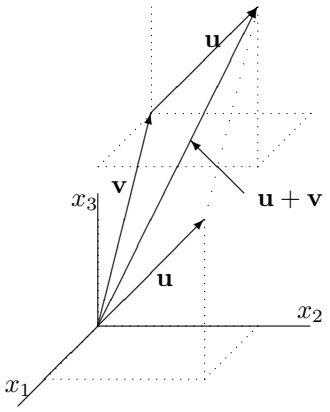
After moving it around, it is regarded as the same vector because it points in the same direction and has the same length.[2]Thus each of the arrows in the above picture is regarded as the same vector. The **components** of this vector are the numbers, $x_1, \cdots, x_n$. You should think of these numbers as directions for obtainng an arrow. Starting at some point, $(a_1, a_2, \cdots, a_n)$ in $\mathbb{R}^n$, you move to the point $(a_1 + x_1, \cdots, a_n)$ and from there to the point $(a_1 + x_1, a_2 + x_2, a_3 \cdots, a_n)$ and then to $(a_1 + x_1, a_2 + x_2, a_3 + x_3, \cdots, a_n)$ and continue this way until you obtain the point $(a_1 + x_1, a_2 + x_2, \cdots, a_n + x_n)$. The arrow having its tail at $(a_1, a_2, \cdots, a_n)$ and its point at $(a_1 + x_1, a_2 + x_2, \cdots, a_n + x_n)$ looks just like the arrow which has its tail at **0** and its point at $(x_1, \cdots, x_n)$ so it is regarded as the same vector.

2.4 Geometric Meaning Of Vector Addition

It was explained earlier that an element of $\mathbb{R}^n$ is an n tuple of numbers and it was also shown that this can be used to determine a point in three dimensional space in the case where $n = 3$ and in two dimensional space, in the case where $n = 2$. This point was specified relative to some coordinate axes.

Consider the case where $n = 3$ for now. If you draw an arrow from the point in three dimensional space determined by $(0, 0, 0)$ to the point (a, b, c) with its tail sitting at the point $(0, 0, 0)$ and its point at the point (a, b, c), this arrow is called the **position vector** of the point determined by $\mathbf{u} \equiv (a, b, c)$. One way to get to this point is to start at $(0, 0, 0)$ and move in the direction of the x_1 axis to $(a, 0, 0)$ and then in the direction of the x_2 axis to $(a, b, 0)$ and finally in the direction of the x_3 axis to (a, b, c). It is evident that the same arrow (vector) would result if you began at the point, $\mathbf{v} \equiv (d, e, f)$, moved in the direction of the x_1 axis to $(d + a, e, f)$, then in the direction of the x_2 axis to $(d + a, e + b, f)$, and finally in the x_3 direction to $(d + a, e + b, f + c)$ only this time, the arrow would have its tail sitting at the point determined by $\mathbf{v} \equiv (d, e, f)$ and its point at $(d + a, e + b, f + c)$. It is said to be the same arrow (vector) because it will point in the same direction and have the same length. It is like you took an actual arrow, the sort of thing you shoot with a bow, and moved it from one location to another keeping it pointing the same direction. This is illustrated in the following picture in which $\mathbf{v} + \mathbf{u}$ is illustrated. Note the parallelogram determined in the picture by the vectors $\mathbf{u}$ and $\mathbf{v}$.

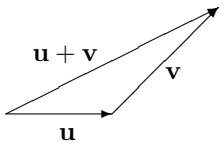
[2]I will discuss how to define length later. For now, it is only necessary to observe that the length should be defined in such a way that it does not change when such motion takes place.



Thus the geometric significance of $(d,e,f)+(a,b,c)=(d+a,e+b,f+c)$ is this. You start with the position vector of the point (d,e,f) and at its point, you place the vector determined by (a,b,c) with its tail at (d,e,f). Then the point of this last vector will be $(d+a,e+b,f+c)$. This is the geometric significance of vector addition. Also, as shown in the picture, $\mathbf{u}+\mathbf{v}$ is the directed diagonal of the parallelogram determined by the two vectors $\mathbf{u}$ and $\mathbf{v}$. A similar interpretation holds in $\mathbb{R}^n, n>3$ but I can't draw a picture in this case.

Since the convention is that identical arrows pointing in the same direction represent the same vector, the geometric significance of vector addition is as follows in any number of dimensions.

Procedure 2.4.1 *Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be two vectors. Slide* $\mathbf{v}$ *so that the tail of* $\mathbf{v}$ *is on the point of* $\mathbf{u}$. *Then draw the arrow which goes from the tail of* $\mathbf{u}$ *to the point of the slid vector,* $\mathbf{v}$. *This arrow represents the vector* $\mathbf{u}+\mathbf{v}$.



2.5 Distance Between Points In $\mathbb{R}^n$

How is distance between two points in $\mathbb{R}^n$ defined?

Definition 2.5.1 *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*

$$\text{distance between } \mathbf{x} \text{ and } \mathbf{y} \equiv |\mathbf{x}-\mathbf{y}| \equiv \left(\sum_{k=1}^{n} |x_k-y_k|^2\right)^{1/2}.$$

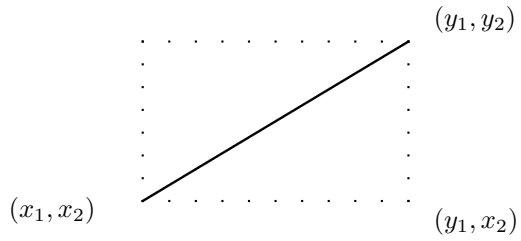
This is called the ***distance formula****. Thus* $|\mathbf{x}| \equiv |\mathbf{x}-\mathbf{0}|$. *The symbol,* $B(\mathbf{a},r)$ *is defined by*

$$B(\mathbf{a},r) \equiv \{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x}-\mathbf{a}| < r\}.$$

This is called an ***open ball*** *of radius* r *centered at* $\mathbf{a}$. *It means all 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. Often people use two lines to denote this distance, $||\mathbf{x}-\mathbf{y}||$. However, I want to emphasize this is really just like the absolute value. Also, the notation I am using is fairly standard.

Consider the following picture in the case that $n = 2$.



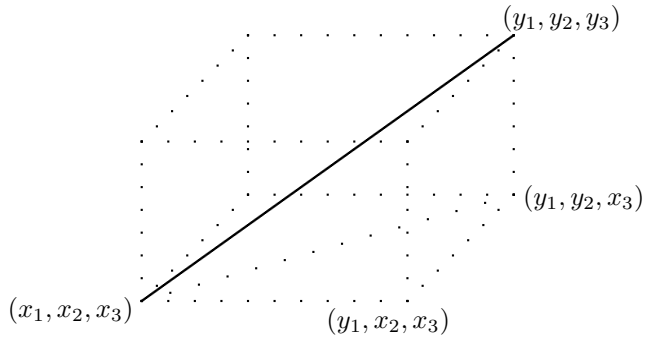
There are two points in the plane whose Cartesian coordinates are (x_1, x_2) and (y_1, y_2) respectively. Then the solid line joining these two points is the hypotenuse of a right triangle which is half of the rectangle shown in dotted lines. What is its length? Note the lengths of the sides of this triangle are $|y_1 - x_1|$ and $|y_2 - x_2|$. Therefore, the Pythagorean theorem implies the length of the hypotenuse equals

$$\left(|y_1 - x_1|^2 + |y_2 - x_2|^2\right)^{1/2} = \left((y_1 - x_1)^2 + (y_2 - x_2)^2\right)^{1/2}$$

which is just the formula for the distance given above. In other words, this distance defined above is the same as the distance of plane geometry in which the Pythagorean theorem holds.

Now suppose $n = 3$ and let (x_1, x_2, x_3) and (y_1, y_2, y_3) be two points in $\mathbb{R}^3$. Consider the following picture in which one of the solid lines joins the two points and a dotted line joins

the points (x_1, x_2, x_3) and (y_1, y_2, x_3).



By the Pythagorean theorem, the length of the dotted line joining (x_1, x_2, x_3) and (y_1, y_2, x_3) equals

$$\left((y_1 - x_1)^2 + (y_2 - x_2)^2\right)^{1/2}$$

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 no problem with the formula for distance in any number of dimensions. Here is an example.

Example 2.5.2 *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 2.5.3 *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

$$2x+2y+2z = -9. \tag{2.11}$$

Since these steps are reversible, the set of points which is at the same distance from the two given points consists of the points, (x, y, z) such that 2.11 holds.

There are certain properties of the distance which are obvious. Two of them which follow directly from the definition are

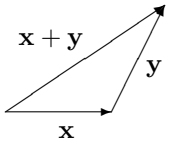
$$|\mathbf{x}-\mathbf{y}| = |\mathbf{y}-\mathbf{x}|,$$

$$|\mathbf{x}-\mathbf{y}| \geq 0 \text{ and equals 0 only if } \mathbf{y} = \mathbf{x}.$$

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. I will show you a proof of this pretty soon. This is usually stated as

$$|\mathbf{x}+\mathbf{y}| \leq |\mathbf{x}|+|\mathbf{y}|.$$

Here is a picture which illustrates the statement of this inequality in terms of geometry.



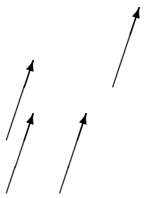
2.6 Geometric Meaning Of Scalar Multiplication

As discussed earlier, $\mathbf{x} = (x_1, x_2, x_3)$ determines a vector. You draw the line from $\mathbf{0}$ to $\mathbf{x}$ placing the point of the vector on $\mathbf{x}$. What is the length of this vector? The length of this vector is defined to equal $|\mathbf{x}|$ as in Definition 2.5.1. Thus the length of $\mathbf{x}$ equals $\sqrt{x_1^2+x_2^2+x_3^2}$. When you multiply $\mathbf{x}$ by a scalar, α, you get $(\alpha x_1, \alpha x_2, \alpha x_3)$ and the length of this vector is defined as $\sqrt{\left((\alpha x_1)^2+(\alpha x_2)^2+(\alpha x_3)^2\right)} = |\alpha|\sqrt{x_1^2+x_2^2+x_3^2}$. Thus the following holds.

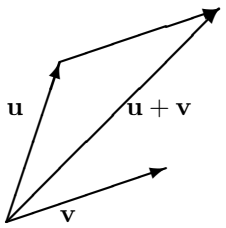
$$|\alpha\mathbf{x}| = |\alpha|\,|\mathbf{x}|.$$

In other words, multiplication by a scalar magnifies the length of the vector. What about the direction? You should convince yourself by drawing a picture that if α is negative, it causes the resulting vector to point in the opposite direction while if $\alpha > 0$ it preserves the direction the vector points.

You can think of vectors as quantities which have direction and magnitude, little arrows. Thus any two little arrows which have the same length and point in the same direction are considered to be the same vector even if their tails are at different points.

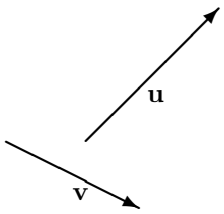


You can always slide such an arrow and place its tail at the origin. If the resulting point of the vector is (a,b,c), it is clear the length of the little arrow is $\sqrt{a^2+b^2+c^2}$. Geometrically, the way you add two geometric vectors is to place the tail of one on the point of the other and then to form the vector which results by starting with the tail of the first and ending with this point as illustrated in the following picture. Also when (a,b,c) is referred to as a vector, you mean any of the arrows which have the same direction and magnitude as the position vector of this point. Geometrically, for $\mathbf{u}=(u_1,u_2,u_3)$, $\alpha\mathbf{u}$ is any of the little arrows which have the same direction and magnitude as $(\alpha u_1,\alpha u_2,\alpha u_3)$.



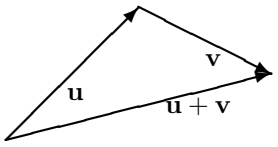
The following example is art which illustrates these definitions and conventions.

Exercise 2.6.1 *Here is a picture of two vectors,* $\mathbf{u}$ *and* $\mathbf{v}$.

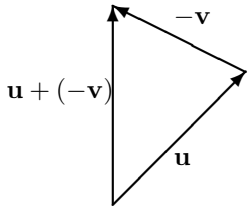


Sketch a picture of $\mathbf{u}+\mathbf{v}, \mathbf{u}-\mathbf{v}$, *and* $\mathbf{u}+2\mathbf{v}$.

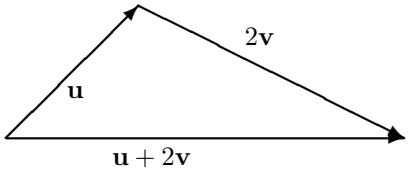
First here is a picture of $\mathbf{u}+\mathbf{v}$. You first draw $\mathbf{u}$ and then at the point of $\mathbf{u}$ you place the tail of $\mathbf{v}$ as shown. Then $\mathbf{u}+\mathbf{v}$ is the vector which results which is drawn in the following pretty picture.



Next consider $\mathbf{u}-\mathbf{v}$. This means $\mathbf{u}+(-\mathbf{v})$. From the above geometric description of vector addition, $-\mathbf{v}$ is the vector which has the same length but which points in the opposite direction to $\mathbf{v}$. Here is a picture.



Finally consider the vector $\mathbf{u}+2\mathbf{v}$. Here is a picture of this one also.



2.7 Unit Vectors

Let $\mathbf{v}$ be a vector,

$$\mathbf{v} = (v_1, \cdots, v_n).$$

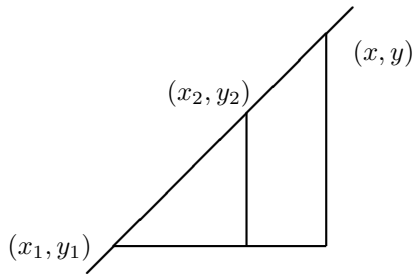
The **direction vector** for $\mathbf{v}$ is defined as $\mathbf{v}/|\mathbf{v}|$. This vector points in the same direction as $\mathbf{v}$ because it consists of the scalar, $1/|\mathbf{v}|$ times $\mathbf{v}$. This vector is called a **unit vector** because $|\mathbf{v}/|\mathbf{v}|| = |\mathbf{v}|/|\mathbf{v}| = 1$. That is, it has length equal to 1. The process of dividing a vector by its length is called **normalizing.** It provides you with a vector which has unit length and the same direction as the original vector.

2.8 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

$$y - y_1 = m\,(x - x_1) = \frac{y_2 - y_1}{x_2 - x_1}\,(x - x_1)\,. \tag{2.12}$$

Now consider points of the form

$$(x, y) = (x_1, y_1) + t\,(x_2 - x_1, y_2 - y_1)\,. \tag{2.13}$$

Do these points satisfy the above equation of the line? Is

$$y_1 + t\,(y_2 - y_1) - y_1 = \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x_1 + t\,(x_2 - x_1) - x_1)?$$

Yes, this is so. Both sides equal $t\,(y_2 - y_1)$. Conversely, if (x, y) is a point which satisfies the equation, 2.12 does there exist a value of t such that this point is of the form $(x_1, y_1) + t\,(x_2 - x_1, y_2 - y_1)$? If the point satisfies 2.12, it is of the form

$$\left(x, y_1 + \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x - x_1)\right).$$

Now let $t = \frac{x - x_1}{x_2 - x_1}$ so

$$x = x_1 + t\,(x_2 - x_1)\,.$$

Then in terms of t, the above reduces to

$$\left(x_1 + t\,(x_2 - x_1)\,, y_1 + \left(\frac{y_2 - y_1}{x_2 - x_1}\right) t\,(x_2 - x_1)\right) = (x_1, y_1) + t\,(x_2 - x_1, y_2 - y_1)\,.$$

It follows the set of points in $\mathbb{R}^2$ obtained from 2.12 and 2.13 are the same. The following is the definition of a line in $\mathbb{R}^n$.

Definition 2.8.1 *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. From now on, you should think of lines in this way. Forget about the stupid special case in $\mathbb{R}^2$ which you had drilled in to your head in high school. The concept of a line is really very simple and it holds in any number of dimensions, not just in two dimensions. It is given in the above definition.

Lemma 2.8.2 *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 2.8.3 *The vector $\mathbf{a}$ in the above lemma is called a **direction vector** for the line.*

Direction vectors are what it is all about, not slope. Slope is fine in two dimensions but we live in three dimensions. Slope is a trivial and stupid concept designed mainly to give children something to do in high school. The correct and worthwhile notion is that of direction vector. This is a new concept. Do not try to fit it in to the stuff you saw earlier. Do not try to put the new wine in the old bottles, to quote the scripture. It only creates confusion and you do not need that.

Example 2.8.4 *Find the line through $(1,2)$ and $(4,7)$.*

A vector equation of this line is $(x, y) = (1,2) + t(3,5)$. Now if you want to get the equation in the form you are used to seeing in high school,

$$x = 1 + 3t, y = 2 + 5t$$

Solving the first one for t, you get $t = (x-1)/3$ and now plugging this in to the second yields,

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

so $y - 2 = \frac{5}{3}(x-1)$ which is the usual point slope form for this line.

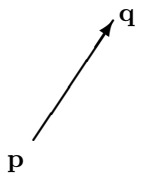
Now that you know about lines, it is possible to give a more analytical description of a vector as a directed line segment.

Definition 2.8.5 *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,*

$$\mathbf{x} = \mathbf{p} + t(\mathbf{q} - \mathbf{p}), \ t \in [0,1]$$

with the direction corresponding to increasing t. In the definition, when $t = 0$, the point $\mathbf{p}$ is obtained and as t increases other points on this line segment are obtained until when $t = 1$, you get the point, $\mathbf{q}$. This is what is meant by saying the direction corresponds 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 2.8.6 *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

$$(x,y,z) = (1,2,0) + t\,(1,-6,6)\,,\ t \in \mathbb{R}.$$

The vector $(1,-6,6)$ is obtained by $(2,-4,6) - (1,2,0)$ as indicated above.

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 is 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 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 2.8.7 *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

$$(x,y,z) = (1,2,0) + t\,(1,2,1)\,,\ t \in \mathbb{R}. \tag{2.14}$$

Sometimes people elect to write a line like the above in the form

$$x = 1 + t,\ y = 2 + 2t,\ z = t,\ t \in \mathbb{R}. \tag{2.15}$$

This is a set of scalar parametric equations which amounts to the same thing as 2.14.

There is one other form for a line which is sometimes considered useful. It is the so called symmetric form. Consider the line of 2.15. 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. Later, it will become clear that this expresses the line as the intersection of two planes but this is not important at this time.

Example 2.8.8 *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

$$(2,1,-3) + t\,(3,2,1) = (x,y,z)\,,\ t \in \mathbb{R}.$$

Example 2.8.9 *A relation such as $x^2 + y^2/4 + z^2/9 = 1$ describes something called a level surface. It consists of the points in $\mathbb{R}^n, (x,y,z)$ which satisfy the relation. Now here are parametric equations for a line: $x = t, y = 1+2t, z = 1-t$. Find where this line intersects the above level surface.*

This sort of problem is not hard if you don't panic. The points on the line are of the form $(t, 1+2t, 1-t)$ where $t \in \mathbb{R}$. All you have to do is to find values of t where this also satisfies the condition for being on the level surface. Thus you need t such that

$$(t)^2 + (1+2t)^2/4 + (1-t)^2/9 = 1.$$

This is just a quadratic equation. Expanding the left side yields $\frac{19}{9}t^2 + \frac{13}{36} + \frac{7}{9}t$ and so you have to solve the quadratic equation,

$$\frac{19}{9}t^2 + \frac{13}{36} + \frac{7}{9}t = 1$$

First simplify this to get the equation

$$76t^2 + 28t - 23 = 0.$$

Then the quadratic formula gives two solutions for $t, t = -\frac{7}{38} + \frac{9}{38}\sqrt{6}, -\frac{7}{38} - \frac{9}{38}\sqrt{6}$. Now you can obtain two points of intersection by plugging these values of t into the equation for the line. The two points are

$$\left(-\frac{7}{38} + \frac{9}{38}\sqrt{6}, \frac{12}{19} + \frac{9}{19}\sqrt{6}, \frac{45}{38} - \frac{9}{38}\sqrt{6}\right)$$

and

$$\left(-\frac{7}{38} - \frac{9}{38}\sqrt{6}, \frac{12}{19} - \frac{9}{19}\sqrt{6}, \frac{45}{38} + \frac{9}{38}\sqrt{6}\right).$$

Possibly you would not have guessed these points. You likely would not have found them by drawing a picture either.

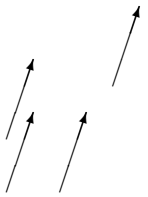
2.9 Vectors And Physics

Suppose you push on something. What is important? There are really two things which are important, how hard you push and the direction you push. This illustrates the concept of force. Also you can see that the concept of a geometric vector is useful for defining something like force.

Definition 2.9.1 ***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 2.8.5 it follows that $\overrightarrow{\mathbf{pq}}$ consists of the points of the form

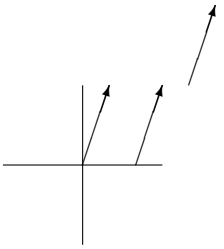
$$\mathbf{p} + t\,(\mathbf{q} - \mathbf{p})$$

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 points of $\mathbb{R}^n$.

Definition 2.9.2 *Let $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$. The* ***position vector*** *of this point is the vector whose point is at $\mathbf{x}$ and whose tail is at the origin, $(0, \cdots, 0)$. If $\mathbf{x} = (x_1, \cdots, x_n)$ is called a vector, the vector which is meant is this position vector just described. Another term associated with this is* ***standard position.*** *A vector is in standard position if the tail is placed at the origin.*

It is customary to identify the point in $\mathbb{R}^n$ with its position vector.

The magnitude of a vector determined by a directed line segment $\overrightarrow{\mathbf{pq}}$ 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 $\left(\sum_{k=1}^{n}v_k^2\right)^{1/2}=|\mathbf{v}|$.

Example 2.9.3 *Consider the vector,* $\mathbf{v}\equiv(1,2,3)$ *in* $\mathbb{R}^n$. *Find* $|\mathbf{v}|$.

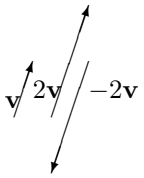
First, the vector is the directed line segment (arrow) which has its base at $\mathbf{0}\equiv(0,0,0)$ and its point at $(1,2,3)$. Therefore,

$$|\mathbf{v}|=\sqrt{1^2+2^2+3^2}=\sqrt{14}.$$

What is the geometric significance of scalar multiplication? As noted earlier, if a vector, $\mathbf{v}$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 $r\mathbf{v}$?

$$\begin{aligned}|r\mathbf{v}|&=\left(\sum_{k=1}^{n}(ra_i)^2\right)^{1/2}=\left(\sum_{k=1}^{n}r^2(a_i)^2\right)^{1/2}\\&=\left(r^2\right)^{1/2}\left(\sum_{k=1}^{n}a_i^2\right)^{1/2}=|r|\,|\mathbf{v}|.\end{aligned}$$

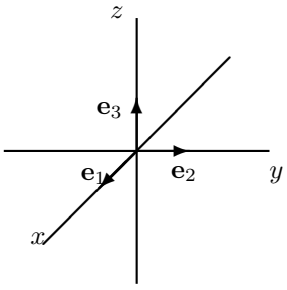
Thus the magnitude of $r\mathbf{v}$ equals $|r|$ times the magnitude of $\mathbf{v}$. If r is positive, then the vector represented by $r\mathbf{v}$ has the same direction as the vector, $\mathbf{v}$ because multiplying by the scalar, r, only has the effect of scaling all the distances. Thus the unit 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.

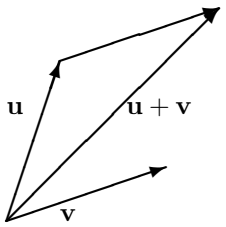
$$\begin{aligned}\mathbf{a} + \mathbf{b} &= (a_1 + b_1, \cdots, a_n + b_n)\,. \\ &= \sum_{i=1}^{n} (a_i + b_i)\,\mathbf{e}_i.\end{aligned}$$

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 2.9.4 *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*

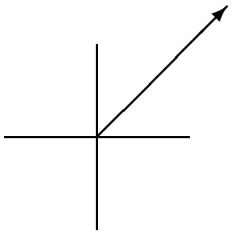
and the third exerts a force of $5\mathbf{i}-\mathbf{j}+2\mathbf{k}$. *Newtons. Find the total force in the direction of* $\mathbf{i}$.

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 2.9.5 *An airplane flies North East at 100 miles per hour. Write this as a vector.*

A picture of this situation follows.



The vector has length 100. Now using that vector as the hypotenuse of a right triangle having equal sides, the sides should be each of length $100/\sqrt{2}$. Therefore, the vector would be $(100/\sqrt{2})\mathbf{i}+(100/\sqrt{2})\mathbf{j}$.

This example also motivates the concept of **velocity**.

Definition 2.9.6

The ***speed*** *of an object is a measure of how fast it is going. It is measured in units of length per unit time. For example, miles per hour, kilometers per minute, feet per second. The* ***velocity*** *is a vector having the speed as the magnitude but also specifing the direction.*

Thus the velocity vector in the above example is $(100/\sqrt{2})\mathbf{i}+(100/\sqrt{2})\mathbf{j}$.

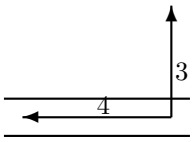
Example 2.9.7 *The velocity of an airplane is* $100\mathbf{i}+\mathbf{j}+\mathbf{k}$ *measured in kilometers per hour and at a certain instant of time its position is* $(1,2,1)$. *Here imagine a Cartesian coordinate system in which the third component is altitude and the first and second components are measured on a line from West to East and a line from South to North. Find the position of this airplane one minute later.*

Consider the vector $(1,2,1)$, is the initial position vector of the airplane. As it moves, the position vector changes. After one minute the airplane has moved in the $\mathbf{i}$ direction a distance of $100\times\frac{1}{60}=\frac{5}{3}$ kilometer. In the $\mathbf{j}$ direction it has moved $\frac{1}{60}$ kilometer during this same time, while it moves $\frac{1}{60}$ kilometer in the $\mathbf{k}$ direction. Therefore, the new displacement vector for the airplane is

$$(1,2,1)+\left(\frac{5}{3},\frac{1}{60},\frac{1}{60}\right)=\left(\frac{8}{3},\frac{121}{60},\frac{121}{60}\right)$$

Example 2.9.8 *A certain river is one half mile wide with a current flowing at 4 miles per hour from East to West. A man swims directly toward the opposite shore from the South bank of the river at a speed of 3 miles per hour. How far down the river does he find himself when he has swam across? How far does he end up swimming?*

Consider the following picture.



You should write these vectors in terms of components. The velocity of the swimmer in still water would be $3\mathbf{j}$ while the velocity of the river would be $-4\mathbf{i}$. Therefore, the velocity of the swimmer is $-4\mathbf{i}+3\mathbf{j}$. Since the component of velocity in the direction across the river is 3, it follows the trip takes $1/6$ hour or 10 minutes. The speed at which he travels is $\sqrt{4^2+3^2}=5$ miles per hour and so he travels $5\times\frac{1}{6}=\frac{5}{6}$ miles. Now to find the distance downstream he finds himself, note that if x is this distance, x and $1/2$ are two legs of a right triangle whose hypotenuse equals $5/6$ miles. Therefore, by the Pythagorean theorem the distance downstream is

$$\sqrt{(5/6)^2-(1/2)^2}=\frac{2}{3}\text{ miles.}$$

Vector Products

3.1 The Dot Product 6 Sept.

Quiz

1. Given two points in $\mathbb{R}^3, (x_1, y_1, z_1)$ and (x_2, y_2, z_2), show the point
$$\left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}, \frac{z_1+z_2}{2}\right)$$
is on the line between these two points and is the same distance from each of them.
2. Given the two points in $\mathbb{R}^3, (x_1, y_1, z_1)$ and (x_2, y_2, z_2), describe the set of all points which are equidistant from these two points in terms of a simple equation.
3. An airplane heads due north at a speed of 120 miles per hour. The wind is blowing north east at a speed of 30 miles per hour. Find the resulting speed of the airplane.

3.1.1 Definition In terms Of Coordinates

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 3.1.1 *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 3.1.2 *The dot product satisfies the following properties.*
$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} \tag{3.1}$$
$$\mathbf{a} \cdot \mathbf{a} \geq 0 \text{ and equals zero if and only if } \mathbf{a} = \mathbf{0} \tag{3.2}$$
$$(\alpha\mathbf{a} + \beta\mathbf{b}) \cdot \mathbf{c} = \alpha(\mathbf{a} \cdot \mathbf{c}) + \beta(\mathbf{b} \cdot \mathbf{c}) \tag{3.3}$$
$$\mathbf{c} \cdot (\alpha\mathbf{a} + \beta\mathbf{b}) = \alpha(\mathbf{c} \cdot \mathbf{a}) + \beta(\mathbf{c} \cdot \mathbf{b}) \tag{3.4}$$
$$|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a} \tag{3.5}$$

You should verify these properties. Also be sure you understand that 3.4 follows from the first three and is therefore redundant. It is listed here for the sake of convenience.

Example 3.1.3 *Find* $(1,2,0,-1)\cdot(0,1,2,3)$.

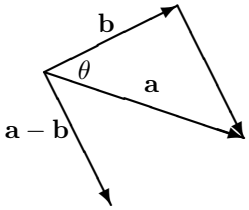
This equals $0+2+0+-3=-1$.

Example 3.1.4 *Find the magnitude of* $\mathbf{a}=(2,1,4,2)$. *That is, find* $|\mathbf{a}|$.

This is $\sqrt{(2,1,4,2)\cdot(2,1,4,2)}=5$.

3.1.2 The Geometric Meaning Of The Dot Product, The Included Angle

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,

$$\mathbf{a}\cdot\mathbf{b}=|\mathbf{a}||\mathbf{b}|\cos\theta. \tag{3.6}$$

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 3.1.5 *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 3.6 the cosine of the included angle equals

$$\cos\theta=\frac{9}{\sqrt{26}\sqrt{6}}=.72058$$

Now the cosine is known, the angle can be determined by solving the equation, $\cos\theta=.72058$. This will involve using a calculator or a table of trigonometric functions. The answer

is $\theta = .766\,16$ 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 3.1.6 *Find the magnitude of the vector* $2\mathbf{i} + 3\mathbf{j} - \mathbf{k}$

As discussed above, this has magnitude equal to

$$\sqrt{(2\mathbf{i} + 3\mathbf{j} - \mathbf{k}) \cdot (2\mathbf{i} + 3\mathbf{j} - \mathbf{k})} = \sqrt{4 + 9 + 1} = \sqrt{14}.$$

Example 3.1.7 *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 3.6

$$\mathbf{u} \cdot \mathbf{v} = 3 \times 4 \times \cos(60°) = 3 \times 4 \times 1/2 = 6.$$

Observation 3.1.8 *Two vectors are said to be* ***perpendicular*** *or* ***orthogonal*** *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 3.1.9 *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 3.1.10 *When two lines intersect, the angle between the two lines is the smaller of the two angles determined.*

Example 3.1.11 *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

$$\cos\theta = \frac{-6}{14} = \frac{-3}{7}. \tag{3.7}$$

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.

3.1.3 The Cauchy Schwarz Inequality

The dot product satisfies a fundamental inequality known as the **Cauchy Schwarz inequality.**

Theorem 3.1.12 *The dot product satisfies the inequality*

$$|\mathbf{a}\cdot\mathbf{b}| \leq |\mathbf{a}|\,|\mathbf{b}|\,. \tag{3.8}$$

Furthermore equality is obtained if and only if one of $\mathbf{a}$ *or* $\mathbf{b}$ *is a scalar multiple of the other.*

Geometric Proof: From the geometric description of the dot product,

$$|\mathbf{a}\cdot\mathbf{b}| = ||\mathbf{a}|\,|\mathbf{b}|\cos\theta| \leq |\mathbf{a}|\,|\mathbf{b}|$$

because $\cos\theta$ is a number between -1 and 1. Equality occurs if and only if $\cos\theta = \pm 1$. This corresponds to $\mathbf{b}$ being a scalar multiple of $\mathbf{a}$. If $\cos\theta = -1$, then $\mathbf{b}$ points in the opposite direction to $\mathbf{a}$ and if $\cos\theta = 1$ then $\mathbf{b}$ points in the same direction as $\mathbf{a}$.

The Cauchy Schwarz inequality is important in many contexts other than vectors in $\mathbb{R}^n$. What follows is a vastly superior algebraic proof. In general it is this way. Algebraic methods are nearly always to be preferred to geometric reasoning.

Algebraic Proof: First note that if $\mathbf{b} = \mathbf{0}$ both sides of 3.8 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 3.2, $f(t) \geq 0$ for all $t \in \mathbb{R}$. Also from 3.3,3.4,3.1, and 3.5

$$\begin{aligned} f(t) &= \mathbf{a}\cdot(\mathbf{a} + t\mathbf{b}) + t\mathbf{b}\cdot(\mathbf{a} + t\mathbf{b}) \\ &= \mathbf{a}\cdot\mathbf{a} + t(\mathbf{a}\cdot\mathbf{b}) + t\mathbf{b}\cdot\mathbf{a} + t^2\mathbf{b}\cdot\mathbf{b} \\ &= |\mathbf{a}|^2 + 2t(\mathbf{a}\cdot\mathbf{b}) + |\mathbf{b}|^2 t^2. \end{aligned}$$

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 $t = -\left(\mathbf{a}\cdot\mathbf{b}/|\mathbf{b}|^2\right)$, the value of t which yields the minimum value of f, which implies

$$\frac{|\mathbf{a}|^2}{|\mathbf{b}|^2} - \left(\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{b}|^2}\right)^2 \geq 0. \tag{3.9}$$

Multiplying both sides by $|\mathbf{b}|^4$,

$$|\mathbf{a}|^2\,|\mathbf{b}|^2 \geq (\mathbf{a}\cdot\mathbf{b})^2$$

which yields 3.8. If equality in the Cauchy Schwarz inequality holds, then the minimum value of $f(t)$ is zero and so for some t, $(\mathbf{a}+t\mathbf{b})\cdot(\mathbf{a}+t\mathbf{b}) = |\mathbf{a}+t\mathbf{b}| = 0$ so that $\mathbf{a} = -t\mathbf{b}$. This proves the theorem.

Another Algebraic Proof: Let $f(t)$ be given as above. Thus as above $f(t) \geq 0$ for all $t \in \mathbb{R}$. Thus as above,

$$f(t) = |\mathbf{a}|^2 + 2t(\mathbf{a}\cdot\mathbf{b}) + |\mathbf{b}|^2 t^2 \geq 0$$

The graph of $f(t)$ is a parabola which must open up and cannot cross the t axis. Thus $f(t) = 0$ has either one real root or no real roots. Now recall the quadratic formula. This condition implies the stuff under the square root sign in the quadratic formula must be nonpositive. Applied to this function of t it says

$$4(\mathbf{a}\cdot\mathbf{b})^2 - 4|\mathbf{a}|^2|\mathbf{b}|^2 \leq 0$$

which is just the Cauchy Schwarz inequality. As before, equality in this inequality implies f has one real zero. Thus the minimum value of f is 0. This means $\mathbf{a}+t\mathbf{b} = \mathbf{0}$ for some t and so one vector is a multiple of the other. This proves the theorem.

You should note that the algebraic arguments were based only on the properties of the dot product listed in 3.1 - 3.5. This means that whenever something satisfies these properties, the Cauchy Schwartz inequality holds. There are many other instances of these properties besides vectors in $\mathbb{R}^n$.

3.1.4 The Triangle Inequality

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 3.1.13 *(Triangle inequality) For* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$

$$|\mathbf{a}+\mathbf{b}| \leq |\mathbf{a}| + |\mathbf{b}| \tag{3.10}$$

and equality holds if and only if one of the vectors is a nonnegative scalar multiple of the other. Also

$$||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a}-\mathbf{b}| \tag{3.11}$$

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 3.10.

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 3.1.12 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,

$$|\mathbf{a}| - |\mathbf{b}| \leq |\mathbf{a} - \mathbf{b}| \tag{3.12}$$

Similarly,

$$|\mathbf{b}| - |\mathbf{a}| \leq |\mathbf{b} - \mathbf{a}| = |\mathbf{a} - \mathbf{b}| . \tag{3.13}$$

It follows from 3.12 and 3.13 that 3.11 holds. This is because $||\mathbf{a}| - |\mathbf{b}||$ equals the left side of either 3.12 or 3.13 and either way, $||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a} - \mathbf{b}|$. This proves the theorem.

3.1.5 Direction Cosines Of A Line

Now that the dot product and distance has been defined, it is possible to mention some archaic terminology which is sometimes found.

Suppose $\mathbf{x} = \mathbf{a} + t\mathbf{b}$ is a vector equation for a line in $\mathbb{R}^n$ where, as explained before, the vector, $\mathbf{b}$ is called a direction vector. When $\mathbf{b}$ is a unit vector ($|\mathbf{b}| = 1$), the components of $\mathbf{b}$ are called **direction cosines.** Say $\mathbf{b} = (b_1, \cdots, b_n)$. Thus, from the definition of the dot product, $b_k = \mathbf{b} \cdot \mathbf{e}_k$ where $\mathbf{e}_k$ is the unit vector for the k^{th} coordinate axis consisting of all zeros except for a 1 in the k^{th} slot. So why in the world do people call these "direction cosines"? It is because the cosine of the angle, θ_k between the unit vector $\mathbf{b}$ and the vector, $\mathbf{e}_k$ is given by

$$\cos\theta_k \equiv \frac{\mathbf{b} \cdot \mathbf{e}_k}{|\mathbf{b}| |\mathbf{e}_k|} = \mathbf{b} \cdot \mathbf{e}_k = b_k.$$

There, isn't that interesting? Now you know why these are called direction cosines. So what importance does it have? If someone gives you the "direction cosines" of a line, they are just using jargon to identify the components of a unit vector which serves as a direction vector for the line.

Example 3.1.14 *A line, l in $\mathbb{R}^3$ contains the point $(1, 2, 3)$ and letting θ_k be the angle between a direction vector and $\mathbf{e}_k$,*

$$\cos(\theta_1) = \frac{1}{\sqrt{5}}, \cos(\theta_2) = \frac{1}{\sqrt{5}}, \cos(\theta_3) = -\frac{\sqrt{15}}{5},$$

find a vector equation for the line.

The information in this example is nothing more than a jargon laden statement that a direction vector for the line is

$$\left(\frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}, \frac{-\sqrt{15}}{5} \right).$$

Therefore, a vector equation for the line is

$$(x,y,z) = (1,2,3) + t\left(\frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}, \frac{-\sqrt{15}}{5}\right).$$

Of course if you like to wallow in terminology, you could also say parametric equations for this line are

$$x = 1 + t\frac{1}{\sqrt{5}}, y = 2 + t\frac{1}{\sqrt{5}}, z = 3 + t\left(\frac{-\sqrt{15}}{5}\right).$$

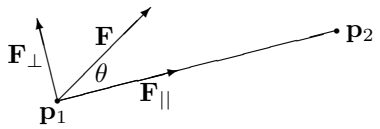
Symmetric equations for the line are obtained by solving for the parameter. Thus symmetric equations for the line are

$$\sqrt{5}\,(x-1) = \sqrt{5}\,(y-2) = -\frac{5}{\sqrt{15}}\,(z-3).$$

Isn't this exciting? No doubt there are other monumental trivialities and stupid observations which could be drawn. The fundamental and significant ingredients of a line are the direction vector and a point on the line. These are the most important things to understand.

3.1.6 Work And Projections

An important application of the dot product is the concept of work. The physical concept of work does not in any way correspond to the notion of work employed in ordinary conversation. For example, if you were to slide a 150 pound weight off a table which is three feet high and shuffle along the floor for 50 yards, sweating profusely and exerting all your strength to keep the weight from falling on your feet, keeping the height always three feet and then deposit this weight on another three foot high table, the physical concept of work would indicate that the force exerted by your arms did no work during this project even though the muscles in your hands and arms would likely be very tired. The reason for such an unusual definition is that even though your arms exerted considerable force on the weight, enough to keep it from falling, the direction of motion was at right angles to the force they exerted. The only part of a force which does work in the sense of physics is the component of the force in the direction of motion (This is made more precise below.). The work is defined to be the magnitude of the component of this force times the distance over which it acts in the case where this component of force points in the direction of motion and (-1) times the magnitude of this component times the distance in case the force tends to impede the motion. Thus the work done by a force on an object as the object moves from one point to another is a measure of the extent to which the force contributes to the motion. This is illustrated in the following picture in the case where the given force contributes to the motion.



In this picture the force, $\mathbf{F}$ is applied to an object which moves on the straight line from $\mathbf{p}_1$ to $\mathbf{p}_2$. There are two vectors shown, $\mathbf{F}_{||}$ 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 $\mathbf{p}_2$. $\mathbf{F}_{||}$ is called the **projection of the force** in the direction of motion. 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 3.1.15 *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 3.1.16 *Let $\mathbf{F}$ and $\mathbf{D}$ be nonzero vectors. Then there exist unique vectors $\mathbf{F}_{||}$ and $\mathbf{F}_\perp$ such that*

$$\mathbf{F} = \mathbf{F}_{||} + \mathbf{F}_\perp \tag{3.14}$$

where $\mathbf{F}_{||}$ is a scalar multiple of $\mathbf{D}$, also referred to as

$$\mathrm{proj}_{\mathbf{D}}(\mathbf{F}),$$

*and $\mathbf{F}_\perp\cdot\mathbf{D}=0$. The vector $\mathrm{proj}_{\mathbf{D}}(\mathbf{F})$ is called the **projection** of $\mathbf{F}$ onto $\mathbf{D}$.*

Proof: Suppose 3.14 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.

Definition 3.1.17 *The component of the vector* $\mathbf{F}$ *in the direction,* $\mathbf{D}$ *equals the scalar*

$$\frac{\mathbf{F}\cdot\mathbf{D}}{|\mathbf{D}|}.$$

Thus

$$\mathrm{proj}_{\mathbf{D}}(\mathbf{F}) = \frac{\mathbf{F}\cdot\mathbf{D}}{|\mathbf{D}|}\frac{\mathbf{D}}{|\mathbf{D}|}.$$

In words, the projection of $\mathbf{F}$ *on* $\mathbf{D}$ *equals the component of* $\mathbf{F}$ *in the direction* $\mathbf{D}$ *times the unit vector in the direction of* $\mathbf{D}$.

Example 3.1.18 *Let* $\mathbf{F} = (1,2,3)$. *Find the projection of* $\mathbf{F}$ *on* $\mathbf{D} = (2,1,1)$ *and also find the component of* $\mathbf{F}$ *in the direction,* $\mathbf{D}$.

$$\begin{aligned}\mathrm{proj}_{\mathbf{D}}(\mathbf{F}) &= \frac{\mathbf{F}\cdot\mathbf{D}}{|\mathbf{D}|}\frac{\mathbf{D}}{|\mathbf{D}|}\\ &= \frac{2+2+3}{\sqrt{4+1+1}}\frac{(2,1,1)}{\sqrt{4+1+1}}\\ &= \frac{7}{6}(2,1,1) = \left(\frac{7}{3},\frac{7}{6},\frac{7}{6}\right)\end{aligned}$$

and the component of $\mathbf{F}$ in the direction of $\mathbf{D}$ is

$$\frac{2+2+3}{\sqrt{4+1+1}} = \frac{7}{6}\sqrt{6}.$$

Example 3.1.19 *Let* $\mathbf{F} = 2\mathbf{i}+7\mathbf{j}-3\mathbf{k}$ *Newtons. Find the work done by this force in moving from the point* $(1,2,3)$ *to the point* $(-9,-3,4)$ *along the straight line segment joining these points where distances are measured in meters.*

According to the definition, this work is

$$\begin{aligned}(2\mathbf{i}+7\mathbf{j}-3\mathbf{k})\cdot(-10\mathbf{i}-5\mathbf{j}+\mathbf{k}) &= -20+(-35)+(-3)\\ &= -58 \text{ Newton meters.}\end{aligned}$$

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 3.1.20 *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 3.1.16, this is just

$$\begin{aligned}&\frac{1}{4+9+16}(\mathbf{i}-2\mathbf{j}+\mathbf{k})\cdot(2\mathbf{i}+3\mathbf{j}-4\mathbf{k})(2\mathbf{i}+3\mathbf{j}-4\mathbf{k})\\ =\ &\frac{-8}{29}(2\mathbf{i}+3\mathbf{j}-4\mathbf{k}) = -\frac{16}{29}\mathbf{i}-\frac{24}{29}\mathbf{j}+\frac{32}{29}\mathbf{k}.\end{aligned}$$

Example 3.1.21 *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,

$$|\mathbf{b}_\perp| = |\mathbf{b}| \sin\theta.$$

3.2 The Cross Product 7 Sept.

Quiz

1. Find the cosine of the angle between the two vectors $(1,2,0)$ and $(2,0,1)$.
2. Suppose $\mathbf{u}, \mathbf{v}$ are vectors. Show the parallelogram identity.

   $$|\mathbf{u}+\mathbf{v}|^2 + |\mathbf{u}-\mathbf{v}|^2 = 2|\mathbf{u}|^2 + 2|\mathbf{v}|^2$$

   You must show this in any dimension, not just in two or three dimensions.
3. Find the projection of the vector $(1,2,3)$ onto the vector $(2,3,1)$.
4. Given two vectors, $\mathbf{u}, \mathbf{v}$ in $\mathbb{R}^n$, show using the properties of the dot product alone that

   $$\mathbf{u} - \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{v}|^2}\mathbf{v}$$

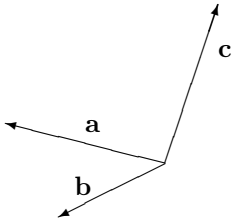
   is perpendicular to $\mathbf{v}$.
5. $\mathbf{x} = \mathbf{u} + t\mathbf{v}$ for $t \in \mathbb{R}$ is a line. Suppose $\mathbf{z}$ is a point in $\mathbb{R}^n$. Find a formula for the distance between $\mathbf{z}$ and this line.

3.2.1 The Geometric Description Of The Cross Product In Terms Of The Included Angle

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 3.2.1 *Three vectors,* $\mathbf{a}, \mathbf{b}, \mathbf{c}$ *form a right handed system if when you extend the fingers of your right hand along the vector,* $\mathbf{a}$ *and close them in the direction of* $\mathbf{b}$*, the thumb points roughly in the direction of* $\mathbf{c}$*.*

For an example of a right handed system of vectors, see the following picture.



In this picture the vector $\mathbf{c}$ points upwards from the plane determined by the other two vectors. You should consider how a right hand system would differ from a left hand system. Try using your left hand and you will see that the vector, $\mathbf{c}$ would need to point in the opposite direction as it would for a right hand system.

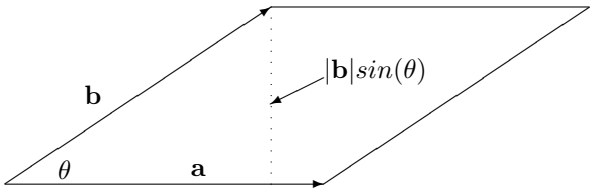
From now on, the vectors, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ will **always** form a right handed system. To repeat, if you extend the fingers of your right hand along $\mathbf{i}$ and close them in the direction $\mathbf{j}$, the thumb points in the direction of $\mathbf{k}$.

The following is the geometric description of the cross product. It gives both the direction and the magnitude and therefore specifies the vector.

Definition 3.2.2 *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two vectors in* $\mathbb{R}^3$*. Then* $\mathbf{a} \times \mathbf{b}$ *is defined by the following two rules.*

1. $|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}|\,|\mathbf{b}| \sin\theta$ *where* 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.

$$\mathbf{a} \times \mathbf{b} = -(\mathbf{b} \times \mathbf{a}) \ , \ \mathbf{a} \times \mathbf{a} = \mathbf{0}, \tag{3.15}$$

For α a scalar,

$$(\alpha\mathbf{a}) \times \mathbf{b} = \alpha(\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (\alpha\mathbf{b}), \tag{3.16}$$

For $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ vectors, one obtains the distributive laws,

$$\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}, \tag{3.17}$$

$$(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}. \tag{3.18}$$

Formula 3.15 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 3.16 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 3.16. 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 3.15,

$$\begin{aligned}(\mathbf{b} + \mathbf{c}) \times \mathbf{a} &= -\mathbf{a} \times (\mathbf{b} + \mathbf{c}) \\ &= -(\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}) \\ &= \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}.\end{aligned}$$

A proof of the distributive law is given in a later section for those who are interested.

3.2.2 The Coordinate Description Of The Cross Product

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 3.2.3 *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*

$$\begin{aligned}\mathbf{a} \times \mathbf{b} = &(a_2b_3 - a_3b_2)\,\mathbf{i} + (a_3b_1 - a_1b_3)\,\mathbf{j} + \\ &+ (a_1b_2 - a_2b_1)\,\mathbf{k}.\end{aligned} \tag{3.19}$$

Proof: From the above table and the properties of the cross product listed,

$$(a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}) \times (b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}) =$$

$$a_1b_2\mathbf{i} \times \mathbf{j} + a_1b_3\mathbf{i} \times \mathbf{k} + a_2b_1\mathbf{j} \times \mathbf{i} + a_2b_3\mathbf{j} \times \mathbf{k} +$$

$$+a_3b_1\mathbf{k} \times \mathbf{i} + a_3b_2\mathbf{k} \times \mathbf{j}$$

$$= a_1b_2\mathbf{k} - a_1b_3\mathbf{j} - a_2b_1\mathbf{k} + a_2b_3\mathbf{i} + a_3b_1\mathbf{j} - a_3b_2\mathbf{i}$$

$$= (a_2b_3 - a_3b_2)\,\mathbf{i} + (a_3b_1 - a_1b_3)\,\mathbf{j} + (a_1b_2 - a_2b_1)\,\mathbf{k} \tag{3.20}$$

This proves the proposition.

It is probably impossible for most people to remember 3.19. Fortunately, there is a somewhat easier way to remember it.

$$\mathbf{a}\times\mathbf{b}=\begin{vmatrix}\mathbf{i}&\mathbf{j}&\mathbf{k}\\a_1&a_2&a_3\\b_1&b_2&b_3\end{vmatrix} \tag{3.21}$$

where you expand the determinant along the top row. This yields

$$(a_2b_3-a_3b_2)\,\mathbf{i}-(a_1b_3-a_3b_1)\,\mathbf{j}+(a_1b_2-a_2b_1)\,\mathbf{k} \tag{3.22}$$

which is the same as 3.20.

You will see determinants later in the course but some of you have already seen them. All you need here is how to evaluate 2×2 and 3×3 determinants.

$$\begin{vmatrix}x&y\\z&w\end{vmatrix}=xw-yz$$

and

$$\begin{vmatrix}a&b&c\\x&y&z\\u&v&w\end{vmatrix}=a\begin{vmatrix}y&z\\v&w\end{vmatrix}-b\begin{vmatrix}x&z\\u&w\end{vmatrix}+c\begin{vmatrix}x&y\\u&v\end{vmatrix}.$$

Some of you are wondering what the rule is. You look at an entry in the top row and cross out the row and column which contain that entry. If the entry is in the i^{th} column, you multiply $(-1)^{1+i}$ times the determinant of the 2×2 which remains. This is the cofactor. You take the element in the top row times this cofactor and add all such up.

Example 3.2.4 *Find* $(\mathbf{i}-\mathbf{j}+2\mathbf{k})\times(3\mathbf{i}-2\mathbf{j}+\mathbf{k})$.

Use 3.21 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 3.2.5 *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 3.2.4.*

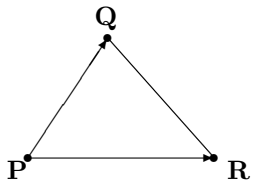
From Example 3.2.4 and the geometric description of the cross product, the area is just the norm of the vector obtained in Example 3.2.4. Thus the area is $\sqrt{9+25+1}=\sqrt{35}$.

Example 3.2.6 *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 $\frac{1}{2}\sqrt{4+49+1}=\frac{3}{2}\sqrt{6}$.

Observation 3.2.7 *In general, if you have three points (vectors) in* $\mathbb{R}^3,\mathbf{P},\mathbf{Q},\mathbf{R}$ *the area of the triangle is given by*

$$\frac{1}{2}\left|(\mathbf{Q}-\mathbf{P})\times(\mathbf{R}-\mathbf{P})\right|.$$



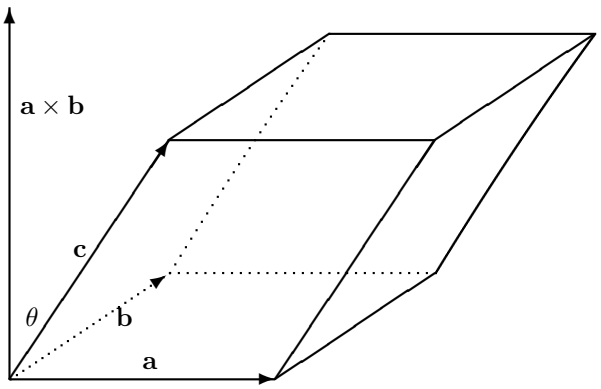
3.2.3 The Box Product, Triple Product

Definition 3.2.8 *A parallelepiped determined by the three vectors,* $\mathbf{a}, \mathbf{b}$, *and* $\mathbf{c}$ *consists of*

$$\{r\mathbf{a}+s\mathbf{b}+t\mathbf{c} : r, s, t \in [0,1]\}.$$

That is, if you pick three numbers, r, s, *and* t *each in* $[0,1]$ *and form* $r\mathbf{a}+s\mathbf{b}+t\mathbf{c}$, *then the collection of all such points is what is meant by the parallelepiped determined by these three vectors.*

The following is a picture of such a thing.



You notice the area of the base of the parallelepiped, the parallelogram determined by the vectors, $\mathbf{a}$ and $\mathbf{b}$ has area equal to $|\mathbf{a}\times\mathbf{b}|$ while the altitude of the parallelepiped is $|\mathbf{c}|\cos\theta$ where θ is the angle shown in the picture between $\mathbf{c}$ and $\mathbf{a}\times\mathbf{b}$. Therefore, the volume of this parallelepiped is the area of the base times the altitude which is just

$$|\mathbf{a}\times\mathbf{b}|\,|\mathbf{c}|\cos\theta = \mathbf{a}\times\mathbf{b}\cdot\mathbf{c}.$$

This expression is known as the box product and is sometimes written as $[\mathbf{a},\mathbf{b},\mathbf{c}]$. You should consider what happens if you interchange the $\mathbf{b}$ with the $\mathbf{c}$ or the $\mathbf{a}$ with the $\mathbf{c}$. You can see geometrically from drawing pictures that this merely introduces a minus sign. In any case the box product of three vectors always equals either the volume of the parallelepiped determined by the three vectors or else minus this volume.

Example 3.2.9 *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.

Observation 3.2.10 *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.*

$$\begin{aligned}\mathbf{u}\cdot\mathbf{v}\times\mathbf{w} &= (a,b,c)\cdot\begin{vmatrix}\mathbf{i}&\mathbf{j}&\mathbf{k}\\d&e&f\\g&h&i\end{vmatrix}\\ &= a\begin{vmatrix}e&f\\h&i\end{vmatrix}-b\begin{vmatrix}d&f\\g&i\end{vmatrix}+c\begin{vmatrix}d&e\\g&h\end{vmatrix}\\ &= \begin{vmatrix}a&b&c\\d&e&f\\g&h&i\end{vmatrix}.\end{aligned}$$

The message is that to take the box product, you can simply take the determinant of the 3×3 *"matrix" as described above.*

Example 3.2.11 *Find the volume of the parallelepiped determined by the vectors,* $(1,2,-1)$, $(2,1,5)$, *and* $(-3,1,2)$.

As just observed, it suffices to take the absolute value of the following determinant.

$$\begin{vmatrix}1&2&-1\\2&1&5\\-3&1&2\end{vmatrix} = -46$$

Thus the volume of this parallelepiped is 46.

There is a fundamental observation which comes directly from the geometric definitions of the cross product and the dot product.

Lemma 3.2.12 *Let* $\mathbf{a},\mathbf{b}$, *and* $\mathbf{c}$ *be vectors. Then* $(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c}=\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})$.

Proof: This follows from observing that either $(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c}$ and $\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})$ both give the volume of the parallelepiped or they both give -1 times the volume.

3.2.4 A Proof Of The Distributive Law For The Cross Product*

Here is a 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.

Example 3.2.13 *Find the volume of the parallelepiped determined by the vectors,*

$$(-1,2,3),(2,-1,1),(3,-2,3)$$

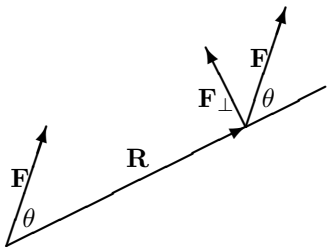
As just explained you only have to find the following 3×3 determinants.

$$\begin{vmatrix}-1&2&3\\2&-1&1\\3&-2&3\end{vmatrix}=-1\begin{vmatrix}-1&1\\-2&3\end{vmatrix}-2\begin{vmatrix}2&1\\3&3\end{vmatrix}+3\begin{vmatrix}2&-1\\3&-2\end{vmatrix}=-8$$

Now volume is always nonnegative so you take the absolute value of this number. The volume of the parallelepiped is 8.

3.2.5 Torque, Moment Of A Force

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 or moment can be obtained by simply adding up the torques associated with the different forces and positions.

Example 3.2.14 *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 3.2.15 *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.

3.2.6 Angular Velocity

Definition 3.2.16 *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 is a simple 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 3.2.17 *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 exercise 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 3.2.18 *A wheel rotates counter clockwise about the vector* $\mathbf{i}+\mathbf{j}+\mathbf{k}$ *at 60 revolutions per minute exactly as in Example 3.2.17. 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.

3.2.7 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 3.2.19 *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,

$$\left(\sum_{k=1}^{p} \mathbf{R}_k g m_k - \mathbf{R}_0 \sum_{k=1}^{p} g m_k \right) \times \mathbf{u} = \mathbf{0} \tag{3.23}$$

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 $\sum_{k=1}^{p} m_k$,

$$\mathbf{R}_0 = \frac{\sum_{k=1}^{p} \mathbf{R}_k m_k}{\sum_{k=1}^{p} m_k}. \tag{3.24}$$

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 3.2.20 *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 3.24

$$\begin{aligned} \mathbf{R}_0 &= \frac{5(2\mathbf{i} + 3\mathbf{j} + \mathbf{k}) + 6(\mathbf{i} - 3\mathbf{j} + 2\mathbf{k}) + 3(2\mathbf{i} - \mathbf{j} + 3\mathbf{k})}{5 + 6 + 3} \\ &= \frac{11}{7}\mathbf{i} - \frac{3}{7}\mathbf{j} + \frac{13}{7}\mathbf{k} \end{aligned}$$

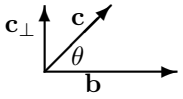
3.3 Further Explanations*

3.3.1 The Distributive Law For The Cross Product*

This section gives a proof for 3.17 which is independent of volume considerations. It is included here for the interested student. If you are satisfied with taking the distributive law on faith or are happy with the other argument given, it is not necessary to read this section. The proof given here is quite clever and follows the one given in [7]. The other approach based on areas is found in [23] and is discussed briefly earlier.

Lemma 3.3.1 *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,

$$\begin{aligned}|\mathbf{b}\times\mathbf{c}_\perp| &= |\mathbf{b}|\,|\mathbf{c}_\perp| \\ &= |\mathbf{b}|\,|\mathbf{c}|\sin\theta \\ &= |\mathbf{b}\times\mathbf{c}|\,.\end{aligned}$$

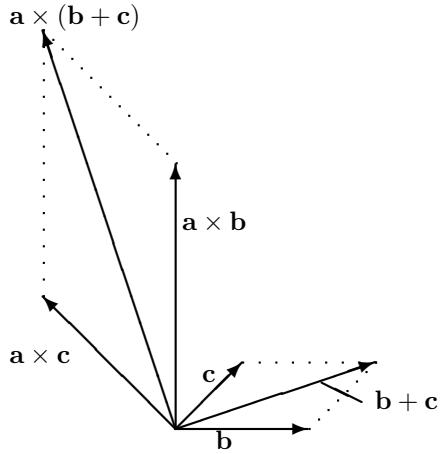
Therefore, $\mathbf{b}\times\mathbf{c}$ and $\mathbf{b}\times\mathbf{c}_\perp$ also have the same magnitude and so they are the same vector.

With this, the proof of the distributive law is in the following theorem.

Theorem 3.3.2 *Let* $\mathbf{a},\mathbf{b}$, *and* $\mathbf{c}$ *be vectors in* $\mathbb{R}^3$. *Then*

$$\mathbf{a}\times(\mathbf{b}+\mathbf{c}) = \mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c} \tag{3.25}$$

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 3.25.

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.

3.3.2 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.

$$\begin{aligned}\mathbf{v}\times\mathbf{w} &= \begin{vmatrix}\mathbf{i} & \mathbf{j} & \mathbf{k}\\ v_1 & v_2 & v_3\\ w_1 & w_2 & w_3\end{vmatrix}\\ &= (v_2w_3-v_3w_2)\,\mathbf{i}+(w_1v_3-v_1w_3)\,\mathbf{j}+(v_1w_2-v_2w_1)\,\mathbf{k}\end{aligned}$$

Next consider $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$ which is given by

$$\mathbf{u}\times(\mathbf{v}\times\mathbf{w})=\begin{vmatrix}\mathbf{i} & \mathbf{j} & \mathbf{k}\\ u_1 & u_2 & u_3\\ (v_2w_3-v_3w_2) & (w_1v_3-v_1w_3) & (v_1w_2-v_2w_1)\end{vmatrix}.$$

When you multiply this out, you get

$$\begin{aligned}&\mathbf{i}\,(v_1u_2w_2+u_3v_1w_3-w_1u_2v_2-u_3w_1v_3)+\mathbf{j}\,(v_2u_1w_1+v_2w_3u_3-w_2u_1v_1-u_3w_2v_3)\\ &+\mathbf{k}\,(u_1w_1v_3+v_3w_2u_2-u_1v_1w_3-v_2w_3u_2)\end{aligned}$$

and if you are clever, you see right away that

$$(\mathbf{i}v_1+\mathbf{j}v_2+\mathbf{k}v_3)\,(u_1w_1+u_2w_2+u_3w_3)-(\mathbf{i}w_1+\mathbf{j}w_2+\mathbf{k}w_3)\,(u_1v_1+u_2v_2+u_3v_3)\,.$$

Thus

$$\mathbf{u}\times(\mathbf{v}\times\mathbf{w})=\mathbf{v}\,(\mathbf{u}\cdot\mathbf{w})-\mathbf{w}\,(\mathbf{u}\cdot\mathbf{v})\,. \tag{3.26}$$

A related formula is

$$\begin{aligned}(\mathbf{u}\times\mathbf{v})\times\mathbf{w} &= -[\mathbf{w}\times(\mathbf{u}\times\mathbf{v})]\\ &= -[\mathbf{u}\,(\mathbf{w}\cdot\mathbf{v})-\mathbf{v}\,(\mathbf{w}\cdot\mathbf{u})]\\ &= \mathbf{v}\,(\mathbf{w}\cdot\mathbf{u})-\mathbf{u}\,(\mathbf{w}\cdot\mathbf{v})\,.\end{aligned} \tag{3.27}$$

This derivation is simply wretched and it does nothing for other identities which may arise in applications. Actually, the above two formulas, 3.26 and 3.27 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 3.3.3 *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 3.3.4 *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 $a_i b_i$ means $\sum_i a_i b_i$. 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 $a_i b_i$ 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 $\sum_i a_i b_i c_i$ it is best to simply use the summation notation. There is a very important reduction identity connecting these two symbols.

Lemma 3.3.5 *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 3.3.6 *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 3.3.7 *Discover a formula which simplifies* $(\mathbf{u} \times \mathbf{v}) \times \mathbf{w}$.

From the above reduction formula,

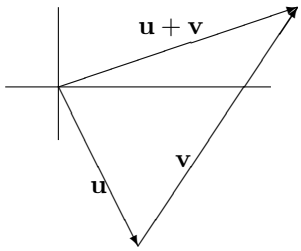
$$\begin{aligned} ((\mathbf{u} \times \mathbf{v}) \times \mathbf{w})_i &= \varepsilon_{ijk} (\mathbf{u} \times \mathbf{v})_j w_k \\ &= \varepsilon_{ijk} \varepsilon_{jrs} u_r v_s w_k \\ &= -\varepsilon_{jik} \varepsilon_{jrs} u_r v_s w_k \\ &= -(\delta_{ir} \delta_{ks} - \delta_{is} \delta_{kr}) u_r v_s w_k \\ &= -(u_i v_k w_k - u_k v_i w_k) \\ &= \mathbf{u} \cdot \mathbf{w} v_i - \mathbf{v} \cdot \mathbf{w} u_i \\ &= ((\mathbf{u} \cdot \mathbf{w}) \mathbf{v} - (\mathbf{v} \cdot \mathbf{w}) \mathbf{u})_i . \end{aligned}$$

Since this holds for all i, it follows that

$$(\mathbf{u} \times \mathbf{v}) \times \mathbf{w} = (\mathbf{u} \cdot \mathbf{w}) \mathbf{v} - (\mathbf{v} \cdot \mathbf{w}) \mathbf{u}.$$

3.3.3 Exercises With Answers

1. Draw the vector $\mathbf{u} = (1, -2)$, the vector $\mathbf{v} = (2, 3)$, and the vector $(1, -2) + (2, 3) = \mathbf{u} + \mathbf{v}$.



2. Let $\mathbf{u} = (1,2,-5), \mathbf{v} = (3,-1,2)$ and $\mathbf{w} = (2,0,3)$ Find the following.

   (a) $(2\mathbf{u}+\mathbf{v})\cdot\mathbf{w}$
   This is $(2(1,2,-5)+(3,-1,2))\cdot(2,0,3) = -14$. Here is why.

   $$2(1,2,-5)+(3,-1,2) = (5,3,-8)$$

   and

   $$(5,3,-8)\cdot(2,0,3) = -14$$

   (b) $(\mathbf{u}-3\mathbf{v})\cdot\mathbf{w}$
   This is $((1,2,-5)-3(3,-1,2))\cdot(2,0,3) = -49$

3. Find the cosine of the angle between the two vectors, $(1,2,5)$ and $(3,-2,1)$.
   $\cos\theta = \frac{(1,2,5)\cdot(3,-2,1)}{\sqrt{1+4+25}\sqrt{9+4+1}} = \frac{1}{105}\sqrt{30}\sqrt{14} = .19518.$

4. Here are two vectors. $(1,2,3)$ and $(3,2,1)$. Find a vector which is perpendicular to both of these vectors.

   One way to do this is to take the cross product of the two vectors. $(1,2,3)\times(3,2,1) = (-4,8,-4)$. A vector perpendicular to both of these vectors is $(-1,2,1)$. Note nothing is changed as far as being perpendicular is concerned by division by 4.

5. Given two vectors in $\mathbb{R}^n, \mathbf{u}, \mathbf{v}$ show that

   $$\mathbf{u}\cdot\mathbf{v} = \frac{1}{4}\left(|\mathbf{u}+\mathbf{v}|^2 - |\mathbf{u}-\mathbf{v}|^2\right).$$

   This is really easy if you remember the axioms for the dot product. Otherwise it is very troublesome. Start with the right side.

   $$\begin{aligned}\frac{1}{4}\left(|\mathbf{u}+\mathbf{v}|^2 - |\mathbf{u}-\mathbf{v}|^2\right) &= \frac{1}{4}\left((\mathbf{u}+\mathbf{v})\cdot(\mathbf{u}+\mathbf{v}) - (\mathbf{u}-\mathbf{v})\cdot(\mathbf{u}-\mathbf{v})\right)\\ &= \frac{1}{4}\left[\mathbf{u}\cdot\mathbf{u}+ + \mathbf{v}\cdot\mathbf{v} - \{\mathbf{u}\cdot\mathbf{u}+\mathbf{v}\cdot\mathbf{v} - 2\mathbf{u}\cdot\mathbf{v}\}\right]\\ &= \frac{1}{4}\left[2\mathbf{u}\cdot\mathbf{v} - (-2\mathbf{u}\cdot\mathbf{v})\right] = \mathbf{u}\cdot\mathbf{v}.\end{aligned}$$

6. If $|\mathbf{u}| = 3$, $|\mathbf{v}| = 4$, and $\mathbf{u}\cdot\mathbf{v} = 5$, find $|\mathbf{u}+\mathbf{v}|$.

   This is easy if you know the properties of the dot product. Otherwise it is trouble.

   $$\begin{aligned}|\mathbf{u}+\mathbf{v}|^2 &= |\mathbf{u}|^2 + |\mathbf{v}|^2 + 2\mathbf{u}\cdot\mathbf{v}\\ &= 9+16+50 = 75.\end{aligned}$$

   Therefore, $|\mathbf{u}+\mathbf{v}| = 5\sqrt{5}$.

7. Find vectors in $\mathbb{R}^3$, $\mathbf{u},\mathbf{v}$ such that $\mathbf{u}\cdot\mathbf{v} = 6$ and $|\mathbf{u}| = 2$ while $|\mathbf{v}| = 3$. You see that equality holds in the Cauchy Schwarz inequality and so one of these vectors must be a multiple of the other. It must be a positive multiple of the other because the dot product is positive which implies the angle between the vectors is 0 and not π. Let $\mathbf{u} = (0,0,2), \mathbf{v} = (0,0,3)$. This appears to work. You should find some other examples. What if $\mathbf{u} = \left(\sqrt{2}/2, \sqrt{2}/2, \sqrt{3}\right)$. In this case $|\mathbf{u}| = 2$ also. Can you find $\mathbf{v}$ such that the above will hold?

8. The projection of $\mathbf{u}$ onto $\mathbf{v}$, denoted by $\mathrm{proj}_{\mathbf{v}}(\mathbf{u})$ is given by the formula

$$\mathrm{proj}_{\mathbf{v}}(\mathbf{u}) = \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{v}|^2}\mathbf{v}$$

Show $\mathbf{u} - \mathrm{proj}_{\mathbf{v}}(\mathbf{u})$ is perpendicular to $\mathbf{v}$. Also show $\mathrm{proj}_{\mathbf{v}}(a\mathbf{u}+b\mathbf{w}) = a(\mathrm{proj}_{\mathbf{v}}(\mathbf{u})) + b(\mathrm{proj}_{\mathbf{v}}(\mathbf{w}))$.

This is another of those things which is very easy if you know the properties of the dot product but lots of trouble if you don't. Of course you can persist in not learning these things if you want. It is up to you.

$$\mathbf{v}\cdot(\mathbf{u} - \mathrm{proj}_{\mathbf{v}}(\mathbf{u})) = \mathbf{v}\cdot\left(\mathbf{u} - \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{v}|^2}\mathbf{v}\right) = \mathbf{v}\cdot\mathbf{u} - \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{v}|^2}\mathbf{v}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{u} - \mathbf{u}\cdot\mathbf{v} = 0$$

This does it. The dot product equals zero and so the two vectors are perpendicular. As to the other claim,

$$\begin{aligned}\frac{(a\mathbf{u}+b\mathbf{w})\cdot\mathbf{v}}{|\mathbf{v}|^2}\mathbf{v} &= a\frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{v}|^2}\mathbf{v} + b\frac{\mathbf{w}\cdot\mathbf{v}}{|\mathbf{v}|^2}\mathbf{v} \\ &\equiv a(\mathrm{proj}_{\mathbf{v}}(\mathbf{u})) + b(\mathrm{proj}_{\mathbf{v}}(\mathbf{w}))\,.\end{aligned}$$

Now that was real easy wasn't it. Note I never said anything about $\mathbf{u}, \mathbf{v}$ being lists of numbers. I just used the properties of the dot product.

9. Find the angle between the vectors $3\mathbf{i} - \mathbf{j} - \mathbf{k}$ and $\mathbf{i} + 4\mathbf{j} + 2\mathbf{k}$.

$\cos\theta = \frac{3-4-2}{\sqrt{9+1+1}\sqrt{1+16+4}} = -.19739$. Therefore, you have to solve the equation $\cos\theta = -.19739$, Solution is : $\theta = 1.7695$ radians. You need to use a calculator or table to solve this.

10. Find $\mathrm{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{v} = (1,3,-2)$ and $\mathbf{u} = (1,2,3)$.

Remember to find this you take $\frac{\mathbf{v}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u}$. Thus the answer is $\frac{1}{14}(1,2,3)$.

11. If $\mathbf{F}$ is a force and $\mathbf{D}$ is a vector, show $\mathrm{proj}_{\mathbf{D}}(\mathbf{F}) = (|\mathbf{F}|\cos\theta)\,\mathbf{u}$ 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, $\mathbf{D}$.

$\mathrm{proj}_{\mathbf{D}}(\mathbf{F}) = \frac{\mathbf{F}\cdot\mathbf{D}}{\mathbf{D}\cdot\mathbf{D}}\mathbf{D} = |\mathbf{F}||\mathbf{D}|\cos\theta\frac{1}{|\mathbf{D}|^2}\mathbf{D} = |\mathbf{F}|\cos\theta\frac{\mathbf{D}}{|\mathbf{D}|}$.

12. A boy drags a sled for 100 feet along the ground by pulling on a rope which is 40 degrees from the horizontal with a force of 10 pounds. How much work does this force do?

The component of force is $10\cos\left(\frac{40}{180}\pi\right)$ and it acts for 100 feet so the work done is

$$10\cos\left(\frac{40}{180}\pi\right)\times 100 = 766.04$$

13. If $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ are vectors. Show that $(\mathbf{b}+\mathbf{c})_{\perp} = \mathbf{b}_{\perp} + \mathbf{c}_{\perp}$ where $\mathbf{b}_{\perp} = \mathbf{b} - \mathrm{proj}_{\mathbf{a}}(\mathbf{b})$.

14. Find $(1,0,3,4)\cdot(2,7,1,3)$. $(1,0,3,4)\cdot(2,7,1,3) = 17$.

15. Prove from the axioms of the dot product the parallelogram identity, $|\mathbf{a}+\mathbf{b}|^2 + |\mathbf{a}-\mathbf{b}|^2 = 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2$.

    Use the properties of the dot product and the definition of the norm in terms of the dot product. Thus the left side is

    $$\mathbf{a}\cdot\mathbf{a}+\mathbf{b}\cdot\mathbf{b}+2(\mathbf{a}\cdot\mathbf{b})+\mathbf{a}\cdot\mathbf{a}+\mathbf{b}\cdot\mathbf{b}-2\mathbf{a}\cdot\mathbf{b} = 2|\mathbf{a}|^2+2|\mathbf{b}|^2.$$

16. Find all vectors, (x,y) which are perpendicular to $(1,2)$.

    You need $x+2y=0$ so $x=-2y$ and you can write all desired vectors in the form

    $$(-2y,y) : y \in \mathbb{R}.$$

17. Find the line through $(1,2,1)$ and $(2,0,3)$.

    First get a direction vector which in this case is $(1,-2,2)$. Then the equation of the line is

    $$(x,y,z) = (1,2,1)+t(1,-2,2) = (1+t,2-2t,1+2t).$$

    Thus a parametric form for this line is $x=1+t, y=2-2t, z=1+2t$ and a vector equation for this line is

    $$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + t\begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$$

    if you want to write the vectors as column vectors.

18. In $\mathbb{R}^2$, the equation of a line is given as $2x+3y=6$. Find a vector equation of this line.

    One way to do it is to get a couple of points on the line and then do as in the previous problem. Two points on this line are $(0,2)$ and $(3,0)$. Then a direction vector for the line is $(-3,2)$ and so a vector equation of the line is

    $$(x,y) = (0,2)+t(-3,2).$$

    Written parametrically, this would be $x=-3t, y=2+2t$.

19. Suppose you have the vector equation for a line joining the two points, $\mathbf{p},\mathbf{q}$. This is

    $$\mathbf{x} = \mathbf{p}+t(\mathbf{q}-\mathbf{p})$$

    Note this works because when $t=0$ the right side is $\mathbf{p}$ and when $t=1$, the right side is $\mathbf{q}$. Now find the point which is $1/3$ of the way between $\mathbf{p}$ and $\mathbf{q}$.

    This point would be obtained by letting $t=1/3$. Thus the point is

    $$\mathbf{x}_{1/3} = \frac{2}{3}\mathbf{p}+\frac{1}{3}\mathbf{q}.$$

    Does it work?

    $$|\mathbf{x}_{1/3}-\mathbf{p}| = \left|-\frac{1}{3}\mathbf{p}+\frac{1}{3}\mathbf{q}\right| = \frac{1}{3}|\mathbf{q}-\mathbf{p}|.$$

    Seems to work just fine. I suppose you could also find points which are $1/5$ of the way between $\mathbf{p}$ and $\mathbf{q}$ also.

20. The wind blows from West to East at a speed of 30 kilometers per hour and an airplane which travels at 300 Kilometers per hour in still air is heading North West. What is the velocity of the airplane relative to the ground? What is the component of this velocity in the direction North?

    Let the positive y axis point in the direction North and let the positive x axis point in the direction East. The velocity of the wind is $30\mathbf{i}$. The plane moves in the direction $\mathbf{i}+\mathbf{j}$. A unit vector in this direction is $\frac{1}{\sqrt{2}}(\mathbf{i}+\mathbf{j})$. Therefore, the velocity of the plane relative to the ground is $30\mathbf{i}+\frac{300}{\sqrt{2}}(\mathbf{i}+\mathbf{j}) = 150\sqrt{2}\mathbf{j}+\left(30+150\sqrt{2}\right)\mathbf{i}$. The component of velocity in the direction North is $150\sqrt{2}$.

21. In the situation of Problem 20 how many degrees to the West of North should the airplane head in order to fly exactly North. What will be the speed of the airplane relative to the ground?

    In this case the unit vector will be $-\sin(\theta)\mathbf{i}+\cos(\theta)\mathbf{j}$. Therefore, the velocity of the plane will be

    $$300(-\sin(\theta)\mathbf{i}+\cos(\theta)\mathbf{j})$$

    and this is supposed to satisfy

    $$300(-\sin(\theta)\mathbf{i}+\cos(\theta)\mathbf{j})+30\mathbf{i} = 0\mathbf{i}+?\mathbf{j}.$$

    Therefore, you need to have $\sin\theta = 1/10$, which means $\theta = .100\,17$ radians. Therefore, the degrees should be $\frac{.1\times 180}{\pi} = 5.729\,6$ degrees. In this case the velocity vector of the plane relative to the ground is $300\left(\frac{\sqrt{99}}{10}\right)\mathbf{j}$.

22. In the situation of 21 suppose the airplane uses 34 gallons of fuel every hour at that air speed and that it needs to fly North a distance of 600 miles. Will the airplane have enough fuel to arrive at its destination given that it has 63 gallons of fuel?

    The airplane needs to fly 600 miles at a speed of $300\left(\frac{\sqrt{99}}{10}\right)$. Therefore, it takes $\frac{600}{\left(300\left(\frac{\sqrt{99}}{10}\right)\right)} = 2.010\,1$ hours to get there. Therefore, the plane will need to use about 68 gallons of gas. It won't make it.

23. A certain river is one half mile wide with a current flowing at 3 miles per hour from East to West. A man swims directly toward the opposite shore from the South bank of the river at a speed of 2 miles per hour. How far down the river does he find himself when he has swam across? How far does he end up swimming?

    The velocity of the man relative to the earth is then $-3\mathbf{i}+2\mathbf{j}$. Since the component of $\mathbf{j}$ equals 2 it follows he takes 1/8 of an hour to get across. Durring this time he is swept downstream at the rate of 3 miles per hour and so he ends up 3/8 of a mile down stream. He has gone $\sqrt{\left(\frac{3}{8}\right)^2+\left(\frac{1}{2}\right)^2} = .625$ miles in all.

24. Three forces are applied to a point which does not move. Two of the forces are $2\mathbf{i}-\mathbf{j}+3\mathbf{k}$ Newtons and $\mathbf{i}-3\mathbf{j}-2\mathbf{k}$ Newtons**.** Find the third force.

    Call it $a\mathbf{i}+b\mathbf{j}+c\mathbf{k}$ Then you need $a+2+1=0, b-1-3=0$, and $c+3-2=0$. Therefore, the force is $-3\mathbf{i}+4\mathbf{j}-\mathbf{k}$.

25. If you only assume 3.23 holds for $\mathbf{u} = \mathbf{i},\mathbf{j},\mathbf{k}$, show that this implies 3.23 holds for all unit vectors, $\mathbf{u}$.

    Suppose than that $(\sum_{k=1}^{p}\mathbf{R}_k gm_k - \mathbf{R}_0\sum_{k=1}^{p}gm_k)\times\mathbf{u} = \mathbf{0}$ for $\mathbf{u} = \mathbf{i},\mathbf{j},\mathbf{k}$. Then if $\mathbf{u}$ is an arbitrary unit vector, $\mathbf{u}$ must be of the form $a\mathbf{i}+b\mathbf{j}+c\mathbf{k}$. Now from the distributive

property of the cross product and letting $\mathbf{w} = (\sum_{k=1}^{p} \mathbf{R}_k g m_k - \mathbf{R}_0 \sum_{k=1}^{p} g m_k)$, this says

$$\begin{aligned}(\textstyle\sum_{k=1}^{p} \mathbf{R}_k g m_k - \mathbf{R}_0 \sum_{k=1}^{p} g m_k) \times \mathbf{u} \\ = \mathbf{w} \times (a\mathbf{i} + b\mathbf{j} + c\mathbf{k}) \\ = a\mathbf{w} \times \mathbf{i} + b\mathbf{w} \times \mathbf{j} + c\mathbf{w} \times \mathbf{k} \\ = \mathbf{0} + \mathbf{0} + \mathbf{0} = \mathbf{0}.\end{aligned}$$

26. Let $m_1 = 4, m_2 = 3$, 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 $2\mathbf{i} - 3\mathbf{j} + \mathbf{k}$ and m_3 is located at $2\mathbf{i} + \mathbf{j} + 3\mathbf{k}$. Find the center of mass of these three masses.

    Let the center of mass be located at $a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$. Then $(4+3+1)(a\mathbf{i} + b\mathbf{j} + c\mathbf{k}) = 4(2\mathbf{i} - \mathbf{j} + \mathbf{k}) + 3(2\mathbf{i} - 3\mathbf{j} + \mathbf{k}) + 1(2\mathbf{i} + \mathbf{j} + 3\mathbf{k}) = 16\mathbf{i} - 12\mathbf{j} + 10\mathbf{k}$. Therefore, $a = 2, b = \frac{-3}{2}$ and $c = \frac{5}{4}$. The center of mass is then $2\mathbf{i} - \frac{3}{2}\mathbf{j} + \frac{5}{4}\mathbf{k}$.

27. Find the angular velocity vector of a rigid body which rotates counter clockwise about the vector $\mathbf{i} - \mathbf{j} + \mathbf{k}$ at 20 revolutions per minute. Assume distance is measured in meters.

    The angular velocity is $20 \times 2\pi = 40\pi$. Then $\mathbf{\Omega} = 40\pi \frac{1}{\sqrt{3}}(\mathbf{i} - \mathbf{j} + \mathbf{k})$.

28. Find the area of the triangle determined by the three points, $(1,2,3), (1,2,6)$ and $(-3,2,1)$.

    The three points determine two displacement vectors from the point $(1,2,3), (0,0,3)$ and $(-4,0,-2)$. To find the area of the parallelogram determined by these two displacement vectors, you simply take the norm of their cross product. To find the area of the triangle, you take one half of that. Thus the area is

$$(1/2)\,|(0,0,3) \times (-4,0,-2)| = \frac{1}{2}\,|(0,-12,0)| = 6.$$

29. Find the area of the parallelogram determined by the vectors, $(1,0,3)$ and $(4,-2,1)$.

    $|(1,0,3) \times (4,-2,1)| = |(6,11,-2)| = \sqrt{26 + 121 + 4} = \sqrt{151}$.

30. Find the volume of the parallelepiped determined by the vectors, $\mathbf{i} - 7\mathbf{j} - 5\mathbf{k}, \mathbf{i} + 2\mathbf{j} - 6\mathbf{k}$,$3\mathbf{i} - 3\mathbf{j} + \mathbf{k}$.

    Remember you just need to take the absolute value of the determinant having the given vectors as rows. Thus the volume is the absolute value of

$$\begin{vmatrix} 1 & -7 & -5 \\ 1 & 2 & -6 \\ 3 & -3 & 1 \end{vmatrix} = 162$$

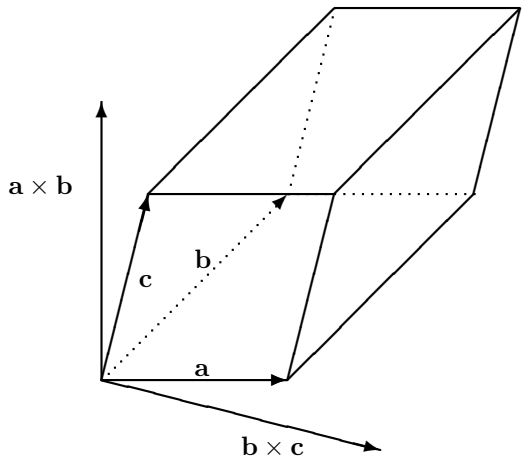
31. Suppose $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ are three vectors whose components are all integers. Can you conclude the volume of the parallelepiped determined from these three vectors will always be an integer?

    Hint: Consider what happens when you take the determinant of a matrix which has all integers.

32. Using the notion of the box product yielding either plus or minus the volume of the parallelepiped determined by the given three vectors, show that

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$$

In other words, the dot and the cross can be switched as long as the order of the vectors remains the same. **Hint:** There are two ways to do this, by the coordinate description of the dot and cross product and by geometric reasoning. It is best if you use the geometric reasoning. Here is a picture which might help.



In this picture there is an angle between $\mathbf{a}\times\mathbf{b}$ and $\mathbf{c}$. Call it θ. Now if you take $|\mathbf{a}\times\mathbf{b}|\,|\mathbf{c}|\cos\theta$ this gives the area of the base of the parallelepiped determined by $\mathbf{a}$ and $\mathbf{b}$ times the altitude of the parallelepiped, $|\mathbf{c}|\cos\theta$. This is what is meant by the volume of the parallelepiped. It also equals $\mathbf{a}\times\mathbf{b}\cdot\mathbf{c}$ by the geometric description of the dot product. Similarly, there is an angle between $\mathbf{b}\times\mathbf{c}$ and $\mathbf{a}$. Call it α. Then if you take $|\mathbf{b}\times\mathbf{c}|\,|\mathbf{a}|\cos\alpha$ this would equal the area of the face determined by the vectors $\mathbf{b}$ and $\mathbf{c}$ times the altitude measured from this face, $|\mathbf{a}|\cos\alpha$. Thus this also is the volume of the parallelepiped. and it equals $\mathbf{a}\cdot\mathbf{b}\times\mathbf{c}$. The picture is not completely representative. If you switch the labels of two of these vectors, say $\mathbf{b}$ and $\mathbf{c}$, explain why it is still the case that $\mathbf{a}\cdot\mathbf{b}\times\mathbf{c}=\mathbf{a}\times\mathbf{b}\cdot\mathbf{c}$. You should draw a similar picture and explain why in this case you get -1 times the volume of the parallelepiped.

33. Discover a vector identity for$(\mathbf{u}\times\mathbf{v})\times\mathbf{w}$.

$$\begin{aligned}((\mathbf{u}\times\mathbf{v})\times\mathbf{w})_i &= \varepsilon_{ijk}\left(\mathbf{u}\times\mathbf{v}\right)_j w_k = \varepsilon_{ijk}\varepsilon_{jrs}u_r v_s w_k = (\delta_{is}\delta_{kr}-\delta_{ir}\delta_{ks})\,u_r v_s w_k\\ &= u_k w_k v_i - u_i v_k w_k = (\mathbf{u}\cdot\mathbf{w})\,v_i - (\mathbf{v}\cdot\mathbf{w})\,u_i.\end{aligned}$$

Therefore, $(\mathbf{u}\times\mathbf{v})\times\mathbf{w}=(\mathbf{u}\cdot\mathbf{w})\,\mathbf{v}-(\mathbf{v}\cdot\mathbf{w})\,\mathbf{u}$.

34. Discover a vector identity for $(\mathbf{u}\times\mathbf{v})\cdot(\mathbf{z}\times\mathbf{w})$.

Start with $\varepsilon_{ijk}u_j v_k\varepsilon_{irs}z_r w_s$ and then go to work on it using the reduction identities for the permutation symbol.

35. Discover a vector identity for $(\mathbf{u}\times\mathbf{v})\times(\mathbf{z}\times\mathbf{w})$ in terms of box products.

You will save time if you use the identity for $(\mathbf{u}\times\mathbf{v})\times\mathbf{w}$ or $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$.

Part II

Planes And Systems Of Equations

Outcomes

Planes

A. 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.

B. Determine a normal vector and the intercepts of a given plane.

C. Sketch the graph of a plane given its equation.

D. Determine the angle between two planes.

E. Find the equation of a plane determined by lines.

Reading: Multivariable Calculus 1.3, Linear Algebra 1.3
Outcome Mapping:

A. 1,3

B. 2,4

C. 4

D. 2

E. Problem 9 in Section 1.5 of Multivariable Calculus

Systems of Linear Equations

A. Define linear equation and system of linear equations. Define solution and solution set for both an linear equation and a system of linear equations.

B. Relate the following 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:

   (a) a unique solution.
   (b) infinitely many solutions.
   (c) no solution.

C. Represent a linear system as an augmented matrix and vice versa.

D. Transform a system to a triangular pattern and then apply back substitution to solve the linear system.

E. Represent the solution set to a linear system using parametric equations.

Reading: Linear Algebra 2.1
Outcome Mapping:

A. 1-6, 7-10

B. 15-18

C. 27-30, 31-32

D. 19-24, 25-26, 33-38

E. 11-14, 39-40

Direct Methods for Solving Linear Systems

A. Identify matrices that are in row echelon form and reduced row echelon form.

B. Determine whether a system of linear equations has no solution, a unique solution or an infinite number of solutions from its echelon form.

C. Apply elementary row operations to transform systems of linear equations.

D. Solve systems of linear equations using Gaussian elimination.

E. Solve systems of linear equations using Gauss-Jordan elimination.

F. Define and evaluate the rank of a matrix.

G. Apply the Rank Theorem relate the rank of an augmented matrix to the solution set of a system in the case of homogeneous and nonhomogeneous systems.

H. Model and solve application problems using linear systems.

Reading: Linear Algebra 2.2
Outcome Mapping:

A. 1-8,24

B. 39-44

C. 9-14,15-16,17-18,19-22

D. 25-34

E. 23

F. 35-38

G. 45-52, (2.4: 1-47)

Planes 11 Sept.

4.1 Finding Planes

Quiz

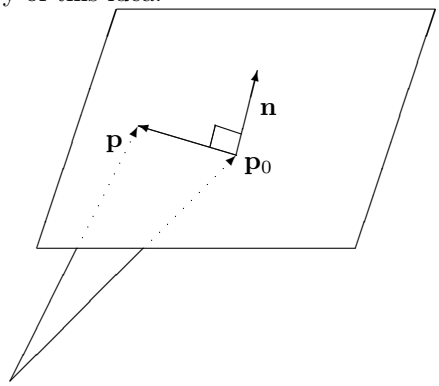
1. Let $\mathbf{a} = (1,2,3)$, $\mathbf{b} = (2,-1,1)$. Find a vector which is perpendicular to both of these vectors.
2. Find the area of the parallelogram determined by the above two vectors.
3. Find the cosine of the angle between the above two vectors.
4. Find the sine of the angle between the above two vectors.
5. Find the volume of the parallelepiped determined by the vectors, $\mathbf{a} = (1,2,3)$, $\mathbf{b} = (2,-1,1)$ and $\mathbf{c} = (1,1,1)$.

4.1.1 Planes From A Normal And A Point

A plane is a long flat thing. 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 4.1.1 *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

$$(1,2,3)\cdot(x-2,y+1,z-5) = x-15+2y+3z = 0$$

Example 4.1.2 $2x+4y-5z=11$ *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.

Proposition 4.1.3 *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 $a\mathbf{i}+b\mathbf{j}+c\mathbf{k}$. Suppose $a\neq 0$. Then the points which satisfy $ax+by+cz=d$ are the same as the points which satisfy

$$a\left(x-\frac{d}{a}\right)+b(y-0)+c(z-0)=0.$$

Thus a point on the plane is $\left(\frac{d}{a},0,0\right)$ and a normal vector is (a,b,c) as claimed. (You can do something similar if $b\neq 0$. Note there are many points on the plane. This just picks out one.)

The converse follows from the above discussion involving the point and a normal vector. This proves the proposition.

Example 4.1.4 *Find a normal vector to the plane* $2x+5y-z=12.3$.

A normal vector is $(2,5,-1)$. A point on this plane is $(0,0,-12.3)$. Of course there are many other points on this plane.

4.1.2 The Angle Between Two Planes

Definition 4.1.5 *Suppose two planes intersect. The angle between the planes is defined to be the angle between their normal vectors.*

Example 4.1.6 *Find the angle between the two planes,* $x+2y-z=6$ *and* $3x+2y-z=7$.

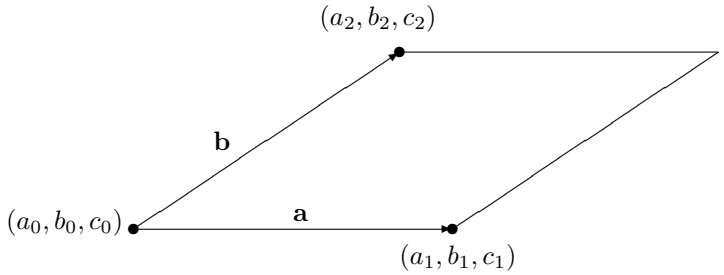
The two normal vectors are $(1,2,-1)$ and $(3,2,-1)$. Therefore, the cosine of the angle desired is

$$\cos\theta = \frac{(1,2,-1)\cdot(3,2,-1)}{\sqrt{1^2+2^2+(-1)^2}\sqrt{3^2+2^2+(-1)^2}} = .872\,87$$

Now use a calculator or table to find what the angle is. $\cos\theta = .872\,87$, Solution is : $\{\theta = .509\,74\}$. This value is in radians.

4.1.3 The Plane Which Contains Three Points

Sometimes you need to find the equation of a plane which contains three points. Consider the following picture.



You have plenty of points but you need a normal. This can be obtained by taking $\mathbf{a} \times \mathbf{b}$ where $\mathbf{a} = (a_1 - a_0, b_1 - b_0, c_1 - c_0)$ and $\mathbf{b} = (a_2 - a_0, b_2 - b_0, c_2 - c_0)$.

Example 4.1.7 *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 $y + 3z = 5$. When you have what you think is the plane containing the three points, you ought to check it by seeing if it really does contain the three points.

Example 4.1.8 *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

$$\begin{vmatrix} 3 & 0 & 0 \\ 2 & -3 & 1 \\ x-1 & y-2 & z-1 \end{vmatrix} = 0.$$

Hence $-9z + 15 - 3y = 0$ and dividing by 3 yields the same answer as the above.

Example 4.1.9 *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 procede as above.

Example 4.1.10 *Find the equation of the plane which contains the two lines, given by the following parametric expressions in which* $t \in \mathbb{R}$.

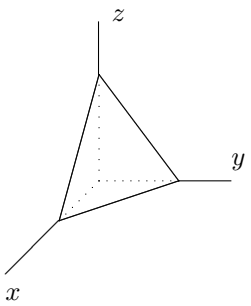
$$(2t, 1+t, 1+2t) = (x, y, z), \ (2t+2, 1, 3+2t) = (x, y, z)$$

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.

4.1.4 Intercepts Of A 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 4.1.11 *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 4.1.12 *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.

In general, to find the intercepts of a plane of the form $ax + by + cz = d$ where $d \neq 0$ and none of a, b, or c are equal to 0, divide by d. This gives

$$\frac{x}{(d/a)} + \frac{y}{(d/b)} + \frac{z}{(d/c)} = 1$$

the intercepts are $\left(\frac{d}{a}, 0, 0\right), \left(0, \frac{d}{b}, 0\right), \left(0, 0, \frac{d}{c}\right)$.

4.1.5 Distance Between A Point And A Plane Or A Point And A Line*

There exists a stupid formula for the distance between a point and a plane. I will first illustrate with an example.

Example 4.1.13 *Find the distance from the point* $(1,2,3)$ *to the plane* $x-y+z=3$.

The distance is the length of the line segment normal to the plane which goes from the given point to the given plane. In this example, a direction vector for this line is $(1,-1,1)$, a normal vector to the plane. Thus the equation for the desired line is

$$\begin{aligned}(x,y,z) &= (1,2,3)+t(1,-1,1)\\ &= (1+t,2-t,3+t)\end{aligned}$$

Lets find the value of t at which the line intersects the plane. Thus

$$(1+t)-(2-t)+(3+t)=3$$

and so $t=\frac{1}{3}$. Therefore, the line segment is the one which joins $(1,2,3)$ to

$$\left(1+\frac{1}{3},2-\frac{1}{3},3+\frac{1}{3}\right)=\left(\frac{4}{3},\frac{5}{3},\frac{10}{3}\right).$$

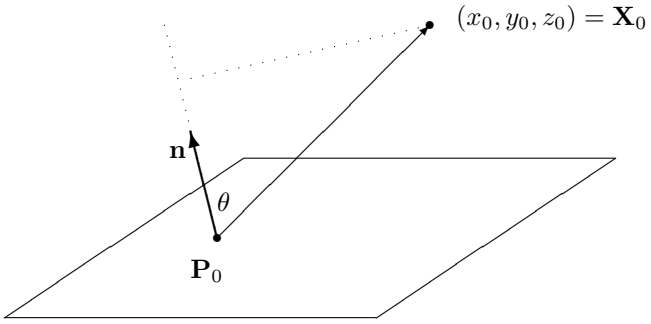
Now it follows the desired distance is

$$\sqrt{\left(1-\frac{4}{3}\right)^2+\left(2-\frac{5}{3}\right)^2+\left(3-\frac{10}{3}\right)^2}=\frac{1}{3}\sqrt{3}$$

In the general case there is a simple and interesting geometrical consideration which will lead to a stupid formula which you can then use with no thought to do an uninteresting task, finding the distance from a point to a plane.

Example 4.1.14 *Find the distance from the point* (x_0,y_0,z_0) *to the plane* $ax+by+cz=d$. *Here* $(a,b,c)\neq(0,0,0)$.

Consider the following picture in which $\mathbf{P}_0$ is a point in the plane and $\mathbf{X}_0=(x_0,y_0,z_0)$ is the point whose distance to the plane is to be found. The normal to the plane is $\mathbf{n}$.



Then from the picture, what you want is to take the projection of the vector $\mathbf{X}_0-\mathbf{P}_0$ onto the line determined by the point, $\mathbf{P}_0$ in the direction, $\mathbf{n}$. That is, you need

$$\begin{aligned}|\mathbf{X}_0-\mathbf{P}_0||\cos\theta| &= |\mathbf{X}_0-\mathbf{P}_0|\left|\frac{(\mathbf{X}_0-\mathbf{P}_0)\cdot\mathbf{n}}{|\mathbf{X}_0-\mathbf{P}_0||\mathbf{n}|}\right|\\ &= \left|(\mathbf{X}_0-\mathbf{P}_0)\cdot\frac{\mathbf{n}}{|\mathbf{n}|}\right|\end{aligned}$$

As drawn in the picture, $|\mathbf{X}_0 - \mathbf{P}_0| \cos\theta$ will be positive but if you had $\mathbf{n}$ pointing the opposite direction this would be negative. However, either way, it's absolute value would give the right answer. This is why the absolute value is taken in the above. From this the stupid formula will follow easily. Suppose $a \neq 0$. Things work the same if b or c are not zero. Then as explained above, you can take $\mathbf{P}_0 = \left(\frac{d}{a}, 0, 0\right)$ and $\mathbf{n} = (a, b, c)$. Therefore, the above expression is

$$\left| \left(x_0 - \frac{d}{a}, y_0, z_0 \right) \cdot \frac{(a, b, c)}{\sqrt{a^2 + b^2 + c^2}} \right| = \frac{|ax_0 + by_0 + cz_0 - d|}{\sqrt{a^2 + b^2 + c^2}}$$

and it is this last expression which is the stupid formula. Here is the same example done in an ad hoc manner earlier but this time through the use of a stupid formula.

Example 4.1.15 *Find the distance from the point* $(1, 2, 3)$ *to the plane* $x - y + z = 3$.

Lets apply the stupid formula. $a = 1, b = -1, c = 1, d = 3, x_0 = 1, y_0 = 2, z_0 = 3$. Then plugging in to the formula, you get

$$\frac{|1 \times 1 + (-1) \times 2 + 1 \times 3 - 3|}{\sqrt{1 + 1 + 1}} = \frac{1}{3}\sqrt{3}$$

which gives the same answer much more easily. Those of you who expect to find the distance from a given point to a plane repeatedly, should certainly cherish and memorize this formula because it will save you lots of time. The rest of you should try to understand its derivation which is genuinely interesting and worth while. Unfortunately, finding the distance from a point to a plane is an excellent test question.

A similar formula holds for the distance from a point to a line in $\mathbb{R}^2$. Recall from high school algebra, a line can be written as

$$ax + by = c$$

Then if (x_0, y_0) is a point and you want the distance from this point to the given line, it equals

$$\frac{|ax_0 + by_0 - c|}{\sqrt{a^2 + b^2}}.$$

You should derive this stupid formula from the same geometric considerations used to get the stupid formula for a point and a plane.

Of course it all generalizes. The same reasoning will yield a stupid formula for the distance between $(y_1, \cdots, y_n)$ and the level surface, called a **hyper**[1] **plane** given by $\sum_{k=1}^{n} a_k x_k = d$. You can probably guess what it is by analogy to the above but it is better to derive it directly using the same sort of geometric reasoning just given.

[1] Words such as "hyper" give an aura of significance to things which are in reality trivial while obfuscating the real issues. They constitute an example of pretentious jargon which militates against correct understanding.

Systems Of Linear Equations 12,13 Sept.

Quiz

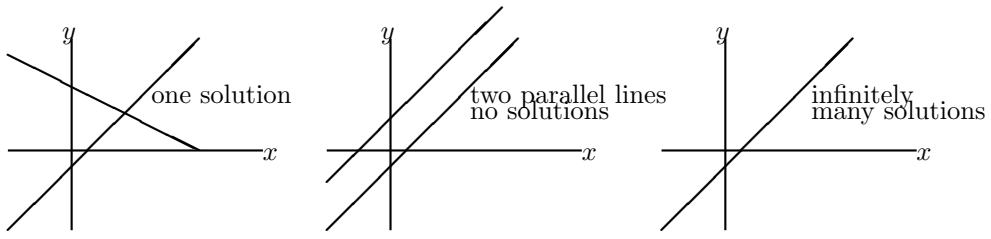
1. The intercepts of a plane are $(1,0,0), (0,2,0)$, and $(0,0,-1)$. Find the equation of the plane.

2. A plane has a normal vector $(1,2,-3)$ and contains the point $(1,1,2)$. Find the equation of the plane.

3. Find the equation of a plane which has the three points, $(1,2,1), (2,-2,1), (0,3,0)$.

5.1 Systems Of Equations, Geometric Interpretations

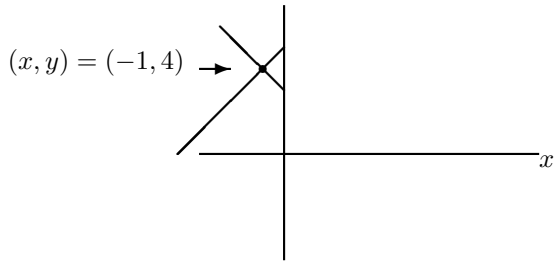
As you know, equations like $2x + 3y = 6$ can be graphed as straight lines. To find the solution to two such equations, you could graph the two straight lines and the ordered pairs identifying the point (or points) of intersection would give the x and y values of the solution to the two equations because such an ordered pair satisfies both equations. The following picture illustrates what can occur with two equations involving two variables.



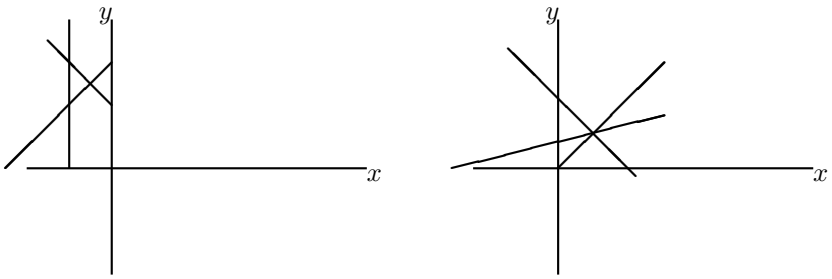
In the first example of the above picture, there is a unique point of intersection. In the second, there are no points of intersection. The other thing which can occur is that the two lines are really the same line. For example, $x + y = 1$ and $2x + 2y = 2$ are relations which when graphed yield the 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 5.1.1 *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.

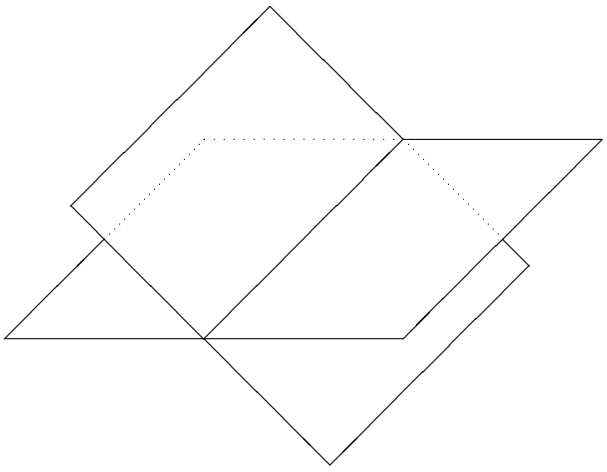


Example 5.1.2 *You can also imagine other situations such as the case of three intersecting lines having no common point of intersection or three intersecting lines which do intersect at a single point as illustrated in the following picture.*



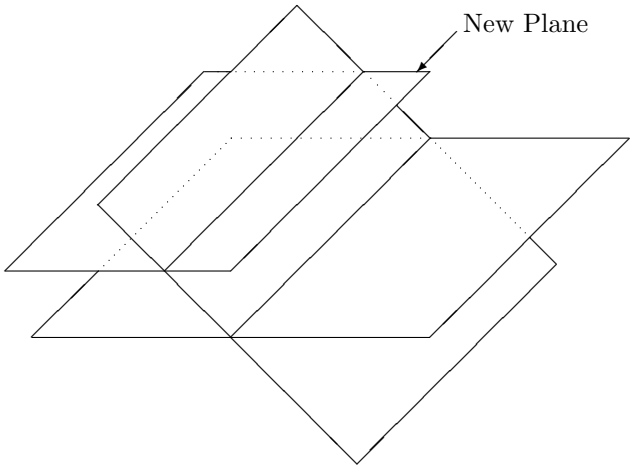
In the case of the first picture above, there would be no solution to the three equations whose graphs are the given lines. In the case of the second picture there is a solution to the three equations whose graphs are the given lines.

The points, (x, y, z) satisfying an equation in three variables like $2x + 4y - 5z = 8$ form a plane in three dimensions and geometrically, when you solve systems of equations involving three variables, you are taking intersections of planes. Consider the following picture involving two planes.



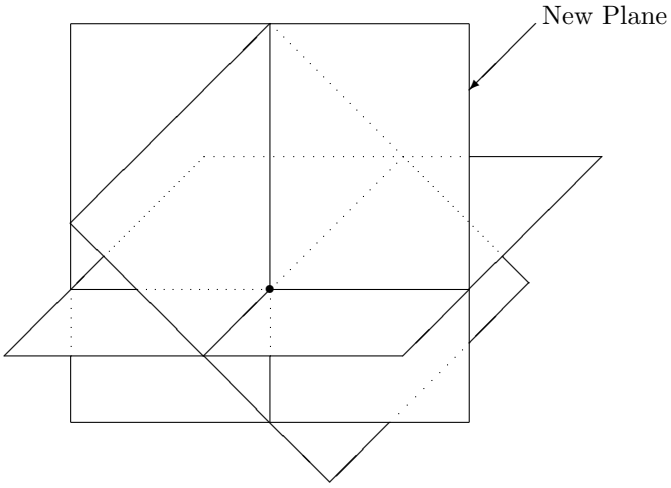
Notice how these two planes intersect in a line. It could also happen the two planes could fail to intersect.

Now imagine a third plane. One thing that could happen is this third plane could have an intersection with one of the first planes which results in a line which fails to intersect the first line as illustrated in the following picture.

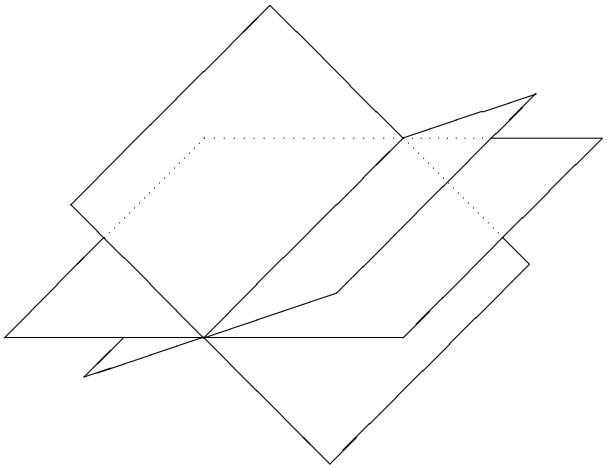


Thus there is no point which lies in all three planes. The picture illustrates the situation in which the line of intersection of the new plane with one of the original planes forms a line parallel to the line of intersection of the first two planes. However, in three dimensions, it is possible for two lines to fail to intersect even though they are not parallel. Such lines are called **skew lines.** You might consider whether there exist two skew lines, each of which is the intersection of a pair of planes selected from a set of exactly three planes such that there is no point of intersection between the three planes. You can also see that if you tilt one of the planes you could obtain every pair of planes having a nonempty intersection in a line and yet there may be no point in the intersection of all three.

It could happen also that the three planes could intersect in a single point as shown in the following picture.



In this case, the three planes have a single point of intersection. The three planes could also intersect in a line.



Thus in the case of three equations having three variables, the planes determined by these equations could intersect in a single point, a line, or even fail to intersect at all. You see that in three dimensions there are many possibilities. If you want to waste some time, you can try to imagine all the things which could happen but this will not help for dimensions higher than 3 which is where many of the important applications lie.

In higher dimensions it is customary to refer to the set of points described by relations like $x + y - 2z + 4w = 8$ as **hyper-planes**.[1] 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. The only rational and useful way to deal with this subject is through the use of algebra. Indeed, a major reason for studying mathematics is to obtain freedom from always having to draw a picture in order to do a computation or find out something important.

5.2 Systems Of Equations, Algebraic Procedures

5.2.1 Elementary Operations

Definition 5.2.1 *A system of linear equations is a set of p equations for the n variables, $x_1, \cdots, x_n$ which is of the form*

$$\sum_{k=1}^{n} a_{mk} x_k = d_m, m = 1, 2, \cdots, p$$

Written less compactly it is a set of equations of the following form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= d_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= d_2 \\ &\vdots \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pn}x_n &= d_p \end{aligned}$$

[1] The evocative semi word, "hyper" conveys absolutely no meaning but is traditional usage which makes the terminology sound more impressive than something like long wide comparatively flat thing which does convey some meaning. However, in such cases as these pretentious jargon is nearly always preferred. Later we will discuss some terms which are not just evocative but yield real understanding.

The problem is to find the values of $x_1, x_2 \cdots, x_n$ which satisfy all p equations. This is called the ***solution set*** *of the system of equations. In other words, $(a_1, \cdots, a_n)$ is in the solution set of the system of equations if when you plug a_1 in place of x_1, a_2 in place of x_2 etc., each equation in the system is satisfied.*

Consider the following example.

Example 5.2.2 *Find x and y such that*

$$x + y = 7 \text{ and } 2x - y = 8. \tag{5.1}$$

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 5.2.3 ***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 5.2.4 *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 5.2.3 do not change the set of solutions to the system of linear equations.

Theorem 5.2.5 *Suppose you have two equations, involving the variables, $(x_1, \cdots, x_n)$*

$$E_1 = f_1, E_2 = f_2 \tag{5.2}$$

where E_1 and E_2 are expressions involving the variables and f_1 and f_2 are constants. (In the above example there are only two variables, x and y and $E_1 = x + y$ while $E_2 = 2x - y$.) Then the system $E_1 = f_1, E_2 = f_2$ has the same solution set as

$$E_1 = f_1,\ E_2 + aE_1 = f_2 + af_1. \tag{5.3}$$

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 5.2.6 *Find the solutions to the system,*

$$\begin{array}{c} x + 3y + 6z = 25 \\ 2x + 7y + 14z = 58 \\ 2y + 5z = 19 \end{array} \tag{5.4}$$

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

$$\begin{array}{c} x + 3y + 6z = 25 \\ y + 2z = 8 \\ 2y + 5z = 19 \end{array} \tag{5.5}$$

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

$$\begin{array}{c} x + 3y + 6z = 25 \\ y + 2z = 8 \\ z = 3 \end{array} \tag{5.6}$$

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 5.6 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.

5.2.2 Gauss Elimination

A less cumbersome way to represent a linear system is to write it as an **augmented matrix.** For example the linear system, 5.4 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 5.4 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 5.5. 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 5.6. By back substitution you obtain the solution $x = 1, y = 6$, and $z = 3$.

In general a linear system is of the form

$$\begin{array}{c} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \end{array}, \tag{5.7}$$

where the x_i are variables and the a_{ij} and b_i are constants. This system can be represented by the augmented matrix,

$$\left(\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right). \tag{5.8}$$

Changes to the system of equations in 5.7 as a result of an elementary operations translate into changes of the augmented matrix resulting from a row operation. Note that Theorem 5.2.5 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 5.2.7 *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.

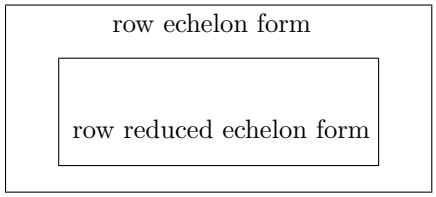
Definition 5.2.8 *An augmented matrix is in* ***echelon form*** *also called* ***row echelon form*** *if*

1. *All nonzero rows are above any rows of zeros.*
2. *Each leading entry of a row is in a column to the right of the leading entries of any rows above it.*

Definition 5.2.9 *An augmented matrix is in* ***row reduced echelon form*** *if*

1. *All nonzero rows are above any rows of zeros.*
2. *Each leading entry of a row is in a column to the right of the leading entries of any rows above it.*
3. *All entries in a column above and below a leading entry are zero.*
4. *Each leading entry is a 1, the only nonzero entry in its column.*

The relation between these two definitions is as described in the following picture.



Thus if the matrix is in row reduced echelon form, it is in row echelon form but not necessarily the other way around. You can usually find the solution to a system of equations by row reducing to row echelon form. You typically don't have to go all the way to the row reduced echelon form but the row reduced echelon form is very important because, unlike a row echelon form, it is unique. It is also easier to use in the case where the system of equations has an infinite solution set.

Example 5.2.10 *Here are some augmented matrices which are in row reduced echelon form.*

$$\left(\begin{array}{ccccc|c} 1 & 0 & 0 & 5 & 8 & 0 \\ 0 & 0 & 1 & 2 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

Example 5.2.11 *Here are augmented matrices in echelon form which are not in row reduced echelon form but which are in echelon form.*

$$\left(\begin{array}{ccccc|c} 1 & 0 & 6 & 5 & 8 & 2 \\ 0 & 0 & 2 & 2 & 7 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{ccc|c} 1 & 3 & 5 & 4 \\ 0 & 2 & 0 & 7 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Example 5.2.12 *Here are some augmented matrices which are not in echelon form.*

$$\left(\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 1 & 2 & 3 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{cc|c} 1 & 2 & 3 \\ 2 & 4 & -6 \\ 4 & 0 & 7 \end{array}\right), \left(\begin{array}{ccc|c} 0 & 2 & 3 & 3 \\ 1 & 5 & 0 & 2 \\ 7 & 5 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right).$$

Definition 5.2.13 *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 5.2.14 *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 following is the algorithm for obtaining a matrix which is in row reduced echelon form.

Algorithm 5.2.15

This algorithm tells how to start with a matrix and do row operations on it in such a way as to end up with a matrix in row reduced echelon form.

1. Find the first nonzero column from the left. This is the first pivot column. The position at the top of the first pivot column is the first pivot position. Switch rows if necessary to place a nonzero number in the first pivot position.
2. Use row operations to zero out the entries below the first pivot position.
3. Ignore the row containing the most recent pivot position identified and the rows above it. Repeat steps 1 and 2 to the remaining submatrix, the rectangular array of numbers obtained from the original matrix by deleting the rows you just ignored. Repeat the process until there are no more rows to modify. The matrix will then be in echelon form.
4. Moving from right to left, use the nonzero elements in the pivot positions to zero out the elements in the pivot columns which are above the pivots.
5. Divide each nonzero row by the value of the leading entry. The result will be a matrix in row reduced echelon form.

This row reduction procedure applies to both augmented matrices and non augmented matrices. There is nothing special about the augmented column with respect to the row reduction procedure.

Example 5.2.16 *Here is a matrix.*

$$\left(\begin{array}{ccccc} 0 & 0 & 2 & 3 & 2 \\ 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \end{array}\right)$$

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.

$$\left(\begin{array}{ccccc} 0 & 1 & 1 & 4 & 3 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \end{array}\right)$$

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.

$$\begin{pmatrix} 0 & 3 & 3 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Take $-1/2$ times the second row and add to the first.

$$\begin{pmatrix} 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Finally, divide by the value of the leading entries in the nonzero rows.

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

The above algorithm is the way a computer would obtain a reduced echelon form for a given matrix. It is not necessary for you to pretend you are a computer but if you like to do so, the algorithm described above will work. The main idea is to do row operations in such a way as to end up with a matrix in echelon form or row reduced echelon form because when this has been done, the resulting augmented matrix will allow you to describe the solutions to the linear system of equations in a meaningful way.

Example 5.2.17 *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 5.2.18 *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. [2]Therefore, the solution set of this system is $x = 3 + 5t, y = 10t$, and $z = t$ where t

[2]In this context t is called a **parameter.**

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 5.2.18 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 5.2.18 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 5.2.19 *Find the solution to the system*

$$\begin{array}{c} x+2y-z+w=3 \\ x+y-z+w=1 \\ x+3y-z+w=5 \end{array}$$

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

$$\left(\begin{array}{cccc|c} 1 & 2 & -1 & 1 & 3 \\ 0 & -1 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right) \tag{5.9}$$

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

$$\left(\begin{array}{c} x \\ y \\ z \\ w \end{array}\right) = \left(\begin{array}{c} -1+s-t \\ 2 \\ s \\ t \end{array}\right). \tag{5.10}$$

This is another example of a system which has an infinite solution set but this time the solution set depends on two parameters, not one. Most people find it less confusing in the case of an infinite solution set to first place the augmented matrix in row reduced echelon form rather than just echelon form before seeking to write down the description of the solution. In the above, this means we don't stop with the echelon form 5.9. 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 5.10.

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

$$\left(\begin{array}{ccc|c} 1 & 2 & | & 3 \\ 0 & 4 & | & -2 \\ 0 & 0 & | & 1 \end{array}\right)$$

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 row reduced echelon form. In every case, the process yields a straightforward way to describe the solutions to the linear system. As indicated above, you are probably less likely to become confused if you place the augmented matrix in row reduced echelon form rather than just echelon form.

In summary,

Definition 5.2.20 *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.

Definition 5.2.21 *A system of linear equations is called* ***consistent*** *if there exists a solution. It is called* ***inconsistent*** *if there is no solution.*

These are reasonable words to describe the situations of having or not having a solution. If you think of each equation as a condition which must be satisfied by the variables, consistent would mean there is some choice of variables which can satisfy all the conditions. Inconsistent means there is no choice of the variables which can satisfy each of the conditions.

5.3 The Rank Of A Matrix 14 Sept.

The notion of an augmented matrix was used to solve systems of equations. In general, a matrix is simply a rectangular array of numbers.

Definition 5.3.1 *A matrix, A is called an $m \times n$ matrix if it has m rows and n columns.*

Example 5.3.2 *The matrix,*

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

is a 3×2 matrix because it has two columns, (These stand upright.) and three rows.

Corresponding to such a rectangular array of numbers, there is a row reduced echelon form discussed above. The following theorem is of fundamental significance.

Theorem 5.3.3 *Given an $m \times n$ matrix, the row reduced echelon form is unique.*

This is a remarkable theorem because there are many ways to do row operations and eventually end up with something in row reduced echelon form. It is remarkable that you always get the same thing. Now it is easy to describe the rank of a matrix.

Definition 5.3.4 *The rank of a matrix, A equals the number of nonzero rows in its row reduced echelon form. This is the same as the number of pivot columns.*

Example 5.3.5 *Find the rank of the matrix,*

$$A = \begin{pmatrix} 1 & 2 & 3 & 1 \\ 0 & 2 & 1 & 1 \\ 1 & 4 & 4 & 2 \end{pmatrix}$$

To find the rank, you obtain the row reduced echelon form and count the number of nonzero rows or equivalently the number of pivot columns. First take -1 times the top row and add to the bottom row. This yields

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

Now add -1 times the second row to the bottom. This yields

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

Now take -1 times the second row and add to the top.

$$\begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Finally, multiply the second row by $1/2$ to get

$$\begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

which is in row reduced echelon form. The rank of this matrix is therefore 2.

Note that from the process used to obtain the row reduced echelon form, once you have obtained an echelon form, you know the correct number of rows in the final result. Thus you can simply take the number of nonzero rows in an echelon form and this will be the rank. Note also that the rank is the number of pivot columns. In this case the pivot columns are the first two.

Definition 5.3.6 *A **homogeneous system of linear equations** is one with augmented matrix of the form*

$$\left(\begin{array}{c|c} A & \mathbf{0} \end{array} \right)$$

where $\mathbf{0}$ *is a column of zeros and* A *is an* $m \times n$ *matrix.*

Example 5.3.7 *An example of a homogeneous system of equations is* $x+y=0, 3x-y=0$. *It has augmented matrix,*

$$\left(\begin{array}{cc|c} 1 & 1 & 0 \\ 3 & -1 & 0 \end{array} \right).$$

The nice thing about homogeneous systems is that they are always consistent. Simply let all the variables equal zero and you obtain a solution. However, there may be other solutions besides this one. This is related to the concept of rank and free variables.

Theorem 5.3.8 *Let* A *be an* $m \times n$ *matrix. Form the augmented matrix,*

$$\left(\begin{array}{c|c} A & \mathbf{0} \end{array} \right)$$

where $\mathbf{0}$ *is the column of zeros. Thus* A *is the coefficient matrix of a system of linear equations with* n *variables. Then the number of free variables* $= n - \operatorname{rank}(A)$.

Proof: The basic variables correspond to the pivot columns of A and the free variables correspond to the other columns. However, the rank of A equals the number of pivot columns.

As a corollary here is a theorem which is called the Rank theorem.

Corollary 5.3.9 *(Rank Theorem) Let* A *be an* $m \times n$ *matrix. Form the augmented matrix,*

$$\left(\begin{array}{c|c} A & \mathbf{b} \end{array} \right)$$

where $\mathbf{b}$ *is an* $m \times 1$ *column. Thus* A *is the coefficient matrix of a system of linear equations with* n *variables. Then if the system of equations represented by the above augmented matrix is consistent, number of free variables* $= n - \operatorname{rank}(A)$.

Proof: Since the equations represented by the above augmented matrix are consistent, the same argument as in Theorem 5.3.8 holds. The leading entry in the last nonzero row cannot be in the last column because if it were, then the system would fail to be consistent.

5.4 Theory Of Row Reduced Echelon Form*

This material will be done much more easily later after the introduction of elementary matrices. You can wait to read it till then. However, if you wish to understand what is going on right now, I am giving an explanation. First recall the row operations.

Definition 5.4.1 *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.*

In rough terms, the following lemma states that linear relationships between columns in a matrix are preserved by row operations.

Definition 5.4.2 *The vector,* $\mathbf{u}$ *is a* ***linear combination*** *of the vectors,* $\mathbf{v}_1, \cdots, \mathbf{v}_m$ *if there exist scalars,* $c_1, \cdots, c_m$ *such that*

$$\begin{aligned} \mathbf{u} &= c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_m\mathbf{v}_m \\ &= \sum_{k=1}^{m} c_k\mathbf{v}_k. \end{aligned}$$

Example 5.4.3

$$3\begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix} + 5\begin{pmatrix} 1 \\ -2 \\ 4 \end{pmatrix} + (-2)\begin{pmatrix} 5 \\ 3 \\ 7 \end{pmatrix} = \begin{pmatrix} 1 \\ -7 \\ 9 \end{pmatrix}$$

Thus $\begin{pmatrix} 1 \\ -7 \\ 9 \end{pmatrix}$ *is a linear combination of the vectors,* $\begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -2 \\ 4 \end{pmatrix}$, *and* $\begin{pmatrix} 5 \\ 3 \\ 7 \end{pmatrix}$. *In this case the scalars are* $3, 5$, *and* -2.

Definition 5.4.4 *When dealing with an* $m \times n$ *matrix,* A*, the element in the* i^{th} *row and the* j^{th} *column is denoted as* A_{ij}*. Thus the* j^{th} *column is*

$$\begin{pmatrix} A_{1j} \\ A_{2j} \\ A_{3j} \\ \vdots \\ A_{mj} \end{pmatrix}$$

Lemma 5.4.5 *Let* B *and* A *be two* $m \times n$ *matrices and suppose* B *results from a row operation applied to* A*. Then the* k^{th} *column of* B *is a linear combination of the* $i_1, \cdots, i_r$ *columns of* B *if and only if the* k^{th} *column of* A *is a linear combination of the* $i_1, \cdots, i_r$ *columns of* A*. Furthermore, the scalars in the linear combination are the same. (The linear relationship between the* k^{th} *column of* A *and the* $i_1, \cdots, i_r$ *columns of* A *is the same as the linear relationship between the* k^{th} *column of* B *and the* $i_1, \cdots, i_r$ *columns of* B*.)*

Proof: This is obvious in the case of the first two row operations and a little less obvious in the case of the third. Therefore, consider the third. Suppose the s^{th} row of B equals the s^{th} row of A added to c times the q^{th} row of A. Therefore,

$$B_{ij} = A_{ij} \text{ if } i \neq s, B_{sj} = A_{sj} + cA_{qj}.$$

The assumption about the k^{th} column of B is equivalent to saying that for each p,

$$B_{pk} = \sum_{j=1}^{r} \alpha_j B_{pi_j}. \tag{5.11}$$

For $p \neq s$, this is equivalent to saying

$$A_{pk} = \sum_{j=1}^{r} \alpha_j A_{pi_j} \tag{5.12}$$

because for these values of p, $B_{pj} = A_{pj}$. For $p = s$, this is equivalent to saying

$$A_{sk} + cA_{qk} = \sum_{j=1}^{r} \alpha_j \left(A_{si_j} + cA_{qi_j} \right). \tag{5.13}$$

but from 5.12, applied to $p = q$,

$$cA_{qk} = c \sum_{j=1}^{r} \alpha_j A_{qi_j}$$

and so from 5.13, it follows 5.11 is equivalent to 5.12 for all p, including $p = s$. This proves the lemma.

Now I will present a review of the row reduced echelon form. It is convenient to describe it slightly differently to use Lemma 5.4.5.

Definition 5.4.6 *Let $\mathbf{e}_i$ denote the column vector which has all zero entries except for the i^{th} slot which is one. An $m \times n$ matrix is said to be in row reduced echelon form if, in viewing succesive columns from left to right, the first nonzero column encountered is $\mathbf{e}_1$ and if you have encountered $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$, the next column is either $\mathbf{e}_{k+1}$ or is a linear combination of the vectors, $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$.*

Theorem 5.4.7 *Let A be an $m \times n$ matrix. Then A has a row reduced echelon form determined by a simple process.*

Proof: Viewing the columns of A from left to right take the first nonzero column. Pick a nonzero entry in this column and switch the row containing this entry with the top row of A. Now divide this new top row by the value of this nonzero entry to get a 1 in this position and then use row operations to make all entries below this element equal to zero. Thus the first nonzero column is now $\mathbf{e}_1$. Denote the resulting matrix by A_1. Consider the submatrix of A_1 to the right of this column and below the first row. Do exactly the same thing for it that was done for A. This time the $\mathbf{e}_1$ will refer to $\mathbb{R}^{m-1}$. Use this 1 and row operations to zero out every element above it in the rows of A_1. Call the resulting matrix, A_2. Thus A_2 satisfies the conditions of the above definition up to the column just encountered. Continue this way till every column has been dealt with and the result must be in row reduced echelon form.

The following diagram illustrates the above procedure. Say the matrix looked something like the following.

$$\begin{pmatrix} 0 & * & * & * & * & * & * \\ 0 & * & * & * & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & * & * & * & * & * & * \end{pmatrix}$$

First step would yield something like

$$\begin{pmatrix} 0 & 1 & * & * & * & * & * \\ 0 & 0 & * & * & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & * & * & * & * & * \end{pmatrix}$$

For the second step you look at the lower right corner as described,

$$\begin{pmatrix} * & * & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ * & * & * & * & * \end{pmatrix}$$

and if the first column consists of all zeros but the next one is not all zeros, you would get something like this.

$$\begin{pmatrix} 0 & 1 & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & * & * & * \end{pmatrix}$$

Thus, after zeroing out the term in the top row above the 1, you get the following for the next step in the computation of the row reduced echelon form for the original matrix.

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

Next you look at the lower right matrix below the top two rows and to the right of the first four columns and repeat the process.

Definition 5.4.8 *The first pivot column of A is the first nonzero column of A. The next pivot column is the first column after this which becomes $\mathbf{e}_2$ in the row reduced echelon form. The third is the next column which becomes $\mathbf{e}_3$ in the row reduced echelon form and so forth.*

There are three choices for row operations at each step in the above theorem. A natural question is whether the same row reduced echelon matrix always results in the end from following the above algorithm applied in any way. The next corollary says this is the case.

Definition 5.4.9 *Two matrices are said to be* ***row equivalent*** *if one can be obtained from the other by a sequence of row operations.*

It has been shown above that every matrix is row equivalent to one which is in row reduced echelon form.

Corollary 5.4.10 *The row reduced echelon form is unique. That is if B, C are two matrices in row reduced echelon form and both are row equivalent to A, then $B = C$.*

Proof: Suppose B and C are both row reduced echelon forms for the matrix, A. Then they clearly have the same zero columns since row operations leave zero columns unchanged. If B has the sequence $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_r$ occuring for the first time in the positions, $i_1, i_2, \cdots, i_r$ the description of the row reduced echelon form means that if $\mathbf{b}_k$ is the k^{th} column of B such that $i_{j-1} < k < i_j$ then $\mathbf{b}_k$ is a linear combination of the columns in positions $i_1, i_2, \cdots, i_r$. By Lemma 5.4.5 the same is true for $\mathbf{c}_k$, the k^{th} column of C. Therefore, $\mathbf{c}_k$ is not equal to $\mathbf{e}_j$ for any j because $\mathbf{e}_j$ is not obtained as a linear combinations of the $\mathbf{e}_i$ for $i < j$. It follows the $\mathbf{e}_j$ for C can only occur in positions $i_1, i_2, \cdots, i_r$. Furthermore, position i_j in C must contain $\mathbf{e}_j$ because if not, then $\mathbf{c}_{i_j}$ would be a linear combination of $\mathbf{e}_1, \cdots, \mathbf{e}_{j-1}$ in C but not in B, thus contradicting Lemma 5.4.5. Therefore, both B and C have the sequence $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_r$ occuring for the first time in the positions, $i_1, i_2, \cdots, i_r$. By Lemma 5.4.5, the columns between the i_k and i_{k+1} position are linear combinations involving the same scalars of the columns in the $i_1, \cdots, i_k$ position. This is equivalent to the assertion that each of these columns is identical and this proves the corollary.

This suggests that to find the rank of a matrix, one should do row operations until a matrix is obtained in which its rank is obvious.

5.4.1 Exercises With Answers

1. Find the distance from the point, $(1, 2, 1)$ to the plane $3x + y - z = 7$.

   You can use the stupid formula for this.

   $$\frac{|3 + 2 - 1 - 7|}{\sqrt{9 + 1 + 1}} = \frac{3}{11}\sqrt{11}$$

2. Find the cosine of the angle between the planes $x - y + z = 7$ and $2x + y - 3z = 4$.

   You just need to consider the normal vectors which are $(1, -1, 1)$ and $(2, 1, -3)$. Then the cosine of the angle desired is

   $$\cos\theta = \left| \frac{(2, 1, -3) \cdot (1, -1, 1)}{\sqrt{1 + 1 + 1}\sqrt{4 + 1 + 9}} \right| = \frac{1}{21}\sqrt{3}\sqrt{14}$$

3. Here are vector equations for two lines. $(x, y, z) = (1, 2, 0) + t(2, 1, 1)$ and $(x, y, z) = (3, 0, 1) + t(1, -2, 1)$. The angle between the direction vectors is not 0 or π and so the lines are not parallel. If they were two lines in $\mathbb{R}^2$, this means they would need to intersect. However, these two lines do not intersect. If they did, there would exist s, t such that

   $$(1, 2, 0) + t(2, 1, 1) = (3, 0, 1) + s(1, -2, 1)$$

   and this would require the following system of equations would need to hold.

   $$\begin{aligned} 1 + 2t &= 3 + s \\ 2 + t &= -2s \\ t &= 1 + s \end{aligned}$$

   The augmented matrix for this system is

   $$\left(\begin{array}{ccc|c} 2 & -1 & | & 2 \\ 1 & 2 & | & -2 \\ 1 & -1 & | & 1 \end{array} \right)$$

The row reduced echelon form is

$$\left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right)$$

and so there is no solution. These lines are called **skew lines**. Imagine two airplanes, one going from South to North and the other going from East to West. The first travels at 40000 feet and the second at 35000 feet. Their paths never cross. Of course the extra dimension is not present in two dimensions and so their paths would cross if they were moving in a plane. Note also that to consider the question whether the lines intersect, you must look at possibly different values for the parameters.

4. Let two skew lines be given in Problem 3. Find two parallel planes which contain the two lines.

   This is easy if you can find the normal vector of the two planes. To say the planes are parallel requires them to have the same normal vector. The two lines were $(x,y,z) = (1,2,0) + t(2,1,1)$ and $(x,y,z) = (3,0,1) + t(1,-2,1)$. Therefore, the normal vector needs to be perpendicular to both direction vectors. You need $\mathbf{n} = (2,1,1) \times (1,-2,1) = (3,-1,-5)$. Now the equation of the first plane is

   $$(3,-1,-5) \cdot (x-1, y-2, z) = 0$$

   and the equation of the second plane is

   $$(3,-1,-5) \cdot (x-3, y, z-1) = 0$$

   The two planes are therefore, $3x - y - 5z = 1$ and $3x - y - 5z = 4$. You see these are parallel planes because they have the same normal vector and the first contains the first line while the second contains the second line.

5. 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?

   $$\left(\begin{array}{ccccc|c} \blacksquare & * & * & * & * & * \\ 0 & \blacksquare & * & * & 0 & * \\ 0 & 0 & \blacksquare & * & * & \blacksquare \\ 0 & 0 & 0 & 0 & \blacksquare & * \end{array}\right)$$

   In this case the system is consistent and there is an infinite set of solutions. To see it is consistent, the bottom equation would yield a unique solution for x_5. Then letting $x_4 = t$, and substituting in to the other equations, beginning with the equation determined by the third row and then proceding up to the next row followed by the first row, you get a solution for each value of t. There is a free variable which comes from the fourth column which is why you can say $x_4 = t$. Therefore, the solution is infinite.

6. 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?

   $$\left(\begin{array}{ccc|c} \blacksquare & * & * & * \\ 0 & 0 & \blacksquare & \blacksquare \\ 0 & 0 & * & 0 \end{array}\right)$$

In this case there is no solution because you could use a row operation to place a 0 in the third row and third column position, like this:

$$\left(\begin{array}{ccc|c} \blacksquare & * & * & * \\ 0 & 0 & \blacksquare & \blacksquare \\ 0 & 0 & 0 & \blacksquare \end{array}\right)$$

This would give a row of zeros equal to something nonzero.

7. Find h such that

$$\left(\begin{array}{cc|c} 1 & h & 4 \\ 3 & 7 & 7 \end{array}\right)$$

is the augmented matrix of an inconsistent matrix.

Doing a row operation by taking -3 times the top row and adding to the bottom, this gives

$$\left(\begin{array}{cc|c} 1 & h & 4 \\ 0 & 7-3h & 7-12 \end{array}\right).$$

The system will be inconsistent if $7-3h=0$ or in other words, $h=7/3$.

8. Determine if the system is consistent.

$$\begin{array}{c} x+2y+3z-w=2 \\ x-y+2z+w=1 \\ 2x+3y-z=1 \\ 4x+2y+z=5 \end{array}$$

The augmented matrix is

$$\left(\begin{array}{cccc|c} 1 & 2 & 3 & -1 & 2 \\ 1 & -1 & 2 & 1 & 1 \\ 2 & 3 & -1 & 0 & 1 \\ 4 & 2 & 1 & 0 & 5 \end{array}\right)$$

A reduced echelon form for this is

$$\left(\begin{array}{cccc|c} 9 & 0 & 0 & 0 & 14 \\ 0 & 9 & 0 & 0 & -6 \\ 0 & 0 & 9 & 0 & 1 \\ 0 & 0 & 0 & 9 & -13 \end{array}\right).$$

Therefore, there is a unique solution. In particular the system is consistent.

9. Find the point, (x_1, y_1) which lies on both lines, $5x+3y=1$ and $4x-y=3$.

You solve the system of equations whose augmented matrix is

$$\left(\begin{array}{cc|c} 5 & 3 & 1 \\ 4 & -1 & 3 \end{array}\right)$$

A reduced echelon form is

$$\left(\begin{array}{ccc} 17 & 0 & 10 \\ 0 & 17 & -11 \end{array}\right)$$

and so the solution is $x=17/10$ and $y=-11/17$.

10. Do the three lines, $3x+2y=1, 2x-y=1$, 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.

    This is asking for the solution to the three equations shown. The augmented matrix is
    $$\left(\begin{array}{cc|c} 3 & 2 & 1 \\ 2 & -1 & 1 \\ 4 & 3 & 3 \end{array}\right)$$
    A reduced echelon form is
    $$\left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right)$$
    and this would require $0x+0y=1$ which is impossible so there is no solution to this system of equations and hence no point on each of the three lines.

11. Find the general solution of the system whose augmented matrix is
    $$\left(\begin{array}{ccc|c} 1 & 2 & 0 & 2 \\ 1 & 1 & 4 & 2 \\ 2 & 3 & 4 & 4 \end{array}\right).$$
    A reduced echelon form for the matrix is
    $$\left(\begin{array}{cccc} 1 & 0 & 8 & 2 \\ 0 & 1 & -4 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right).$$
    Therefore, $y=4z$ and $x=2-8z$. Apparently z can equal anything so we let $z=t$ and then the solution is
    $$x=2-8t, y=4t, z=t.$$

12. Find the point, (x_1, y_1) which lies on both lines, $x+2y=1$ and $3x-y=3$.

    The solution is $y=0$ and $x=1$.

13. Find the point of intersection of the two lines $x+y=3$ and $x+2y=1$.

    The solution is $(5,-2)$.

14. 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.

    To solve this set up the augmented matrix and go to work on it. The augmented matrix is
    $$\left(\begin{array}{cc|c} 1 & 2 & 1 \\ 2 & -1 & 1 \\ 4 & 3 & 3 \end{array}\right)$$
    A reduced echelon matrix for this is
    $$\left(\begin{array}{cc|c} 1 & 0 & \frac{3}{5} \\ 0 & 1 & \frac{1}{5} \\ 0 & 0 & 0 \end{array}\right)$$
    Therefore, there is a point in the intersection of these and it is $y=1/5$ and $x=3/5$. Thus the point is $(3/5, 1/5)$.

15. Do the three planes, $x + 2y - 3z = 2$, $x + y + z = 1$, 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.

    You need to find (x, y, z) which solves each equation. The augmented matrix is

$$\left(\begin{array}{ccc|c} 1 & 2 & -3 & 2 \\ 1 & 1 & 1 & 1 \\ 3 & 2 & 2 & 0 \end{array}\right)$$

    A reduced echelon form for the matrix is

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & -2 \\ 0 & 1 & 0 & \frac{13}{5} \\ 0 & 0 & 1 & \frac{2}{5} \end{array}\right)$$

    and so you should let $(x, y, z) = (-2, 13/5, 2/5)$.

16. 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?

$$\left(\begin{array}{ccccc|c} \blacksquare & * & * & * & * & * \\ 0 & \blacksquare & * & * & 0 & * \\ 0 & 0 & \blacksquare & * & * & * \\ 0 & 0 & 0 & 0 & \blacksquare & * \end{array}\right)$$

    You could do another set of row operations and reduce the matrix to one of the form

$$\left(\begin{array}{ccccc|c} \blacksquare & * & * & * & 0 & * \\ 0 & \blacksquare & * & * & 0 & * \\ 0 & 0 & \blacksquare & * & 0 & * \\ 0 & 0 & 0 & 0 & \blacksquare & * \end{array}\right)$$

    It follows there exists a solution but the solution is not unique because x_4 is a free variable. You can pick it to be anything you like and the system will yield values for the other variables.

17. 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?

$$\left(\begin{array}{ccc|c} \blacksquare & * & * & * \\ 0 & \blacksquare & * & * \\ 0 & 0 & \blacksquare & * \end{array}\right)$$

    In this case there is a unique solution to the system. To see this, you could do more row operations and reduce this to something of the form

$$\left(\begin{array}{ccc|c} \blacksquare & 0 & 0 & * \\ 0 & \blacksquare & 0 & * \\ 0 & 0 & \blacksquare & * \end{array}\right).$$

18. 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?

$$\left(\begin{array}{ccccc|c} \blacksquare & * & * & * & * & * \\ 0 & \blacksquare & 0 & * & 0 & * \\ 0 & 0 & 0 & \blacksquare & * & * \\ 0 & 0 & 0 & 0 & \blacksquare & * \end{array}\right)$$

In this case, you could do more row operations and get something of the form

$$\left(\begin{array}{ccccc|c} \blacksquare & 0 & * & 0 & 0 & * \\ 0 & \blacksquare & 0 & 0 & 0 & * \\ 0 & 0 & 0 & \blacksquare & 0 & * \\ 0 & 0 & 0 & 0 & \blacksquare & * \end{array}\right)$$

Now you can determine the answer.

19. 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.

Take -3 times the top row and add to 2 times the bottom. This yields

$$\left(\begin{array}{cc|c} 2 & h & 4 \\ 0 & 12-3h & 2 \end{array}\right)$$

Now if $h = 4$ the system is inconsistent because it would have the bottom row equal to $\left(\begin{array}{cc|c} 0 & 0 & 2 \end{array}\right)$.

20. 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.

$$\left(\begin{array}{cc|c} 1 & h & 2 \\ 2 & 4 & k \end{array}\right).$$

If $h \neq 2$ then k can be anything and the system represented by the augmented matrix will have a unique solution. Suppose then that $h = 2$. Then taking -2 times the top row and adding to the bottom row gives

$$\left(\begin{array}{cc|c} 1 & 2 & 2 \\ 0 & 0 & k-4 \end{array}\right)$$

If $k \neq 4$ there is no solution. However, if $k = 4$ you are left with the single equation, $x + 2y = 2$ and there are infinitely many solutions to this. In fact anything of the form $(2 - 2y, y)$ will work just fine.

21. Determine if the system is consistent.

$$\begin{array}{c} x + 2y + z - w = 2 \\ x - y + z + w = 1 \\ 2x + y - z = 1 \\ 4x + 2y + z = 5 \end{array}$$

This system is inconsistent. To see this, write the augmented matrix and do row operations. The augmented matrix is

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & -1 & 2 \\ 1 & -1 & 1 & 1 & 1 \\ 2 & 1 & -1 & 0 & 1 \\ 4 & 2 & 1 & 0 & 5 \end{array}\right)$$

A reduced echelon form for this matrix is

$$\left(\begin{array}{cccc|c} 1 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 1 & 0 & -\frac{2}{3} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

and the bottom row shows there is no solution.

22. Find the general solution of the system whose augmented matrix is

$$\left(\begin{array}{ccc|c} 1 & 2 & 0 & 2 \\ 1 & 3 & 4 & 2 \\ 1 & 0 & 2 & 1 \end{array}\right)$$

A reduced echelon form for this matrix is

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & \frac{6}{5} \\ 0 & 1 & 0 & \frac{2}{5} \\ 0 & 0 & 1 & -\frac{1}{10} \end{array}\right)$$

and so the solution is unique and is $z = -1/10, y = 2/5$, and $x = 6/5$.

23. Find the general solution of the system whose augmented matrix is

$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & 5 \\ 1 & 0 & 3 & 2 \end{array}\right).$$

A reduced echelon form for this matrix is

$$\left(\begin{array}{ccc|c} 1 & 0 & 3 & 2 \\ 0 & 1 & -3 & 3 \end{array}\right)$$

and so the general solution is of the form $y = 3 + 3z, x = 2 - 3z$ with z arbitrary.

24. Find the general solution of the system whose augmented matrix is

$$\left(\begin{array}{ccccc|c} 1 & 0 & 2 & 1 & 1 & 3 \\ 0 & 1 & 0 & 4 & 2 & 1 \\ 2 & 2 & 0 & 0 & 1 & 3 \\ 1 & 0 & 1 & 0 & 2 & 2 \end{array}\right).$$

You do the usual thing, row operations on the matrix to obtain a reduced echelon form. A reduced echelon form is

$$\left(\begin{array}{ccccc|c} 1 & 0 & 0 & 0 & \frac{9}{2} & \frac{7}{6} \\ 0 & 1 & 0 & 0 & -4 & \frac{1}{3} \\ 0 & 0 & 1 & 0 & -\frac{5}{2} & \frac{5}{6} \\ 0 & 0 & 0 & 1 & \frac{3}{2} & \frac{1}{6} \end{array}\right)$$

Therefore, the general solution is $x_4 = 1/6 - 3/2x_5, x_3 = 5/6 + 5/2x_5, x_2 = 1/3 + 4x_5$, and $x_1 = 7/6 - 9/2x_5$ with x_5 arbitrary.

Part III

Linear Independence And Matrices

Outcomes

Spanning sets and Linear Independence

A. Explain what is meant by the span of a set of vectors both geometrically and algebraically.

B. Determine the span of a set of vectors. Determine if a given vector is in the span of a set of vectors.

C. Define linear independence.

D. Determine whether a set of vectors is linearly dependent or linearly independent. For sets that are linearly dependent, determine a dependence relation.

E. Prove theorems about span and linear independence.

Reading: Linear Algebra 2.3
Supplemental Problems:

A1. Study the definition of linear independence. Write it from memory.

Outcome Mapping:

A. 13-16,17

B. 1-6,7-12

C. A1

D. 22-31

E. 18-21,42-48

Spanning Sets And Linear Independence 18,19 Sept.

Quiz

1. Find a parametric equation for the line determined by the two points, $(1,2,1)$ and $(2,-1,3)$.

2. Find an equation of the plane containing the point $(0,1,0)$ and the line $(1,1,1)+t(2,-1,1)$.

3. An equation contains the point $(0,0,0)$ and is perpendicular to the vector $(1,1,1)$. Find an equation of this plane.

4. Here are some equations. Find the complete solution.

$$\begin{array}{c} x+y+4z=1 \\ -2x+y-2z=-2 \\ x+2z=1 \end{array}$$

6.0.2 Spanning Sets

Definition 6.0.11 *The vector,* $\mathbf{u}$ *is a* ***linear combination*** *of the vectors,* $\mathbf{v}_1,\cdots,\mathbf{v}_m$ *if there exist scalars,* $c_1,\cdots,c_m$ *such that*

$$\begin{aligned} \mathbf{u} &= c_1\mathbf{v}_1+c_2\mathbf{v}_2+\cdots+c_m\mathbf{v}_m \\ &= \sum_{k=1}^{m} c_k\mathbf{v}_k. \end{aligned}$$

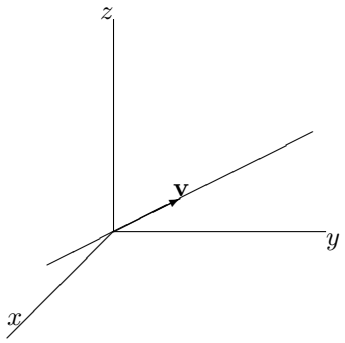
When $\mathbf{u}$ *is a linear combination of* $\{\mathbf{v}_1,\cdots,\mathbf{v}_m\}$, *we say* $\mathbf{u}$ *is in the* ***span*** *of* $\mathbf{v}_1,\cdots,\mathbf{v}_m$ *written*

$$\mathbf{u}\in\operatorname{span}(\mathbf{v}_1,\cdots,\mathbf{v}_m).$$

Equivalently, $\operatorname{span}(\mathbf{v}_1,\cdots,\mathbf{v}_m)$ *equals the set of all linear combinations of the vectors* $\mathbf{v}_1,\cdots,\mathbf{v}_m$. *If* $V=\operatorname{span}(\mathbf{v}_1,\cdots,\mathbf{v}_m)$, *then* $\{\mathbf{v}_1,\cdots,\mathbf{v}_m\}$ *is called a* ***spanning set*** *for* V.

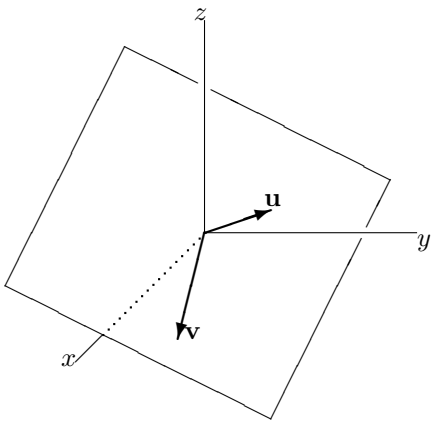
You can consider the geometric significance of the span of a few vectors in three or two dimensional space.

Example 6.0.12 *Consider the span of one vector in* $\mathbb{R}^3$.

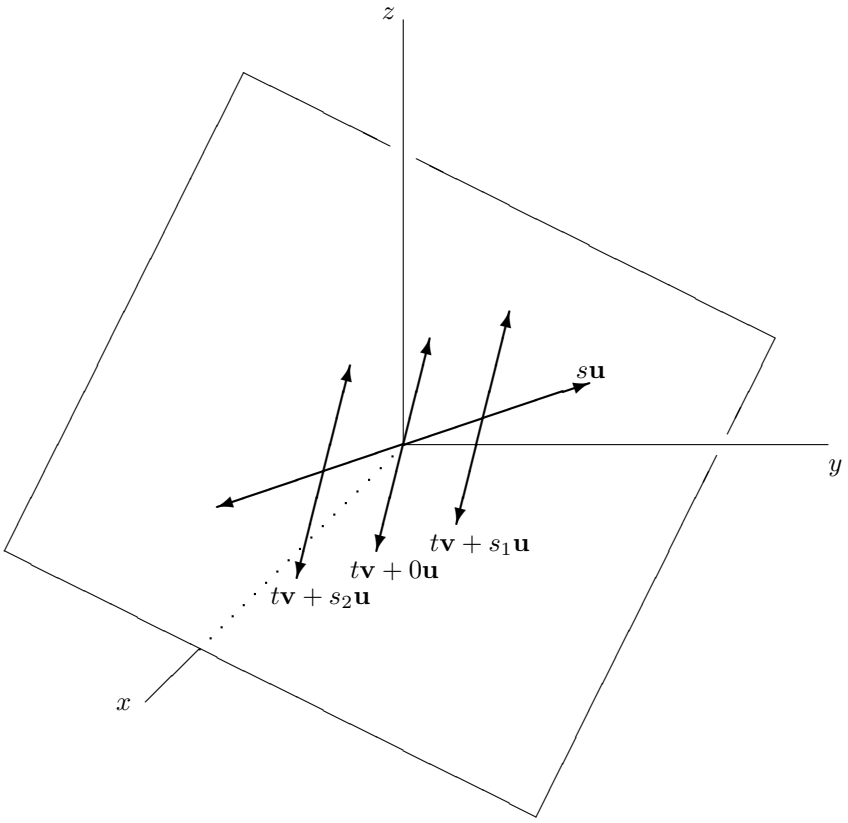


You see there is a vector, $\mathbf{v}$ and the span of this single vector, $\{t\mathbf{v} \text{ such that } t \in \mathbb{R}\}$ gives the indicated line which goes through the origin, $(0,0,0)$ having $\mathbf{v}$ as a direction vector.

Example 6.0.13 *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.

Geometric considerations such as these don't take you anywhere because as soon as you encounter more than three dimensions, you can't draw a meaningful picture. The notions of span and spanning set and linear combination are best understood according to the above definition and are algebraic in nature. Here is an example.

Example 6.0.14

$$3\begin{pmatrix}2\\3\\1\end{pmatrix}+5\begin{pmatrix}1\\-8\\2\end{pmatrix}+(-2)\begin{pmatrix}1\\3\\7\end{pmatrix}=\begin{pmatrix}9\\-37\\-1\end{pmatrix}$$

Thus $\begin{pmatrix}9\\-37\\-1\end{pmatrix}$ *is a linear combination of the vectors,* $\begin{pmatrix}2\\3\\1\end{pmatrix}$, $\begin{pmatrix}1\\-8\\2\end{pmatrix}$, *and* $\begin{pmatrix}1\\3\\7\end{pmatrix}$. *In this case the scalars are* $3, 5,$ *and* -2.

The following theorem is nothing but a restatement of the definition of what it means for a vector to be in the span of some other vectors.

Theorem 6.0.15 *Let A be an $m\times n$ matrix and let $\mathbf{b}$ be an $m\times 1$ vector. Then if the columns of A are $\mathbf{a}_1,\cdots,\mathbf{a}_n$, $\mathbf{b}\in\operatorname{span}(\mathbf{a}_1,\cdots,\mathbf{a}_n)$ if and only if the system of equations*

represented by the augmented matrix,

$$\left(\begin{array}{c|c} A & \mathbf{b} \end{array} \right) \tag{6.1}$$

is consistent.

Proof: Suppose first the system of equations just described is consistent. Let

$$\mathbf{a}_k = \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{mk} \end{pmatrix}, \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

To say the system is consistent is to say there exist $x_1, \cdots, x_n$ solving the following system of equations.

$$\begin{array}{c} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{array} \tag{6.2}$$

But from the way we add vectors, this can be written as

$$x_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \tag{6.3}$$

which says the same thing as $\mathbf{b} \in \text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_n)$.

Next suppose $\mathbf{b} \in \text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_n)$. This says there exist scalars, $x_1, \cdots, x_n$ such that 6.3 holds. This says the same thing as 6.2 and so the system of equations represented by 6.1 is consistent. This proves the theorem.

Example 6.0.16 *Show that a spanning set for* $\mathbb{R}^3$ *is* $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ *where*

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

This is really easy. If $\begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3$, you can write it as a linear combination of the above three vectors as follows.

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = x \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Of course it isn't always so easy.

Example 6.0.17 *Is*

$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

a spanning set for $\mathbb{R}^3$*?*

From Theorem 6.0.15 it is required to show the system of equations represented by the augmented matrix,

$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & a \\ 1 & 0 & 1 & b \\ 0 & 1 & 0 & c \end{array}\right)$$

has a solution for any choice of a, b, c. Take (-1) times the top row and add to the middle row.

$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & a \\ 0 & -1 & 1 & b-a \\ 0 & 1 & 0 & c \end{array}\right)$$

Now take the second row and add to the bottom.

$$\left(\begin{array}{ccc|c} 1 & 1 & 0 & a \\ 0 & -1 & 1 & b-a \\ 0 & 0 & 1 & c \end{array}\right)$$

You can see at this point that there will be a solution which you can obtain by back substitution. Therefore, the vectors are a spanning set for $\mathbb{R}^3$.

Example 6.0.18 *Is*

$$\left\{\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\right\}$$

a spanning set for $\mathbb{R}^3$?

By Theorem 6.0.15 you must consider the system of equations represented by

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & a \\ 1 & 0 & 1 & b \\ 0 & 1 & 1 & c \end{array}\right)$$

and see if there is a solution for any choice of a, b, c. Take (-1) times the top row and add to the second.

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & a \\ 0 & -1 & -1 & b-a \\ 0 & 1 & 1 & c \end{array}\right)$$

Now add the second row to the bottom.

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & a \\ 0 & -1 & -1 & b-a \\ 0 & 0 & 0 & c+b-a \end{array}\right) \tag{6.4}$$

It follows that to obtain a solution to this system you must have $c + b - a = 0$. Therefore, the vector, $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ fails to be in

$$\operatorname{span}\left(\left\{\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\right\}\right)$$

along with many others.

Example 6.0.19 *In the above example what is the span of the three given vectors?*

Let $a = b + c$ so 6.4 reduces to

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & a \\ 0 & -1 & -1 & b-a \\ 0 & 0 & 0 & 0 \end{array}\right)$$

and now you can add the middle row to the top row to obtain

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & b \\ 0 & -1 & -1 & b-a \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Multiply the second row by (-1) to get the result in row reduced echelon form

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & b \\ 0 & 1 & 1 & a-b \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Now there exists a solution to this. So what is the span of these vectors? It is

$$\left(\begin{array}{c} b+c \\ b \\ c \end{array}\right) : b, c \in \mathbb{R}.$$

That is, you can take either b or c to be anything you want and put it in the above formula for a vector and it will be in the span of the three vectors. Thus

$$\left(\begin{array}{c} 2 \\ 1 \\ 1 \end{array}\right), \left(\begin{array}{c} 0 \\ 1 \\ -1 \end{array}\right), \left(\begin{array}{c} 3 \\ 2 \\ 1 \end{array}\right)$$

are examples of vectors in the span of

$$\left\{\left(\begin{array}{c} 1 \\ 1 \\ 0 \end{array}\right), \left(\begin{array}{c} 1 \\ 0 \\ 1 \end{array}\right), \left(\begin{array}{c} 2 \\ 1 \\ 1 \end{array}\right)\right\}.$$

6.0.3 Linear Independence

When a vector is in the span of some other vectors you can say it is dependent on these other vectors and that all the vectors involved are a dependent set of vectors. The precise definition follows.

Definition 6.0.20 *A set of vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is **dependent** if there exist scalars, $c_1, c_2, \cdots, c_n$ not all zero such that*

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}.$$

*People often refer to this as: There exists a nontrivial linear combination of the vectors which equals zero. (It is nontrivial because some c_k is nonzero.) The set of vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is **independent** if it is not dependent. Thus there is no nontrivial linear combination which equals zero. Or equivalently, if*

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$$

then each $c_i = 0$. If you find scalars, $c_1, c_2, \cdots, c_n$ not all zero such that

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$$

this equation is called a ***dependence relation****.*

The following theorem is important.

Theorem 6.0.21 *A set of vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is dependent if and only if one of the vectors is a linear combination of the others.*

Proof: Suppose first that $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is dependent. Then there exist scalars, $c_1, \cdots, c_n$ not all zero such that

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$$

Let c_k be one of the scalars which is not zero. Then from the above equation,

$$c_k\mathbf{v}_k = -c_1\mathbf{v}_1 - c_2\mathbf{v}_2 \cdots - c_{k-1}\mathbf{v}_{k-1} - c_{k+1}\mathbf{v}_k \cdots - c_n\mathbf{v}_n$$

Now divide both sides by c_k to obtain

$$\begin{aligned}\mathbf{v}_k &= (-c_1/c_k)\,\mathbf{v}_1 + (-c_2/c_k)\,\mathbf{v}_2 + \cdots + (-c_{k-1}/c_k)\,\mathbf{v}_{k-1} \\ &\quad + (-c_{k+1}/c_k)\,\mathbf{v}_k + \cdots + (-c_n/c_k)\,\mathbf{v}_n.\end{aligned}$$

and this shows $\mathbf{v}_k$ is a linear combination of the other vectors.

Now suppose

$$\mathbf{v}_k = d_1\mathbf{v}_1 + \cdots + d_{k-1}\mathbf{v}_{k-1} + d_{k+1}\mathbf{v}_{k+1} + \cdots + d_n\mathbf{v}_n.$$

Then

$$\mathbf{0} = d_1\mathbf{v}_1 + \cdots + d_{k-1}\mathbf{v}_{k-1} + (-1)\,\mathbf{v}_k + d_{k+1}\mathbf{v}_{k+1} + \cdots + d_n\mathbf{v}_n$$

and so there is a nontrivial linear combination which equals zero. In fact $(-1) \neq 0$. This proves the theorem.

Observation 6.0.22 *Any set of vectors containing the zero vector is dependent. To see this, multiply the zero vector by 1 and all the other vectors by 0 and then add them together. You have a nontrivial linear combination equal to zero.*

How can you tell if $\{\mathbf{a}_1, \cdots, \mathbf{a}_n\}$ is independent or dependent? It must be one or the other. How can you determine which it is? Let

$$\mathbf{a}_k = \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{mk} \end{pmatrix}$$

Then a linear combination of the $\mathbf{a}_j$ where $\mathbf{a}_j$ is multiplied by the scalar, x_j is of the form

$$x_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix}$$

which is the same as

$$\begin{pmatrix} x_1a_{11} + x_2a_{12} + \cdots + x_na_{1n} \\ x_1a_{21} + x_2a_{22} + \cdots + x_na_{2n} \\ \vdots \\ x_1a_m + x_2a_{2m} + \cdots + x_na_{mn} \end{pmatrix}$$

Therefore,

$$x_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

if and only if

$$\begin{pmatrix} x_1a_{11} + x_2a_{12} + \cdots + x_na_{1n} = 0 \\ x_1a_{21} + x_2a_{22} + \cdots + x_na_{2n} = 0 \\ \vdots \\ x_1a_m + x_2a_{2m} + \cdots + x_na_{mn} = 0 \end{pmatrix}$$

Thus if A is an $m \times n$ matrix, the columns of A are dependent if and only if there exists a nonzero (nontrivial) solution to the system of equations represented by the augmented matrix,

$$\left(\begin{array}{c|c} A & \mathbf{0} \end{array} \right).$$

If $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ are dependent, it follows from Theorem 6.0.21 that one of the vectors is a linear combination of the others. Say $\mathbf{v}_k$ is a linear combination of the others. This means a suitable linear combination of the other vectors added to $\mathbf{v}_k$ yields $\mathbf{0}$.

This leads directly to the following theorem.

6.0.4 Recognizing Linear Dependence

Theorem 6.0.23 *Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ be vectors in $\mathbb{R}^n$. Make these vectors the rows of a matrix, A. Thus A is of the form*

$$\begin{pmatrix} - & \mathbf{v}_1 & - \\ & \vdots & \\ - & \mathbf{v}_m & - \end{pmatrix}.$$

Then the vectors are dependent if and only if $\operatorname{rank}(A) < m$.

Proof: If the vectors are dependent, then a linear combination gives the zero vector. Thus the row reduced echelon form has at least one row of zeros and so $\operatorname{rank}(A) < m$.

If $\operatorname{rank}(A) < m$, then the row reduced echelon form has at least one row of zeros. This row of zeros was obtained from doing row operations and so the rows are dependent. This proves the theorem.

Now recall Theorem 5.3.8 on Page 95 which is listed here for convenience.

Theorem 6.0.24 *Let A be an $m \times n$ matrix. Form the augmented matrix,*

$$\left(\begin{array}{c|c} A & \mathbf{0} \end{array} \right) \tag{6.5}$$

so that A is the coefficient matrix of a system of linear equations with n variables. Then the number of free variables $= n - \operatorname{rank}(A)$.

This theorem will be used to establish the following.

Theorem 6.0.25 *Any set of n vectors in $\mathbb{R}^m$ is linearly dependent if $n > m$.*

Proof: Let the set of n vectors be $\{\mathbf{a}_1, \cdots, \mathbf{a}_n\}$ and make them the column vectors of a matrix, A. Thus A is of the form

$$\begin{pmatrix} | & & | \\ \mathbf{a}_1 & \cdots & \mathbf{a}_n \\ | & & | \end{pmatrix}.$$

Consider the augmented matrix of Theorem 5.3.8 listed above. The number of free variables equals $n - \text{rank}(A)$ and $\text{rank}(A)$ is no more than m because there are only m rows in the matrix. Therefore, the number of free variables in the system of equations represented by 6.5 equals $n - \text{rank}(A) > n - m > 0$. Since there exist free variables, there exist non zero solutions to the system represented by 6.5 which implies a nontrivial linear combination of the vectors equals zero. Thus the set of vectors is dependent. This proves the theorem.

6.0.5 Discovering Dependence Relations

Suppose you have some vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and you wonder whether they are independent or dependent. If dependent, can you find a dependence relation? How do you go about answering this question? Recall the definition. You are looking for scalars, $x_1, \cdots, x_n$ such that

$$x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n = \mathbf{0}$$

and you are trying to find whether there are any nonzero solutions to this vector equation. If the vectors, $\mathbf{v}_i$ are in $\mathbb{F}^m$, then the above is really just a homogeneous system of equations because there is an equation for each component. Thus you form the augmented matrix,

$$\begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n & | & \mathbf{0} \end{pmatrix}$$

and row reduce to find the solution. If the only solution is $x_1 = x_2 = \cdots = x_n = 0$, then the vectors are linearly independent. If there exists something nonzero in the system of equations, then you have produced a dependence relation.

Example 6.0.26 *Here are some vectors in $\mathbb{R}^3$.* $\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 5 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$. *I know by Theorem 6.0.25 that these vectors are dependent. However, I would like to find a dependence relation.*

To do this, I let them be the columns of an augmented matrix as shown.

$$\begin{pmatrix} 1 & 0 & 2 & 1 & | & 0 \\ 2 & 1 & 5 & 1 & | & 0 \\ 0 & 1 & 1 & 1 & | & 0 \end{pmatrix}$$

Then I row reduce this in order to find the solutions. The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 2 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Therefore, The solutions are $x_4 = 0, x_2 = -x_3$, and $x_1 = -2x_3$. Thus, letting $x_3 = t$, all the solutions are

$$(-2t, -t, t, 0) : t \in \mathbb{R}$$

and so if you let $t = 1$, you find the dependence relation,

$$-2\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} + (-1)\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + 1\begin{pmatrix} 2 \\ 5 \\ 1 \end{pmatrix} + 0\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

There is nothing new here at all! It is just another way of asking for the solution to a homogeneous system of linear equations.[1]

[1] This is the style in linear algebra books these days, ask the same question over and over again in disguised form to give the illusion of learning something new. However, there is much more to linear algebra than row reduction of augmented matrices.

Matrices

7.1 Matrix Operations And Algebra 20,21 Sept.

Quiz

1. Here are three points: $(1,1,1), (2,0,1), (0,1,0)$. Find an equation of a plane which contains all three points.

2. Find the equation of a plane which is parallel to the plane whose equations is $x+2y+z=7$ which contains the point $(1,2,1)$.

3. Here are three vectors: $(1,2,1), (2,1,0), (-2,0,1)$. Find the volume of the parallelepiped determined by these three vectors.

4. Here are three vectors. $\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}$. Determine whether the vectors are dependent. If they are dependent, find a dependence relation.

5. Here is a system of equations.

$$\begin{aligned} 3x+4y+z &= 4 \\ x+2y+z &= 2 \\ y+z &= 1 \end{aligned}$$

   Find the complete solution.

7.1.1 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}) refersto 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 are identical. Thus

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

because they are different sizes. As noted above, you write (c_{ij}) for the matrix C whose ij^{th} entry is c_{ij}. In doing arithmetic with matrices you must define what happens in terms of the c_{ij} sometimes called the **entries** of the matrix or the **components** of the matrix.

The above discussion stated for general matrices is given in the following definition.

Definition 7.1.1 *(Scalar Multiplication) If $A = (a_{ij})$ and k is a scalar, then $kA = (ka_{ij})$.*

Example 7.1.2 $7\begin{pmatrix} 2 & 0 \\ 1 & -4 \end{pmatrix} = \begin{pmatrix} 14 & 0 \\ 7 & -28 \end{pmatrix}$.

Definition 7.1.3 *(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 7.1.4

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 2 & 3 \\ -6 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 4 & 6 \\ -5 & 2 & 5 \end{pmatrix}$$

To save on notation, we will often use A_{ij} to refer to the ij^{th} entry of the matrix, A.

Definition 7.1.5 *(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 7.1.6 *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 7.1.7 *(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.

$$A + B = B + A, \tag{7.1}$$

- Associative Law for Addition.

$$(A + B) + C = A + (B + C), \tag{7.2}$$

- Existence of an Additive Identity

$$A + 0 = A, \tag{7.3}$$

- Existence of an Additive Inverse

$$A + (-A) = 0, \tag{7.4}$$

Also for α, β scalars, the following additional properties hold.

- Distributive law over Matrix Addition.

$$\alpha (A + B) = \alpha A + \alpha B, \tag{7.5}$$

- Distributive law over Scalar Addition

$$(\alpha + \beta) A = \alpha A + \beta A, \tag{7.6}$$

- Associative law for Scalar Multiplication

$$\alpha (\beta A) = \alpha\beta (A), \tag{7.7}$$

- Rule for Multiplication by 1.

$$1A = A. \tag{7.8}$$

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.

7.1.2 Multiplication Of Matrices

Definition 7.1.8 *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,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{pmatrix}.$$

Another way to think of this is

$$x_1 \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} + x_3 \begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix}$$

Thus you take x_1 times the first column, add to x_2 times the second column, and finally x_3 times the third column. In general, here is the definition of how to multiply an $(m \times n)$ matrix times a $(n \times 1)$ matrix.

Definition 7.1.9 *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

$$A\mathbf{v} = \begin{pmatrix} \sum_{j=1}^{n} A_{1j}v_j \\ \vdots \\ \sum_{j=1}^{n} A_{mj}v_j \end{pmatrix}. \tag{7.9}$$

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 7.9 and the observation that the j^{th} column of A is

$$\begin{pmatrix} A_{1j} \\ A_{2j} \\ \vdots \\ A_{mj} \end{pmatrix}$$

so 7.9 reduces to

$$v_1 \begin{pmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{m1} \end{pmatrix} + v_2 \begin{pmatrix} A_{12} \\ A_{22} \\ \vdots \\ A_{m2} \end{pmatrix} + \cdots + v_n \begin{pmatrix} A_{1n} \\ A_{2n} \\ \vdots \\ A_{mn} \end{pmatrix}$$

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 7.1.10 *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.

$$(m \times \overbrace{\quad n)\,(n}^{\text{these must match!}} \times p \quad) = m \times p$$

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 7.1.11 *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:

$$AB \equiv (A\mathbf{b}_1, \cdots, A\mathbf{b}_p) \tag{7.10}$$

where $A\mathbf{b}_k$ is an $m \times 1$ matrix or column vector which gives the k^{th} column of AB.

Example 7.1.12 *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 7.1.13 *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!

7.1.3 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

$$\sum_{k=1}^{n} A_{ik} B_{kj}. \tag{7.11}$$

In terms of pictures of the matrix, you are doing

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{np} \end{pmatrix}$$

Then as explained above, the j^{th} column is of the form

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} B_{1j} \\ B_{2j} \\ \vdots \\ B_{nj} \end{pmatrix}$$

which is a $m \times 1$ matrix or column vector which equals

$$\begin{pmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{m1} \end{pmatrix} B_{1j} + \begin{pmatrix} A_{12} \\ A_{22} \\ \vdots \\ A_{m2} \end{pmatrix} B_{2j} + \cdots + \begin{pmatrix} A_{1n} \\ A_{2n} \\ \vdots \\ A_{mn} \end{pmatrix} B_{nj}.$$

The second entry of this $m \times 1$ matrix is

$$A_{21}B_{1j} + A_{22}B_{2j} + \cdots + A_{2n}B_{nj} = \sum_{k=1}^{m} A_{2k}B_{kj}.$$

Similarly, the i^{th} entry of this $m \times 1$ matrix is

$$A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{in}B_{nj} = \sum_{k=1}^{m} A_{ik}B_{kj}.$$

This shows the following definition for matrix multiplication in terms of the ij^{th} entries of the product coincides with Definition 7.1.11.

Definition 7.1.14 *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*

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}. \tag{7.12}$$

Example 7.1.15 *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 7.1.16 *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 7.1.17 *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 7.1.18 *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

$$\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 4 & 2 & 0 \\ 1 & 2 & 1 & 0 \end{pmatrix}$$

7.1.4 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 7.1.19 *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 7.1.20 *If all multiplications and additions make sense, the following hold for matrices, A, B, C and a, b scalars.*

$$A(aB + bC) = a(AB) + b(AC) \tag{7.13}$$

$$(B + C)A = BA + CA \tag{7.14}$$

$$A(BC) = (AB)C \tag{7.15}$$

Proof: Using Definition 7.1.14,

$$\begin{aligned}
\left(A\left(aB+bC\right)\right)_{ij} &= \sum_k A_{ik}\left(aB+bC\right)_{kj} \\
&= \sum_k A_{ik}\left(aB_{kj}+bC_{kj}\right) \\
&= a\sum_k A_{ik}B_{kj} + b\sum_k A_{ik}C_{kj} \\
&= a\left(AB\right)_{ij} + b\left(AC\right)_{ij} \\
&= \left(a\left(AB\right)+b\left(AC\right)\right)_{ij}.
\end{aligned}$$

Thus $A\left(B+C\right) = AB+AC$ as claimed. Formula 7.14 is entirely similar.

Formula 7.15 is the associative law of multiplication. Using Definition 7.1.14,

$$\begin{aligned}
\left(A\left(BC\right)\right)_{ij} &= \sum_k A_{ik}\left(BC\right)_{kj} \\
&= \sum_k A_{ik}\sum_l B_{kl}C_{lj} \\
&= \sum_l \left(AB\right)_{il}C_{lj} \\
&= \left(\left(AB\right)C\right)_{ij}.
\end{aligned}$$

This proves 7.15.

7.1.5 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 7.1.21 *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 7.1.22

$$\begin{pmatrix} 1 & 2 & -6 \\ 3 & 5 & 4 \end{pmatrix}^T = \begin{pmatrix} 1 & 3 \\ 2 & 5 \\ -6 & 4 \end{pmatrix}.$$

The transpose of a matrix has the following important properties.

Lemma 7.1.23 *Let A be an $m\times n$ matrix and let B be a $n\times p$ matrix. Then*

$$\left(AB\right)^T = B^TA^T \tag{7.16}$$

and if α and β are scalars,

$$(\alpha A + \beta B)^T = \alpha A^T + \beta B^T \tag{7.17}$$

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 7.17 is left as an exercise and this proves the lemma.

Definition 7.1.24 *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 7.1.25 *Let*

$$A = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 5 & -3 \\ 3 & -3 & 7 \end{pmatrix}.$$

Then A is symmetric.

Example 7.1.26 *Let*

$$A = \begin{pmatrix} 0 & 1 & 3 \\ -1 & 0 & 2 \\ -3 & -2 & 0 \end{pmatrix}$$

Then A is skew symmetric.

7.1.6 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 7.1.27 *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 7.1.28 *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 7.1.29 *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 7.1.30 *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 7.1.31 *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.

7.2 Finding The Inverse Of A Matrix, Gauss Jordan Method 21,22 Sept.

Quiz

1. Multiply the matrices if possible.

$$\begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 2 \end{pmatrix}$$

2. Multiply the matrices if possible.

$$\begin{pmatrix} 1 & 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

3. Multiply the matrices if possible.

$$\begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 & 0 \end{pmatrix}$$

4. True or False. In each case the capital letters are matrices of an appropriate size and the lower case letters represent numbers.

   (a) $A^2 - B^2 = (A - B)(A + B)$

   (b) $(AB)^T = A^T B^T$

   (c) $(aA + bB)C = aAC + bCB$

   (d) If $AB = 0$, then either $A = 0$ or $B = 0$.

   (e) $A/A = 1$

   (f) $(AB)C = A(BC)$

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

$$\left(\begin{array}{cc|c} 1 & 1 & 1 \\ 1 & 2 & 0 \end{array}\right) \tag{7.18}$$

for the first system and

$$\left(\begin{array}{cc|c} 1 & 1 & 0 \\ 1 & 2 & 1 \end{array}\right) \tag{7.19}$$

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, 7.19 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

$$\left(\begin{array}{cc} 2 & -1 \\ -1 & 1 \end{array}\right).$$

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

$$\left(\begin{array}{cc|cc} 1 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{array}\right)$$

and row reduced till you obtained

$$\left(\begin{array}{cc|cc} 1 & 0 & 2 & -1 \\ 0 & 1 & -1 & 1 \end{array}\right)$$

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 7.2.1 *Suppose A is an $n \times n$ matrix. To find A^{-1} if it exists, form the augmented $n \times 2n$ matrix,*

$$(A|I)$$

and then, if possible do row operations until you obtain an $n \times 2n$ matrix of the form

$$(I|B)\,. \tag{7.20}$$

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 7.2.2 *Let* $A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 3 & 1 & -1 \end{pmatrix}$. *Find* A^{-1} *if it exists.*

Set up the augmented matrix, $(A|I)$

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 & 1 & 0 \\ 3 & 1 & -1 & 0 & 0 & 1 \end{array}\right)$$

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

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -2 & 0 & -1 & 1 & 0 \\ 0 & -5 & -7 & -3 & 0 & 1 \end{array}\right).$$

Then take 5 times the second row and add to -2 times the last row.

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{array}\right)$$

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

$$\begin{pmatrix} -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{pmatrix}$$

Example 7.2.3 *Let* $A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 4 \end{pmatrix}$. *Find* A^{-1} *if it exists.*

Write the augmented matrix, $(A|I)$

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 2 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 & 1 & 0 \\ 2 & 2 & 4 & 0 & 0 & 1 \end{array}\right)$$

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 7.2.4 *Let* $A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}$. *Find* A^{-1} *if it exists.*

Form the augmented matrix,

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 & 1 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 \end{array}\right).$$

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.

Example 7.2.5 *In this example, it is shown how to use the inverse of a matrix to find the solution to a system of equations. Consider the following system of equations. Use the inverse of a suitable matrix to give the solutions to this system.*

$$\begin{pmatrix} x+z=1 \\ x-y+z=3 \\ x+y-z=2 \end{pmatrix}.$$

The system of equations can be written in terms of matrices as

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}. \tag{7.21}$$

More simply, this is of the form $A\mathbf{x} = \mathbf{b}$. Suppose you find the inverse of the matrix, A^{-1}. Then you could multiply both sides of this equation by A^{-1} to obtain

$$\mathbf{x} = \left(A^{-1}A\right)\mathbf{x} = A^{-1}\left(A\mathbf{x}\right) = A^{-1}\mathbf{b}.$$

This gives the solution as $\mathbf{x} = A^{-1}\mathbf{b}$. Note that once you have found the inverse, you can easily get the solution for different right hand sides without any effort. It is always just $A^{-1}\mathbf{b}$. In the given example, the inverse of the matrix is

$$\begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & -1 & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix}$$

This was shown in Example 7.2.4. Therefore, from what was just explained the solution to the given system is

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & -1 & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} = \begin{pmatrix} \frac{5}{2} \\ -2 \\ -\frac{3}{2} \end{pmatrix}.$$

What if the right side of 7.21 had been

$$\begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix}?$$

What would be the solution to

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix}?$$

By the above discussion, it is just

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & -1 & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ -2 \end{pmatrix}.$$

This illustrates why once you have found the inverse of a given matrix, you can use it to solve many different systems easily.

Here is a formula for the inverse of a 2×2 matrix.

Theorem 7.2.6 *Let* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *where* $ad - bc \neq 0$. *Then*

$$A^{-1} = \frac{1}{ad-cb}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Proof: Just multiply and verify it works.

$$\frac{1}{ad-cb}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Therefore, this is indeed the inverse.

The expression, $ad-cb$ is the determinant of the given matrix. Recall, this was discussed in connection with the cross product. This will be discussed in more generality later.

7.3 Elementary Matrices 22 Sept.

Quiz

1. Here is a matrix.
$$\begin{pmatrix} 1 & -1 & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \\ -1 & \frac{3}{2} & \frac{1}{2} \end{pmatrix}$$
Find its inverse.

2. The inverse of $\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & -1 & 0 \end{pmatrix}$ is $\begin{pmatrix} 1 & -1 & 0 \\ 1 & -1 & -1 \\ -1 & 3 & 1 \end{pmatrix}$. Use this fact to write the solution to the system
$$\begin{pmatrix} x + \frac{1}{2}y + \frac{1}{2}z = a \\ \frac{1}{2}y + \frac{1}{2}z = b \\ x - y = c \end{pmatrix}$$
in terms of a, b, c.

The elementary matrices result from doing a row operation to the identity matrix.

Definition 7.3.1 *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.*

The elementary matrices are given in the following definition.

Definition 7.3.2 *The elementary matrices consist of those matrices which result by applying a row operation to an identity matrix. Those which involve switching rows of the identity are called permutation matrices*[1].

[1] More generally, a permutation matrix is a matrix which comes by permuting the rows of the identity matrix, not just switching two rows.

As an example of why these elementary matrices are interesting, consider the following.

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b & c & d \\ x & y & z & w \\ f & g & h & i \end{pmatrix} = \begin{pmatrix} x & y & z & w \\ a & b & c & d \\ f & g & h & i \end{pmatrix}$$

A 3×4 matrix was multiplied on the left by an elementary matrix which was obtained from row operation 1 applied to the identity matrix. This resulted in applying the operation 1 to the given matrix. This is what happens in general.

Now consider what these elementary matrices look like. First consider the one which involves switching row i and row j where $i < j$. This matrix is of the form

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & & & & & & & & & \vdots \\ \vdots & & 1 & & & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 & 1 & \cdots & \cdots & 0 \\ \vdots & & & \vdots & 1 & 0 & \cdots & 0 & & & \vdots \\ \vdots & & & \vdots & & \ddots & & \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 & \cdots & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & \cdots & 0 & \cdots & \cdots & 0 \\ \vdots & & & & & & & & 1 & & \vdots \\ \vdots & & & & & & & & & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

The two exceptional rows are shown. The i^{th} row was the j^{th} and the j^{th} row was the i^{th} in the identity matrix. Now consider what this does to a column vector.

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & & & & & & & & & \vdots \\ \vdots & & 1 & & & & & & & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 & 1 & \cdots & \cdots & 0 \\ \vdots & & & \vdots & 1 & 0 & \cdots & 0 & & & \vdots \\ \vdots & & & \vdots & & \ddots & & \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 & \cdots & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & \cdots & 0 & \cdots & \cdots & 0 \\ \vdots & & & & & & & & 1 & & \vdots \\ \vdots & & & & & & & & & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ \vdots \\ v_i \\ \vdots \\ \vdots \\ \vdots \\ v_j \\ \vdots \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ \vdots \\ v_j \\ \vdots \\ \vdots \\ \vdots \\ v_i \\ \vdots \\ \vdots \\ v_n \end{pmatrix}$$

Now denote by P^{ij} the elementary matrix which comes from the identity from switching rows i and j. From what was just explained consider multiplication on the left by this

elementary matrix.

$$P^{ij}\begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{j1} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}$$

From the way you multiply matrices this is a matrix which has the indicated collumns.

$$\begin{pmatrix} P^{ij}\begin{pmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{j1} \\ \vdots \\ a_{n1} \end{pmatrix}, P^{ij}\begin{pmatrix} a_{12} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots, P^{ij}\begin{pmatrix} a_{1p} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{np} \end{pmatrix} \end{pmatrix}$$

$$= \left(\begin{pmatrix} a_{11} \\ \vdots \\ a_{j1} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{n1} \end{pmatrix}, \begin{pmatrix} a_{12} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots, \begin{pmatrix} a_{1p} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{np} \end{pmatrix} \right)$$

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{j1} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}$$

This has established the following lemma.

Lemma 7.3.3 *Let P^{ij} denote the elementary matrix which involves switching the i^{th} and the j^{th} rows. Then*

$$P^{ij}A = B$$

where B is obtained from A by switching the i^{th} and the j^{th} rows.

Next consider the row operation which involves multiplying the i^{th} row by a nonzero constant, c. The elementary matrix which results from applying this operation to the i^{th}

row of the identity matrix is of the form

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & & & & & \vdots \\ \vdots & & 1 & & & & \vdots \\ \vdots & & & c & & & \vdots \\ \vdots & & & & 1 & & \vdots \\ \vdots & & & & & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

Now consider what this does to a column vector.

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & & & & & \vdots \\ \vdots & & 1 & & & & \vdots \\ \vdots & & & c & & & \vdots \\ \vdots & & & & 1 & & \vdots \\ \vdots & & & & & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_{i-1} \\ v_i \\ v_{i+1} \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_{i-1} \\ cv_i \\ v_{i+1} \\ \vdots \\ v_n \end{pmatrix}$$

Denote by $E(c,i)$ this elementary matrix which multiplies the i^{th} row of the identity by the nonzero constant, c. Then from what was just discussed and the way matrices are multiplied,

$$E(c,i) \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{j2} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}$$

equals a matrix having the columns indicated below.

$$= \left(E(c,i) \begin{pmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{j1} \\ \vdots \\ a_{n1} \end{pmatrix}, E(c,i) \begin{pmatrix} a_{12} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots, E(c,i) \begin{pmatrix} a_{1p} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{np} \end{pmatrix} \right)$$

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ ca_{i1} & ca_{i2} & \cdots & \cdots & \cdots & \cdots & ca_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{j2} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}$$

This proves the following lemma.

Lemma 7.3.4 *Let $E(c,i)$ denote the elementary matrix corresponding to the row operation in which the i^{th} row is multiplied by the nonzero constant, c. Thus $E(c,i)$ involves multiplying the i^{th} row of the identity matrix by c. Then*

$$E(c,i)\,A = B$$

where B is obtained from A by multiplying the i^{th} row of A by c.

Finally consider the third of these row operations. Denote by $E(c \times i + j)$ the elementary matrix which replaces the j^{th} row with itself added to c times the i^{th} row added to it. In case $i < j$ this will be of the form

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\ 0 & \ddots & & & & & \vdots \\ \vdots & & 1 & & & & \vdots \\ \vdots & & \vdots & \ddots & & & \vdots \\ \vdots & & c & \cdots & 1 & & \vdots \\ \vdots & & & & & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

Now consider what this does to a column vector.

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\ 0 & \ddots & & & & & \vdots \\ \vdots & & 1 & & & & \vdots \\ \vdots & & \vdots & \ddots & & & \vdots \\ \vdots & & c & \cdots & 1 & & \vdots \\ \vdots & & & & & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_j \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ cv_i + v_j \\ \vdots \\ v_n \end{pmatrix}$$

Now from this and the way matrices are multiplied,

$$E(c \times i + j) \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{j2} & a_{j2} & \cdots & \cdots & \cdots & \cdots & a_{jp} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}$$

equals a matrix of the following form having the indicated columns.

$$\left(E(c \times i + j) \begin{pmatrix} a_{11} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{n1} \end{pmatrix}, E(c \times i + j) \begin{pmatrix} a_{12} \\ \vdots \\ a_{i2} \\ \vdots \\ a_{j2} \\ \vdots \\ a_{n2} \end{pmatrix}, \cdots E(c \times i + j) \begin{pmatrix} a_{1p} \\ \vdots \\ a_{ip} \\ \vdots \\ a_{jp} \\ \vdots \\ a_{np} \end{pmatrix} \right)$$

$$= \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & \cdots & a_{1p} \\ \vdots & \vdots & & & & & \vdots \\ a_{i1} & a_{i2} & \cdots & \cdots & \cdots & \cdots & a_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{j2} + ca_{i1} & a_{j2} + ca_{i2} & \cdots & \cdots & \cdots & \cdots & a_{jp} + ca_{ip} \\ \vdots & \vdots & & & & & \vdots \\ a_{n1} & a_{n2} & \cdots & \cdots & \cdots & \cdots & a_{np} \end{pmatrix}$$

The case where $i > j$ is handled similarly. This proves the following lemma.

Lemma 7.3.5 *Let $E(c \times i + j)$ denote the elementary matrix obtained from I by replacing the j^{th} row with c times the i^{th} row added to it. Then*

$$E(c \times i + j)\, A = B$$

where B is obtained from A by replacing the j^{th} row of A with itself added to c times the i^{th} row of A.

The next theorem is the main result.

Theorem 7.3.6 *To perform any of the three row operations on a matrix, A it suffices to do the row operation on the identity matrix obtaining an elementary matrix, E and then take the product, EA. Furthermore, each elementary matrix is invertible and its inverse is an elementary matrix.*

Proof: The first part of this theorem has been proved in Lemmas 7.3.3 - 7.3.5. It only remains to verify the claim about the inverses. Consider first the elementary matrices corresponding to row operation of type three.

$$E(-c\times i+j)\,E(c\times i+j)=I$$

This follows because the first matrix takes c times row i in the identity and adds it to row j. When multiplied on the left by $E(-c\times i+j)$ it follows from the first part of this theorem that you take the i^{th} row of $E(c\times i+j)$ which coincides with the i^{th} row of I since that row was not changed, multiply it by $-c$ and add to the j^{th} row of $E(c\times i+j)$ which was the j^{th} row of I added to c times the i^{th} row of I. Thus $E(-c\times i+j)$ multiplied on the left, undoes the row operation which resulted in $E(c\times i+j)$. The same argument applied to the product

$$E(c\times i+j)\,E(-c\times i+j)$$

replacing c with $-c$ in the argument yields that this product is also equal to I. Therefore, $E(c\times i+j)^{-1}=E(-c\times i+j)$.

Similar reasoning shows that for $E(c,i)$ the elementary matrix which comes from multiplying the i^{th} row by the nonzero constant, c,

$$E(c,i)^{-1}=E\left(c^{-1},i\right).$$

Finally, consider P^{ij} which involves switching the i^{th} and the j^{th} rows.

$$P^{ij}P^{ij}=I$$

because by the first part of this theorem, multiplying on the left by P^{ij} switches the i^{th} and j^{th} rows of P^{ij} which was obtained from switching the i^{th} and j^{th} rows of the identity. First you switch them to get P^{ij} and then you multiply on the left by P^{ij} which switches these rows again and restores the identity matrix. Thus $\left(P^{ij}\right)^{-1}=P^{ij}$.

Corollary 7.3.7 *Let A be an invertible $n\times n$ matrix. Then A equals a finite product of elementary matrices.*

Proof: Since A^{-1} is given to exist, it follows A must have rank n and so the row reduced echelon form of A is I. Therefore, by Theorem 7.3.6 there is a sequence of elementary matrices, $E_1,\cdots,E_p$ which accomplish successive row operations such that

$$(E_pE_{p-1}\cdots E_1)\,A=I.$$

But now multiply on the left on both sides by E_p^{-1} then by E_{p-1}^{-1} and then by E_{p-2}^{-1} etc. until you get

$$A=E_1^{-1}E_2^{-1}\cdots E_{p-1}^{-1}E_p^{-1}$$

and by Theorem 7.3.6 each of these in this product is an elementary matrix.

7.4 Block Multiplication Of Matrices

Consider the following problem

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix}$$

You know how to do this. You get

$$\begin{pmatrix} AE+BG & AF+BH \\ CE+DG & CF+DH \end{pmatrix}.$$

Now what if instead of numbers, the entries, A, B, C, D, E, F, G are matrices of a size such that the multiplications and additions needed in the above formula all make sense. Would the formula be true in this case? I will show below that it is true.

Suppose A is a matrix of the form

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{r1} & \cdots & A_{rm} \end{pmatrix} \tag{7.22}$$

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,** also a **partitioned 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 7.4.1 *Let A be an $m \times n$ block matrix as in 7.22 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 block matrix of the form

$$\begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix} \tag{7.23}$$

and A is a block matrix of the form

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{p1} & \cdots & A_{pm} \end{pmatrix} \tag{7.24}$$

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, B_{is} and A_{sj} 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 7.4.2 *Let B be a block matrix as in 7.23 and let A be a block matrix as in 7.24 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 block matrix such that the ij^{th} block is of the form*

$$\sum_s B_{is}A_{sj}. \tag{7.25}$$

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 7.4.1 applied twice,

$$\left(\begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix}\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{p1} & \cdots & A_{pm} \end{pmatrix}\right)\begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{pmatrix}$$

$$= \begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix}\begin{pmatrix} \sum_j A_{1j}\mathbf{x}_j \\ \vdots \\ \sum_j A_{rj}\mathbf{x}_j \end{pmatrix}$$

$$= \begin{pmatrix} \sum_s \sum_j B_{1s}A_{sj}\mathbf{x}_j \\ \vdots \\ \sum_s \sum_j B_{rs}A_{sj}\mathbf{x}_j \end{pmatrix} = \begin{pmatrix} \sum_j \left(\sum_s B_{1s}A_{sj}\right)\mathbf{x}_j \\ \vdots \\ \sum_j \left(\sum_s B_{rs}A_{sj}\right)\mathbf{x}_j \end{pmatrix}$$

$$= \begin{pmatrix} \sum_s B_{1s}A_{s1} & \cdots & \sum_s B_{1s}A_{sm} \\ \vdots & \ddots & \vdots \\ \sum_s B_{rs}A_{s1} & \cdots & \sum_s B_{rs}A_{sm} \end{pmatrix}\begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{pmatrix}$$

By Lemma 7.4.1, this shows that $(BA)\,\mathbf{x}$ equals the block matrix whose ij^{th} entry is given by 7.25 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.

7.4.1 Exercises With Answers

1. Here are some matrices:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 7 \\ 1 & 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 3 & -1 & 2 \\ -3 & 2 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{pmatrix}, D = \begin{pmatrix} -1 & 2 \\ 2 & -3 \end{pmatrix}, E = \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

Find if possible $-3A, 3B - A, AC, CB, EA, DC^T$ If it is not possible explain why.

$$-3A = -3\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 7 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -3 & -6 & -9 \\ -6 & -9 & -21 \\ -3 & 0 & -3 \end{pmatrix}$$

$3B - A$ is nonsense because the matrices B and A are not of the same size.

$$AC = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 7 \\ 1 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 10 & 7 \\ 18 & 14 \\ 2 & 3 \end{pmatrix}$$

There is no problem here because you are doing $(3 \times 3)\,(3 \times 2)$.

$$CB = \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 3 & -1 & 2 \\ -3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} -3 & 3 & 4 \\ 6 & -1 & 7 \\ 0 & 1 & 3 \end{pmatrix}$$

There is no problem here because you are doing $(3\times 2)(2\times 3)$ and the inside numbers match. EA is nonsense because it is of the form $(2\times 1)(3\times 3)$ so since the inside numbers do not match the matrices are not conformable.

$$DC^T = \begin{pmatrix} -1 & 2 \\ 2 & -3 \end{pmatrix}\begin{pmatrix} 1 & 3 & 1 \\ 2 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & -1 & 1 \\ -4 & 3 & -1 \end{pmatrix}.$$

2. Let $A = \begin{pmatrix} 0 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 1 & 2 \\ 1 & k \end{pmatrix}$. Is it possible to choose k such that $AB = BA$? If so, what should k equal?

   We just multiply and see if it can happen.

$$AB = \begin{pmatrix} 0 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 1 & k \end{pmatrix} = \begin{pmatrix} 2 & 2k \\ 7 & 6+4k \end{pmatrix}.$$

   On the other hand,

$$BA = \begin{pmatrix} 1 & 2 \\ 1 & k \end{pmatrix}\begin{pmatrix} 0 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 6 & 10 \\ 3k & 2+4k \end{pmatrix}.$$

   If these were equal you would need to have $6 = 2$ which is not the case. Therefore, there is no way to choose k such that these two matrices will commute.

3. Let $\mathbf{x} = (-1, 0, 3)$ and $\mathbf{y} = (3, 1, 2)$. Find $\mathbf{x}^T\mathbf{y}$.

$$\mathbf{x}^T\mathbf{y} = \begin{pmatrix} -1 \\ 0 \\ 3 \end{pmatrix}\begin{pmatrix} 3 & 1 & 2 \end{pmatrix} = \begin{pmatrix} -3 & -1 & -2 \\ 0 & 0 & 0 \\ 9 & 3 & 6 \end{pmatrix}.$$

4. Write $\begin{pmatrix} 4x_1 - x_2 + 2x_3 \\ 2x_3 + 7x_1 \\ 2x_3 \\ 3x_3 + 3x_2 + x_1 \end{pmatrix}$ in the form $A\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$ where A is an appropriate matrix.

$$\begin{pmatrix} 4 & -1 & 2 & 0 \\ 7 & 0 & 2 & 0 \\ 0 & 0 & 2 & 0 \\ 1 & 3 & 3 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}.$$

5. Let

$$A = \begin{pmatrix} 1 & 2 & 5 \\ 2 & 1 & 4 \\ 1 & 0 & 2 \end{pmatrix}.$$

   Find A^{-1} if possible. If A^{-1} does not exist, determine why.

$$\begin{pmatrix} 1 & 2 & 5 \\ 2 & 1 & 4 \\ 1 & 0 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} -\frac{2}{3} & \frac{4}{3} & -1 \\ 0 & 1 & -2 \\ \frac{1}{3} & -\frac{2}{3} & 1 \end{pmatrix}.$$

6. Let

$$A = \begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 5 & 2 & 0 \\ 2 & 1 & -3 & 2 \\ 1 & 2 & 1 & 2 \end{pmatrix}$$

Find A^{-1} if possible. If A^{-1} does not exist, determine why.

$$\begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 5 & 2 & 0 \\ 2 & 1 & -3 & 2 \\ 1 & 2 & 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} -3 & \frac{1}{6} & \frac{5}{6} & \frac{13}{6} \\ 1 & \frac{1}{6} & -\frac{1}{6} & -\frac{5}{6} \\ -1 & 0 & 0 & 1 \\ 1 & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \end{pmatrix}.$$

7. 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 $B = A^{-1}$.

   From $AB = I$, multiply both sides by A^{-1}. Thus $A^{-1}(AB) = A^{-1}$. Then from the associative property of matrix multiplication, $A^{-1} = A^{-1}(AB) = \left(A^{-1}A\right)B = IB = B$.

8. Suppose A, B are two matrices. Show $(AB)^{-1} = B^{-1}A^{-1}$.

   All you have to do is multiply it. If it acts like the inverse, it is the inverse.

   $$\left(B^{-1}A^{-1}\right)(AB) = B^{-1}\left(A^{-1}A\right)B = B^{-1}IB = B^{-1}B = I.$$

   Therefore, $B^{-1}A^{-1} = (AB)^{-1}$.

9. Show $\left(A^{-1}\right)^T = \left(A^T\right)^{-1}$.

   $A^T\left(A^{-1}\right)^T = \left(A^{-1}A\right)^T = I^T = I$ and so $\left(A^{-1}\right)^T = \left(A^T\right)^{-1}$.

10. Here are elementary matrices. Find their inverses.

    (a) $$\begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
    $$\begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -a & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

    (b) $$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
    $$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

    (c) $$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
    $$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

11. When you have an $n \times n$ matrix, $A, A^n = \overbrace{A \times A \times \cdots \times A}^{n \text{ times}}$. If A^{-1} exists, show $\left(A^{-1}\right)^n = (A^n)^{-1}$.

The equation is true if $n = 1$. Suppose it is true for n. Then by the induction hypothesis,

$$\begin{aligned} \left(A^{-1}\right)^{n+1} &= \left(A^{-1}\right)^{n} A^{-1} = \left(A^{n}\right)^{-1} A^{-1} \\ &= \left(A\left(A^{n}\right)\right)^{-1} = \left(A^{n+1}\right)^{-1}. \end{aligned}$$

Part IV

LU Decomposition, Subspaces, Linear Transformations

Outcomes

A. Find the LU factorization of a matrix.

B. Use the LU factorization of a matrix to solve a system of linear equations.

*C. Find the P^TLU factorization of a matrix.

*D. Use that P^TLU factorization of a matrix to solve a system of linear equations.

*E. Find the inverse of a matrix using the LU factorization.

Reading: Linear Algebra 3.4

Outcome Mapping:

A. 7-12,13-14

B. 1-6

*C. 19-22,23-25

*D. 27-28

*E. 15-18,30

A. Define subspace of $\mathbb{R}^n$. Determine whether or not a given set of vectors forms a subspace of $\mathbb{R}^n$.

B. Define row space, column space, and null space for a matrix. Determine whether or not a given vector is in one of these spaces.

C. Define basis and dimension. Given a subspace, determine its dimension and a basis. Verify whether or not a given set of vectors is a basis for the subspace.

D. Define rank and nullity. Determine the rank and nullity of a given matrix.

*E. Prove and recall theorems involving the rank, nullity, and invertibility of matrices.

*F. Find the coordinates of a vector with respect to a given basis.

Reading: Linear Algebra 3.5
Outcome Mapping:

A. 1-10

B. 11-16

C. 17-20,21-26,27-30,31-34,45-48

D. 35-42,43-44

*E. 55-64

*F. 49-50

A. Define linear transformation. Determine whether or not a given transformation is linear.

B. Determine the matrix that represents a given linear transformation of vectors.

C. Prove and recall theorems involving linear transformations.

*D. Find compositions and inverses of linear transformations.

Reading: Linear Algebra 3.6

Outcome Mapping:

A. 1-10,46-51

B. 11-14,15-28

C. 29,40-45,52-55

*D. 30-35,36-39

The LU Factorization 25 Sept.

8.0.2 Definition Of An LU Decomposition

An LU decomposition of a matrix involves writing the given matrix as the product of a lower triangular matrix which has the main diagonal consisting entirely of ones L, and an upper triangular matrix U in the indicated order. This is the version discussed here but it is sometimes the case that the L has numbers other than 1 down the main diagonal. It is still a useful concept. The L goes with "lower" and the U with "upper". It turns out many matrices can be written in this way and when this is possible, people get excited about slick ways of solving the system of equations, $A\mathbf{x} = \mathbf{y}$. It is for this reason that you want to study the LU decomposition. It allows you to work only with triangular matrices. It turns out that it takes about $2n^3/3$ operations to use Gauss elimination but only $n^3/3$ to obtain an LU factorization.

First it should be noted not all matrices have an LU decomposition and so we will emphasize the techniques for achieving it rather than formal proofs.

Example 8.0.3 *Can you write* $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ *in the form* LU *as just described?*

To do so you would need

$$\begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix} \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} = \begin{pmatrix} a & b \\ xa & xb+c \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Therefore, $b = 1$ and $a = 0$. Also, from the bottom rows, $xa = 1$ which can't happen and have $a = 0$. Therefore, you can't write this matrix in the form LU. It has no LU decomposition. This is what we mean above by saying the method lacks generality.

8.0.3 Finding An LU Decomposition By Inspection

Which matrices have an LU decomposition? It turns out it is those whose row reduced echelon form can be achieved without switching rows and which only involve row operations of type 3 in which row j is replaced with a multiple of row i added to row j for $i < j$.

Example 8.0.4 *Find an* LU *decomposition of* $A = \begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 3 & 2 & 1 \\ 2 & 3 & 4 & 0 \end{pmatrix}$.

One way to find the LU decomposition is to simply look for it directly. You need

$$\begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 3 & 2 & 1 \\ 2 & 3 & 4 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ x & 1 & 0 \\ y & z & 1 \end{pmatrix} \begin{pmatrix} a & d & h & j \\ 0 & b & e & i \\ 0 & 0 & c & f \end{pmatrix}.$$

Then multiplying these you get

$$\begin{pmatrix} a & d & h & j \\ xa & xd+b & xh+e & xj+i \\ ya & yd+zb & yh+ze+c & yj+iz+f \end{pmatrix}$$

and so you can now tell what the various quantities equal. From the first column, you need $a = 1, x = 1, y = 2$. Now go to the second column. You need $d = 2, xd + b = 3$ so $b = 1, yd + zb = 3$ so $z = -1$. From the third column, $h = 0, e = 2, c = 6$. Now from the fourth column, $j = 2, i = -1, f = -5$. Therefore, an LU decomposition is

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 & 2 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & 6 & -5 \end{pmatrix}.$$

You can check whether you got it right by simply multiplying these two.

8.0.4 Using Multipliers To Find An LU Decomposition

There is also a convenient procedure for finding an LU decomposition. It turns out that it is only necessary to keep track of the **multipliers** which are used to row reduce to upper triangular form. This procedure is described in the following examples.

Example 8.0.5 *Find an LU decomposition for* $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & -4 \\ 1 & 5 & 2 \end{pmatrix}$

Write the matrix next to the identity matrix as shown.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & -4 \\ 1 & 5 & 2 \end{pmatrix}.$$

The process involves doing row operations to the matrix on the right while simultaneously updating successive columns of the matrix on the left. First take -2 times the first row and add to the second in the matrix on the right.

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -10 \\ 1 & 5 & 2 \end{pmatrix}$$

Note the way we updated the matrix on the left. We put a 2 in the second entry of the first column because we used -2 times the first row added to the second row. Now replace the third row in the matrix on the right by -1 times the first row added to the third. Thus the next step is

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -10 \\ 0 & 3 & -1 \end{pmatrix}$$

Finally, we will add the second row to the bottom row and make the following changes

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -10 \\ 0 & 0 & -11 \end{pmatrix}.$$

At this point, we stop because the matrix on the right is upper triangular. An LU decomposition is the above.

The justification for this gimmick is in my linear algebra book on the web.

Example 8.0.6 *Find an LU decomposition for* $A = \begin{pmatrix} 1 & 2 & 1 & 2 & 1 \\ 2 & 0 & 2 & 1 & 1 \\ 2 & 3 & 1 & 3 & 2 \\ 1 & 0 & 1 & 1 & 2 \end{pmatrix}$.

We will use the same procedure as above. However, this time we will do everything for one column at a time. First multiply the first row by (-1) and then add to the last row. Next take (-2) times the first and add to the second and then (-2) times the first and add to the third.

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

This finishes the first column of L and the first column of U. Now take $-(1/4)$ times the second row in the matrix on the right and add to the third followed by $-(1/2)$ times the second added to the last.

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

This finishes the second column of L as well as the second column of U. Since the matrix on the right is upper triangular, stop. The LU decomposition has now been obtained. This technique is called Dolittle's method.

This process is entirely typical of the general case. The matrix U is just the first upper triangular matrix you come to in your quest for the row reduced echelon form using only the row operation which involves replacing a row by itself added to a multiple of another row. The matrix, L is what you get by updating the identity matrix as illustrated above.

You should note that for a square matrix, the number of row operations necessary to reduce to LU form is about half the number needed to place the matrix in row reduced echelon form. This is why an LU decomposition is of interest in solving systems of equations.

8.0.5 Solving Systems Using The LU Decomposition

The reason people care about the LU decomposition is it allows the quick solution of systems of equations. Here is an example.

Example 8.0.7 *Suppose you want to find the solutions to* $\begin{pmatrix} 1 & 2 & 3 & 2 \\ 4 & 3 & 1 & 1 \\ 1 & 2 & 3 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$.

Of course one way is to write the augmented matrix and grind away. However, this involves more row operations than the computation of the LU decomposition and it turns out that the LU decomposition can give the solution quickly. Here is how. The following is an LU decomposition for the matrix.

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 4 & 3 & 1 & 1 \\ 1 & 2 & 3 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix}.$$

Let $U\mathbf{x} = \mathbf{y}$ and consider $L\mathbf{y} = \mathbf{b}$ where in this case, $\mathbf{b} = (1,2,3)^T$. Thus

$$\begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

which yields very quickly that $\mathbf{y} = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$. Now you can find $\mathbf{x}$ by solving $U\mathbf{x} = \mathbf{y}$. Thus in this case,

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$$

which yields

$$\mathbf{x} = \begin{pmatrix} -\frac{3}{5} + \frac{7}{5}t \\ \frac{9}{5} - \frac{11}{5}t \\ t \\ -1 \end{pmatrix}, t \in \mathbb{R}.$$

Rank Of A Matrix 26,27 Sept.

Quiz

1. Here is a matrix. Find an LU factorization.

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

2. An LU factorization for $\begin{pmatrix} 1 & 2 & 0 & 3 \\ -1 & -4 & -4 & 1 \\ -1 & -5 & 1 & 6 \end{pmatrix}$ is

$$\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & \frac{3}{2} & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 & 3 \\ 0 & -2 & -4 & 4 \\ 0 & 0 & 7 & 3 \end{pmatrix}$$

Use this to solve the system

$$\begin{pmatrix} 1 & 2 & 0 & 3 \\ -1 & -4 & -4 & 1 \\ -1 & -5 & 1 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Show your work. You must use the LU factorization to receive any credit.

9.1 The Row Reduced Echelon Form Of A Matrix

Recall the **row operations** used to solve a system of equations which were presented earlier.

Definition 9.1.1 *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.*

Recall that putting a matrix in row reduced echelon form involves doing row operations as described on Page 88. In this section we review the description of the row reduced echelon form and prove the row reduced echelon form for a given matrix is unique. That is, every matrix can be row reduced to a unique row reduced echelon form. Of course this is not true

of the echelon form. The significance of this is that it becomes possible to use the definite article in referring to **the** row reduced echelon form and hence important conclusions about the original matrix may be logically deduced from an examination of its unique row reduced echelon form. Also recall the definition of linear combination and span.

Definition 9.1.2 *Let* $\mathbf{v}_1, \cdots, \mathbf{v}_k, \mathbf{u}$ *be vectors. Then* $\mathbf{u}$ *is said to be a* ***linear combination*** *of the vectors* $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ *if there exist scalars,* $c_1, \cdots, c_k$ *such that*

$$\mathbf{u} = \sum_{i=1}^{k} c_i \mathbf{v}_i.$$

The collection of all linear combinations of the vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ *is known as the* ***span*** *of these vectors and is written as* $\operatorname{span}(\mathbf{v}_1, \cdots, \mathbf{v}_k)$.

Another way to say the same thing as expressed in the earlier definition of row reduced echelon form found on Page 86 is the following which is a more useful description when proving the major assertions about the row reduced echelon form.

Definition 9.1.3 *Let* $\mathbf{e}_i$ *denote the column vector which has all zero entries except for the* i^{th} *slot which is one. An* $m \times n$ *matrix is said to be in* ***row reduced echelon form*** *if, in viewing successive columns from left to right, the first nonzero column encountered is* $\mathbf{e}_1$ *and if you have encountered* $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$, *the next column is either* $\mathbf{e}_{k+1}$ *or is a linear combination of the vectors,* $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$.

Theorem 9.1.4 *Let* A *be an* $m \times n$ *matrix. Then* A *has a row reduced echelon form determined by a simple process.*

Proof: Viewing the columns of A from left to right take the first nonzero column. Pick a nonzero entry in this column and switch the row containing this entry with the top row of A. Now divide this new top row by the value of this nonzero entry to get a 1 in this position and then use row operations to make all entries below this entry equal to zero. Thus the first nonzero column is now $\mathbf{e}_1$. Denote the resulting matrix by A_1. It has been obtained from A through a sequence of row operations.

Consider the submatrix of A_1 to the right of this column and below the first row. Do exactly the same thing for this submatrix that was done for A. This time the $\mathbf{e}_1$ will refer to $\mathbb{F}^{m-1}$. Use the first 1 obtained by the above process which is in the top row of this submatrix and row operations to zero out every entry above it in the rows of A_1. Call the resulting matrix, A_2. Thus A_2 satisfies the conditions of the above definition up to the column just encountered. Continue this way till every column has been dealt with and the result must be in row reduced echelon form. This proves the theorem.

The following diagram illustrates the above procedure. Say the matrix looked something like the following.

$$\begin{pmatrix} 0 & * & * & * & * & * & * \\ 0 & * & * & * & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & * & * & * & * & * & * \end{pmatrix}$$

First step would yield something like

$$\begin{pmatrix} 0 & 1 & * & * & * & * & * \\ 0 & 0 & * & * & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & * & * & * & * & * \end{pmatrix}$$

For the second step you look at the lower right corner as described,

$$\begin{pmatrix} * & * & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ * & * & * & * & * \end{pmatrix}$$

and if the first column consists of all zeros but the next one is not all zeros, you would get something like this.

$$\begin{pmatrix} 0 & 1 & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & * & * & * \end{pmatrix}$$

Thus, after zeroing out the term in the top row above the 1, you get the following for the next step in the computation of the row reduced echelon form for the original matrix.

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

Next you look at the lower right matrix below the top two rows and to the right of the first four columns and repeat the process.

Recall the following definition which was discussed earlier.

Definition 9.1.5 *The first **pivot column** of A is the first nonzero column of A. The next pivot column is the first column after this which becomes $\mathbf{e}_2$ in the row reduced echelon form. The third is the next column which becomes $\mathbf{e}_3$ in the row reduced echelon form and so forth.*

There are three choices for row operations at each step in the above theorem. A natural question is whether the same row reduced echelon matrix always results in the end from following the above algorithm applied in any way. The next corollary says this is the case. To prove this corollary, the following fundamental lemma will be used.

In rough terms, this lemma states that **linear relationships** between columns in a matrix are preserved by row operations.

Lemma 9.1.6 *Let A and B be two $m \times n$ matrices and suppose B results from a row operation applied to A. Then the k^{th} column of B is a linear combination of the $i_1, \cdots, i_r$ columns of B if and only if the k^{th} column of A is a linear combination of the $i_1, \cdots, i_r$ columns of A. Furthermore, the scalars in the linear combination are the same. (The linear relationship between the k^{th} column of A and the $i_1, \cdots, i_r$ columns of A is the same as the linear relationship between the k^{th} column of B and the $i_1, \cdots, i_r$ columns of B.)*

Proof: Let A equal the following matrix in which the $\mathbf{a}_k$ are the columns

$$\begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{pmatrix}$$

and let B equal the following matrix in which the columns are given by the $\mathbf{b}_k$

$$\begin{pmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \end{pmatrix}$$

Then by Theorem 7.3.6 on Page 144 $\mathbf{b}_k = E\mathbf{a}_k$ where E is an elementary matrix. Suppose then that one of the columns of A is a linear combination of some other columns of A. Say

$$\mathbf{a}_k = \sum_{r \in S} c_r \mathbf{a}_r.$$

Then multiplying by E,

$$\mathbf{b}_k = E\mathbf{a}_k = \sum_{r\in S} c_r E\mathbf{a}_r = \sum_{r\in S} c_r \mathbf{b}_r.$$

This proves the lemma.

Definition 9.1.7 *Two matrices are said to be* ***row equivalent*** *if one can be obtained from the other by a sequence of row operations.*

It has been shown above in Theorem 9.1.4 that every matrix is row equivalent to one which is in row reduced echelon form.

Corollary 9.1.8 *The row reduced echelon form is unique. That is if B, C are two matrices in row reduced echelon form and both are row equivalent to A, then $B = C$.*

Proof: Suppose B and C are both row reduced echelon forms for the matrix, A. Then they clearly have the same zero columns since row operations leave zero columns unchanged. If B has the sequence $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_r$ occuring for the first time in the positions, $i_1, i_2, \cdots, i_r$ the description of the row reduced echelon form means that if $\mathbf{b}_k$ is the k^{th} column of B such that $i_{j-1} < k < i_j$ then $\mathbf{b}_k$ is a linear combination of the columns in positions $i_1, i_2, \cdots, i_{k-1}$. By Lemma 9.1.6 the same is true for $\mathbf{c}_k$, the k^{th} column of C. Therefore, $\mathbf{c}_k$ is not equal to $\mathbf{e}_j$ for any j because $\mathbf{e}_j$ is not obtained as a linear combinations of the $\mathbf{e}_i$ for $i < j$. It follows the $\mathbf{e}_j$ for C can only occur in positions $i_1, i_2, \cdots, i_r$. Furthermore, position i_j in C must contain $\mathbf{e}_j$ because if not, then $\mathbf{c}_{i_j}$ would be a linear combination of $\mathbf{e}_1, \cdots, \mathbf{e}_{j-1}$ in C but not in B, thus contradicting Lemma 9.1.6. Therefore, both B and C have the sequence $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_r$ occuring for the first time in the positions, $i_1, i_2, \cdots, i_r$. By Lemma 9.1.6, the columns between the i_k and i_{k+1} position are linear combinations involving the same scalars of the columns in the $i_1, \cdots, i_k$ position. This is equivalent to the assertion that each of these columns is identical and this proves the corollary.

Example 9.1.9 *Find the row reduced echelon form of the matrix,*

$$\begin{pmatrix} 0 & 0 & 2 & 3 \\ 0 & 2 & 0 & 1 \\ 0 & 1 & 1 & 5 \end{pmatrix}$$

The first nonzero column is the second in the matrix. Switch the third and first rows to obtain

$$\begin{pmatrix} 0 & 1 & 1 & 5 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 3 \end{pmatrix}$$

Now we multiply the top row by -2 and add to the second.

$$\begin{pmatrix} 0 & 1 & 1 & 5 \\ 0 & 0 & -2 & -9 \\ 0 & 0 & 2 & 3 \end{pmatrix}$$

Next, add the second row to the bottom and then divide the bottom row by -6

$$\begin{pmatrix} 0 & 1 & 1 & 5 \\ 0 & 0 & -2 & -9 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Next use the bottom row to obtain zeros in the last column above the 1 and divide the second row by -2

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

Finally, add -1 times the middle row to the top.

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

This is in row reduced echelon form.

Example 9.1.10 *Find the row reduced echelon form for the matrix,*

$$\begin{pmatrix} 1 & 2 & 0 & 2 \\ -1 & 3 & 4 & 3 \\ 0 & 5 & 4 & 5 \end{pmatrix}$$

You should verify that the row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & -\frac{8}{5} & 0 \\ 0 & 1 & \frac{4}{5} & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

9.2 The Rank Of A Matrix

9.2.1 The Definition Of Rank

To begin, here is a definition to introduce some terminology.

Definition 9.2.1 *Let A be an $m \times n$ matrix. The **column space** of A is the subspace of $\mathbb{F}^m$ spanned by the columns. The **row space** is the subspace of $\mathbb{F}^n$ spanned by the rows.*

Earlier the rank was defined to be the number of nonzero rows in the row reduced echelon form. This is fine. However, it is useful to tie this in to the notion of spans of columns and rows.

Definition 9.2.2 *The **row space** of a matrix, A is the span of the rows, denoted as $\mathrm{row}(A)$ and the **column space** of a matrix is the span of the columns, denoted as $\mathrm{col}(A)$. The **row rank** of a matrix is the number of nonzero rows in the row reduced echelon form and the **column rank** is the number of columns in the row reduced echelon form which are one of the $\mathbf{e}_k$. Thus the column rank equals the number of pivot columns. Thus the row rank equals the column rank. This is also called the rank of the matrix. The rank of a matrix, A is denoted by* $\mathrm{rank}(A)$.

Example 9.2.3 *Consider the matrix,*

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix}$$

What is its rank?

The row reduced echelon form is $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \end{pmatrix}$. There is one pivot column so the column rank is 1. There is one nonzero row so the row rank is 1.

Example 9.2.4 *Find the rank of the matrix,*

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 0 \\ -4 & 3 & 2 & 1 & 2 \\ 3 & 2 & 1 & 6 & 5 \\ 4 & -3 & -2 & 1 & 7 \end{pmatrix}.$$

From the above definition, all you have to do is find the row reduced echelon form and then count up the number of nonzero rows. But the row reduced echelon form of this matrix is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & -\frac{17}{4} \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & -\frac{45}{4} \\ 0 & 0 & 0 & 1 & \frac{9}{2} \end{pmatrix}$$

and so the rank of this matrix is 4.

Example 9.2.5 *Find the rank of the matrix*

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 0 \\ -4 & 3 & 2 & 1 & 2 \\ 3 & 2 & 1 & 6 & 5 \\ 0 & 7 & 4 & 10 & 7 \end{pmatrix}$$

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 0 & \frac{3}{2} & \frac{5}{2} \\ 0 & 1 & 0 & -4 & -17 \\ 0 & 0 & 1 & \frac{19}{2} & \frac{63}{2} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and so this time the rank is 3.

9.2.2 Finding The Row And Column Space Of A Matrix

The row reduced echelon form also can be used to obtain an efficient description of the row and column space of a matrix. Of course you can get the column space by simply saying that it equals the span of all the columns but often you can get the column space as the span of fewer columns than this. This is what we mean by an "efficient description". This is illustrated in the next example.

Example 9.2.6 *Find the rank of the following matrix and describe the column and row spaces efficiently.*

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 1 & 3 & 6 & 0 & 2 \\ 3 & 7 & 8 & 6 & 6 \end{pmatrix} \tag{9.1}$$

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & -9 & 9 & 2 \\ 0 & 1 & 5 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Therefore, the rank of this matrix equals 2. All columns of this row reduced echelon form are in

$$\text{span}\left(\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\1\\0\end{pmatrix}\right).$$

For example,

$$\begin{pmatrix}-9\\5\\0\end{pmatrix}=-9\begin{pmatrix}1\\0\\0\end{pmatrix}+5\begin{pmatrix}0\\1\\0\end{pmatrix}.$$

By Lemma 9.1.6, all columns of the original matrix, are similarly contained in the span of the first two columns of **that matrix**. For example, consider the third column of the original matrix.

$$\begin{pmatrix}1\\6\\8\end{pmatrix}=-9\begin{pmatrix}1\\1\\3\end{pmatrix}+5\begin{pmatrix}2\\3\\7\end{pmatrix}.$$

How did I know to use -9 and 5 for the coefficients? This is what Lemma 9.1.6 says! It says linear relationships are all preserved. Therefore, the column space of the original matrix equals the span of the first two columns. This is the desired efficient description of the column space.

What about an efficient description of the row space? When row operations are used, the resulting vectors remain in the row space. Thus the rows in the row reduced echelon form are in the row space of the original matrix. Furthermore, by reversing the row operations, each row of the original matrix can be obtained as a linear combination of the rows in the row reduced echelon form. It follows that the span of the nonzero rows in the row reduced echelon equals the span of the original rows. In the above example, the row space equals the span of the two vectors, $\begin{pmatrix}1&0&-9&9&2\end{pmatrix}$ and $\begin{pmatrix}0&1&5&-3&0\end{pmatrix}$.

Example 9.2.7 *Find the rank of the following matrix and describe the column and row spaces efficiently.*

$$\begin{pmatrix}1&2&1&3&2\\1&3&6&0&2\\1&2&1&3&2\\1&3&2&4&0\end{pmatrix} \tag{9.2}$$

The row reduced echelon form is

$$\begin{pmatrix}1&0&0&0&\frac{13}{2}\\0&1&0&2&-\frac{5}{2}\\0&0&1&-1&\frac{1}{2}\\0&0&0&0&0\end{pmatrix}.$$

and so the rank is 3, the row space is the span of the vectors,

$$\begin{pmatrix}0&0&1&-1&\frac{1}{2}\end{pmatrix},\begin{pmatrix}0&1&0&2&-\frac{5}{2}\end{pmatrix},$$

and

$$\begin{pmatrix}1&0&0&0&\frac{13}{2}\end{pmatrix}$$

and the column space is the span of the first three columns in the **original matrix**,

$$\text{span}\left(\begin{pmatrix}1\\1\\1\\1\end{pmatrix},\begin{pmatrix}2\\3\\2\\3\end{pmatrix},\begin{pmatrix}1\\6\\1\\2\end{pmatrix}\right).$$

Example 9.2.8 *Find the rank of the following matrix and describe the column and row spaces efficiently.*

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

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 1 & 0 & \frac{21}{17} \\ 0 & 1 & 1 & 0 & -\frac{2}{17} \\ 0 & 0 & 0 & 1 & \frac{14}{17} \end{pmatrix}.$$

It follows the rank is three and the column space is the span of the first, second and fourth columns of the **original matrix.**

$$\text{span}\left(\begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 3 \end{pmatrix}\right)$$

while the row space is the span of the vectors $\begin{pmatrix} 0 & 0 & 0 & 1 & \frac{14}{17} \end{pmatrix}, \begin{pmatrix} 0 & 1 & 1 & 0 & -\frac{2}{17} \end{pmatrix}$, and $\begin{pmatrix} 1 & 0 & 1 & 0 & \frac{21}{17} \end{pmatrix}$.

Procedure 9.2.9 *To find the rank of a matrix, obtain the row reduced echelon form for the matrix. Then count the number of nonzero rows or equivalently the number of pivot columns. This is the rank. The row space is the span of the nonzero rows in the row reduced echelon form and the column space is the span of the pivot columns of the* ***original matrix****.*

9.3 Linear Independence And Bases

Quiz

1. Let $\mathbf{u} = (1,2,1)$ and $\mathbf{v} = (1,1,2)$. Find $\text{proj}_{\mathbf{u}}(\mathbf{v})$.
2. Here are two vectors: $\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$. Find an equation of the plane which equals the span of these two vectors.
3. Here is a matrix: $\begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \\ 2 & 1 & 3 \end{pmatrix}$. Find an LU factorization of this matrix.
4. Here are three points: $(1,2,1), (2,1,0), (0,1,1)$. Find the area of the triangle determined by these three points.

9.3.1 Linear Independence And Dependence

First we review the concept of linear independence and dependence which was presented earlier. We define what it means for vectors in $\mathbb{F}^n$ to be linearly independent and then give equivalent descriptions. In the following discussion, the symbol,

$$\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{pmatrix}$$

denotes the matrix which has the vector, $\mathbf{v}_1$ as the first column, $\mathbf{v}_2$ as the second column and so forth until $\mathbf{v}_k$ is the k^{th} column.

Definition 9.3.1 *Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ be vectors in $\mathbb{F}^n$. Then this collection of vectors is said to be* ***linearly independent*** *if whenever*

$$\sum_{i=1}^{k} c_i \mathbf{v}_i = \mathbf{0}$$

it follows each $c_i = 0$. If this condition does not hold, then the set of vectors is said to be dependent.

The following theorem is very important.

Theorem 9.3.2 *A set of vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is linearly independent if and only if none of the vectors is a linear combination of the others.*

Proof: Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is linearly independent. If $\mathbf{v}_k = \sum_{i=1}^{k-1} c_i \mathbf{v}_i$, then $\mathbf{0} = \sum_{i=1}^{k-1} c_i \mathbf{v}_i + (-1)\,\mathbf{v}_k$ and this would mean the vectors are not linearly independent after all. Therefore, $\mathbf{v}_k$ cannot be a linear combination of the other vectors. Similarly, none of the other $\mathbf{v}_i$ can be a linear combination of the other vectors.

Next suppose none of the vectors is a linear combination of the others. If $\sum_{i=1}^{k} c_i \mathbf{v}_i = \mathbf{0}$, then if some $c_l \neq 0$, you could write

$$c_l \mathbf{v}_l = -\sum_{i \neq l} c_i \mathbf{v}_i$$

and then divide by c_l to obtain

$$\mathbf{v}_l = -\sum_{i \neq l} \left(\frac{c_i}{c_l} \right) \mathbf{v}_i$$

showing that one of the vectors is a linear combination of the others which, by assumption, does not happen. Therefore, each of the $c_i = 0$ which shows $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is linearly independent. This proves the theorem.

In words, this says a set of vectors is linearly independent if and only if none of the vectors is "dependent" on the other vectors.

Lemma 9.3.3 *The set of vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\} \subseteq \mathbb{F}^m$ is linearly independent if and only if whenever $\{\mathbf{w}_1, \cdots, \mathbf{w}_r\}$ is a set of vectors in $\mathbb{F}^m$, each of the first k columns of the $n \times (k+r)$ matrix $\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k & \mathbf{w}_1 & \cdots & \mathbf{w}_r \end{pmatrix}$ is a pivot column.*

Proof: Suppose first $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is linearly independent. Then by Theorem 9.3.2 none of the $\mathbf{v}_k$ is a linear combination of the others. It follows each must be a pivot column $\mathbf{v}_k$ becoming $\mathbf{e}_k$ in the row reduced echelon form. If this didn't happen, then you could apply Lemma 9.1.6 and conclude one of the $\mathbf{v}_k$ is a combination of the others.

Next suppose each are pivot columns. Then the row reduced echelon form of the above matrix is of the form

$$\begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_k & \mathbf{w}_1' & \cdots & \mathbf{w}_r' \end{pmatrix}.$$

None of the $\mathbf{e}_k$ is a linear combination of the others and so by Lemma 9.1.6, the same is true of the $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$. In other words $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is independent.

Corollary 9.3.4 *Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ be a set of vectors in $\mathbb{F}^n$. Then if $k > n$, it must be the case that $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is not linearly independent. In other words, if $k > n$, then $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is dependent.*

Proof: If $k > n$, then the columns of $\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{pmatrix}$ cannot each be a pivot column because there are at most n pivot columns due to the fact the matrix has only n rows.

Example 9.3.5 *Determine whether the vectors,* $\left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \\ 0 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 2 \\ -1 \end{pmatrix} \right\}$ *are linearly independent. If they are linearly dependent, exhibit one of the vectors as a linear combination of the others.*

Form the matrix mentioned above.

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

Then the row reduced echelon form of this matrix is

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

Thus not all the columns are pivot columns and so the vectors are not linear independent. Note the fourth column is of the form

$$1 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + 1 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + (-1) \begin{pmatrix} 1 \\ 1 \\ -1 \\ 0 \end{pmatrix}$$

From Lemma 9.1.6, the same linear relationship exists between the columns of the original matrix. Thus

$$1 \begin{pmatrix} 1 \\ 2 \\ 3 \\ 0 \end{pmatrix} + 1 \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix} + (-1) \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 2 \\ -1 \end{pmatrix}.$$

Note the usefullness of the row reduced echelon form in discovering hidden linear relationships in collections of vectors.

Example 9.3.6 *Determine whether the vectors,* $\left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \\ 0 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 2 \\ 0 \end{pmatrix} \right\}$ *are linearly independent. If they are linearly dependent, exhibit one of the vectors as a linear combination of the others.*

The matrix used to find this is

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

The row reduced echelon form is

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

and so every column is a pivot column. Therefore, these vectors are linearly independent and there is no way to obtain one of the vectors as a linear combination of the others.

9.3.2 Subspaces

It turns out that the span of a set of vectors is something called a subspace. What follows is an easier to remember description of subspaces. Furthermore, every such thing is the span of a set of vectors.

Definition 9.3.7 *Let V be a nonempty collection of vectors in $\mathbb{F}^n$. Then V is called a subspace if whenever α, β are scalars and $\mathbf{u}, \mathbf{v}$ are vectors in V, the linear combination, $\alpha\mathbf{u} + \beta\mathbf{v}$ is also in V.*

Theorem 9.3.8 *V is a subspace of $\mathbb{F}^n$ if and only if there exist vectors of V, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ such that $V = \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$.*

Proof: Pick a vector of $V, \mathbf{u}_1$. If $V = \operatorname{span}\{\mathbf{u}_1\}$, then stop. You have found your list of vectors. If $V \neq \operatorname{span}(\mathbf{u}_1)$, then there exists $\mathbf{u}_2$ a vector of V which is not a vector in $\operatorname{span}(\mathbf{u}_1)$. Consider $\operatorname{span}(\mathbf{u}_1, \mathbf{u}_2)$. If $V = \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2)$, stop. Otherwise, pick $\mathbf{u}_3 \notin \operatorname{span}(\mathbf{u}_1, \mathbf{u}_2)$. Continue this way. Here is why $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ is a linearly independent set.

Suppose it is not so. Then you can let l be the largest index such that $\mathbf{u}_l$ is a linear combination of the other vectors. Then

$$\mathbf{u}_l = \sum_{i=1}^{l-1} c_i \mathbf{u}_i + \sum_{j=l+1}^{k} d_j \mathbf{u}_j.$$

By the construction, at least one of the d_j must be nonzero since otherwise, $\mathbf{u}_l$ would be a linear combination of the preceding vectors which is not allowed by the construction. But then you could solve for that $\mathbf{u}_j$ in terms of the other vectors and contradict the choice of l. Therefore, from Corollary 9.3.4 the process stops when k is no larger than n. This proves one half of the theorem.

For the other half, suppose $V = \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ and let $\sum_{i=1}^{k} c_i \mathbf{u}_i$ and $\sum_{i=1}^{k} d_i \mathbf{u}_i$ be two vectors in V. Now let α and β be two scalars. Then

$$\alpha \sum_{i=1}^{k} c_i \mathbf{u}_i + \beta \sum_{i=1}^{k} d_i \mathbf{u}_i = \sum_{i=1}^{k} (\alpha c_i + \beta d_i) \mathbf{u}_i$$

which is one of the things in $\operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ showing that $\operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ has the properties of a subspace. This proves the theorem.

Contained within the proof is the following corollary.

Corollary 9.3.9 *If V is a subspace of $\mathbb{F}^n$, then there exist vectors of $V, \{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ such that $V = \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ and $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ is linearly independent.*

Proof: In the proof we eventually obtain $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ such that $V = \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ and $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ is linearly independent.

The message is that subspaces of $\mathbb{F}^n$ consist of spans of finite, linearly independent collections of vectors of $\mathbb{F}^n$.

9.3.3 The Basis Of A Subspace

It was just shown in Corollary 9.3.9 that every subspace of $\mathbb{F}^n$ is equal to the span of a linearly independent collection of vectors of $\mathbb{F}^n$. Such a collection of vectors is called a basis.

Definition 9.3.10 *Let V be a subspace of $\mathbb{F}^n$. Then $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ is a **basis** for V if the following two conditions hold.*

1. $\mathrm{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k) = V$.
2. *$\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ is linearly independent.*

*The plural of basis is **bases**.*[1]

The main theorem about bases is the following.

Theorem 9.3.11 *Let V be a subspace of $\mathbb{F}^n$ and suppose $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ and $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ are two bases for V. Then $k = m$.*

Proof: Suppose $k < m$. Then since $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ is a basis for V, each $\mathbf{v}_i$ is a linear combination of the vectors of $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$. Consider the matrix

$$\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_k & \mathbf{v}_1 & \cdots & \mathbf{v}_m \end{pmatrix}$$

in which each of the $\mathbf{u}_i$ is a pivot column by Lemma 9.3.3 because the $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ are linearly independent. Therefore, the row reduced echelon form of this matrix is

$$\begin{pmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_k & \mathbf{w}_1 & \cdots & \mathbf{w}_m \end{pmatrix} \tag{9.3}$$

where each $\mathbf{w}_j$ has zeroes below the k^{th} row. This is because of Lemma 9.1.6 which implies each $\mathbf{w}_i$ is a linear combination of the $\mathbf{e}_1, \cdots, \mathbf{e}_k$ due to the fact each $\mathbf{v}_k$ is a linear combination of the $\mathbf{u}_j$ vectors. Discarding the bottom $n-k$ rows of zeroes in the above, yields the matrix,

$$\begin{pmatrix} \mathbf{e}'_1 & \cdots & \mathbf{e}'_k & \mathbf{w}'_1 & \cdots & \mathbf{w}'_m \end{pmatrix}$$

in which all vectors are in $\mathbb{F}^k$. Since $m > k$, it follows from Corollary 9.3.4 that the vectors, $\{\mathbf{w}'_1, \cdots, \mathbf{w}'_m\}$ are dependent. Therefore, some $\mathbf{w}'_j$ is a linear combination of the other $\mathbf{w}'_i$. Therefore, $\mathbf{w}_j$ is a linear combination of the other $\mathbf{w}_i$ in 9.3. By Lemma 9.1.6 again, the same linear relationship exists between the $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ showing that $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ is not linearly independent and contradicting the assumption that $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ is a basis. It follows $k \leq m$. Similarly, $m \leq k$. This proves the theorem.

The following definition can now be stated.

Definition 9.3.12 *Let V be a subspace of $\mathbb{F}^n$. Then the **dimension** of V is defined to be the number of vectors in a basis.*

Corollary 9.3.13 *The dimension of $\mathbb{F}^n$ is n.*

Proof: You only need to exhibit a basis for $\mathbb{F}^n$ which has n vectors. Such a basis is $\{\mathbf{e}_1, \cdots, \mathbf{e}_n\}$.

Corollary 9.3.14 *Let A be an $m \times n$ matrix. Then $\mathrm{rank}(A)$ equals the dimension of $\mathrm{col}(A)$ and this equals the dimension of $\mathrm{row}(A)$.*

[1] To see why the plural of basis is bases, try to say basiss. It involves much hissing.

Proof: The rank of A equals the number of pivot columns. By Lemma 9.1.6 these pivot columns are linearly independent and span $col(A)$. Therefore, the dimension of $col(A)$ equals the rank of A. The number of nonzero rows in the row reduced echelon form equals the rank of A also. Furthermore, these rows are independent and span $row(A)$. Thus the rank of A equals the dimension of the row space as claimed.

From this corollary, the following is obvious.

Corollary 9.3.15 *rank*$(A) = \text{rank}\left(A^T\right)$.

Corollary 9.3.16 *Suppose* $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ *is linearly independent and each* $\mathbf{v}_i$ *is a vector in* $\mathbb{F}^n$. *Then* $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ *is a basis for* $\mathbb{F}^n$. *Suppose* $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ *spans* $\mathbb{F}^n$. *Then* $m \geq n$. *If* $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ *spans* $\mathbb{F}^n$, *then* $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ *is linearly independent.*

Proof: First suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is linearly independent. Let $\mathbf{u}$ be a vector of $\mathbb{F}^n$ and consider the matrix,

$$\begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n & \mathbf{u} \end{pmatrix}.$$

By Lemma 9.3.3, on Page 167 each $\mathbf{v}_i$ is a pivot column, the row reduced echelon form is

$$\begin{pmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_n & \mathbf{w} \end{pmatrix}$$

and so, since $\mathbf{w}$ is in span $(\mathbf{e}_1, \cdots, \mathbf{e}_n)$, it follows from Lemma 9.1.6 that $\mathbf{u}$ is in span $(\mathbf{v}_1, \cdots, \mathbf{v}_n)$. Therefore, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis as claimed.

For the second claim, if any of the $\mathbf{v}_i$ is a linear combination of the others, then delete that vector from the list. This yields a shorter list of vectors which has the same span. Now do the same with this shorter list eventually obtaining vectors $\{\mathbf{v}_1', \cdots, \mathbf{v}_l'\}$ with $l \leq m$ that spans $\mathbb{F}^n$ and has the property that no vector is a linear combination of the others. Thus $\{\mathbf{v}_1', \cdots, \mathbf{v}_l'\}$ is a basis for $\mathbb{F}^n$ and so by Theorem 9.3.11, $l = n$. Therefore, $m \geq l = n$.

Finally consider the third claim. If $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is not linearly independent, then some vector is a linear combination of the others. Delete that vector from the list. The new list of vectors still has the same span. If it is linearly independent, stop. If not, some vector is a linear combination of the others. Delete that vector. This does not change the span. Continue this way, finally obtaining a shorter list of vectors, $\{\mathbf{v}_1', \cdots, v_m'\}$ which spans $\mathbb{F}^n$ and is also linearly independent. But then this contradicts Theorem 9.3.11 because this would yield two bases having different sizes. This proves the corollary.

By way of review, here are a few more examples of the sort worked earlier on which you can use the new terminology, $row(A)$, $col(A)$ and basis.

Example 9.3.17 *Find the rank of the following matrix. If the rank is* r, *identify* r *columns* ***in the original matrix*** *which have the property that every other column may be written as a linear combination of these. Also find a basis for the row and column spaces of the matrices.*

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 5 & -4 & -1 \\ -2 & 3 & 1 & 0 \end{pmatrix}$$

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 0 & \frac{27}{70} \\ 0 & 1 & 0 & \frac{1}{10} \\ 0 & 0 & 1 & \frac{33}{70} \end{pmatrix}$$

and so the rank of the matrix is 3. A basis for the column space is the first three columns of the **original matrix.** I know they span because the first three columns of the row reduced

echelon form above span the column space of that matrix. They are linearly independent because the first three columns of the row reduced echelon form are linearly independent. By Lemma 9.1.6 all linear relationships are preserved and so these first three vectors form a basis for the column space. The four rows of the row reduced echelon form form a basis for the row space of the original matrix.

Example 9.3.18 *Find the rank of the following matrix. If the rank is r, identify r columns* ***in the original matrix*** *which have the property that every other column may be written as a linear combination of these. Also find a basis for the row and column spaces of the matrices.*

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

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 1 & 0 & -\frac{1}{7} \\ 0 & 1 & 1 & 0 & \frac{4}{7} \\ 0 & 0 & 0 & 1 & -\frac{11}{42} \end{pmatrix}.$$

A basis for the column space of this row reduced echelon form is the first second and fourth columns. Therefore, a basis for the column space in the **original matrix** is the first second and fourth columns. The rank of the matrix is 3. A basis for the row space of the original matrix is the columns of the row reduced echelon form.

9.3.4 Finding The Null Space Or Kernel Of A Matrix

Let A be an $m \times n$ matrix.

Definition 9.3.19 $\ker(A)$, *also referred to as the null space of A is defined as follows.*

$$\ker(A) = \{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$$

and to find $\ker(A)$ *one must solve the system of equations* $A\mathbf{x} = \mathbf{0}$. *This is also denoted as*

$$\text{null}(A).$$

That is, $\text{null}(A) = \ker(A) = \{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$ *which equals all the vectors which A sends to* $\mathbf{0}$.

This is not new! There is just some new terminology being used. To repeat, $\ker(A)$ is the solution to the system $A\mathbf{x} = \mathbf{0}$.

Example 9.3.20 *Let*

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 1 \\ 2 & 3 & 3 \end{pmatrix}.$$

Find $\ker(A)$.

You need to solve the equation $A\mathbf{x} = \mathbf{0}$. To do this you write the augmented matrix and then obtain the row reduced echelon form and the solution. The augmented matrix is

$$\left(\begin{array}{ccc|c} 1 & 2 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ 2 & 3 & 3 & 0 \end{array}\right)$$

Next place this matrix in row reduced echelon form,

$$\left(\begin{array}{ccc|c} 1 & 0 & 3 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Note that x_1 and x_2 are basic variables while x_3 is a free variable. Therefore, the solution to this system of equations, $A\mathbf{x} = \mathbf{0}$ is given by

$$\left(\begin{array}{c} 3t \\ t \\ t \end{array}\right) : t \in \mathbb{R}.$$

Example 9.3.21 *Let*

$$A = \left(\begin{array}{ccccc} 1 & 2 & 1 & 0 & 1 \\ 2 & -1 & 1 & 3 & 0 \\ 3 & 1 & 2 & 3 & 1 \\ 4 & -2 & 2 & 6 & 0 \end{array}\right)$$

Find the null space of A.

You need to solve the equation, $A\mathbf{x} = \mathbf{0}$. The augmented matrix is

$$\left(\begin{array}{ccccc|c} 1 & 2 & 1 & 0 & 1 & 0 \\ 2 & -1 & 1 & 3 & 0 & 0 \\ 3 & 1 & 2 & 3 & 1 & 0 \\ 4 & -2 & 2 & 6 & 0 & 0 \end{array}\right)$$

Its row reduced echelon form is

$$\left(\begin{array}{ccccc|c} 1 & 0 & \frac{3}{5} & \frac{6}{5} & \frac{1}{5} & 0 \\ 0 & 1 & \frac{1}{5} & -\frac{3}{5} & \frac{2}{5} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

It follows x_1 and x_2 are basic variables and x_3, x_4, x_5 are free variables. Therefore, $\ker(A)$ is given by

$$\left(\begin{array}{c} \left(-\frac{3}{5}\right) s_1 + \left(\frac{-6}{5}\right) s_2 + \left(\frac{1}{5}\right) s_3 \\ \left(-\frac{1}{5}\right) s_1 + \left(\frac{3}{5}\right) s_2 + \left(-\frac{2}{5}\right) s_3 \\ s_1 \\ s_2 \\ s_3 \end{array}\right) : s_1, s_2, s_3 \in \mathbb{R}.$$

We write this in the form

$$s_1 \left(\begin{array}{c} -\frac{3}{5} \\ -\frac{1}{5} \\ 1 \\ 0 \\ 0 \end{array}\right) + s_2 \left(\begin{array}{c} \frac{-6}{5} \\ \frac{3}{5} \\ 0 \\ 1 \\ 0 \end{array}\right) + s_3 \left(\begin{array}{c} \frac{1}{5} \\ -\frac{2}{5} \\ 0 \\ 0 \\ 1 \end{array}\right) : s_1, s_2, s_3 \in \mathbb{R}.$$

In other words, the null space of this matrix equals the span of the three vectors above. Thus

$$\ker(A) = \text{span}\left(\left(\begin{array}{c} -\frac{3}{5} \\ -\frac{1}{5} \\ 1 \\ 0 \\ 0 \end{array}\right), \left(\begin{array}{c} \frac{-6}{5} \\ \frac{3}{5} \\ 0 \\ 1 \\ 0 \end{array}\right), \left(\begin{array}{c} \frac{1}{5} \\ -\frac{2}{5} \\ 0 \\ 0 \\ 1 \end{array}\right)\right).$$

This is the same as

$$\ker(A) = \text{span}\left(\left\{\begin{pmatrix}\frac{3}{5}\\ \frac{1}{5}\\ -1\\ 0\\ 0\end{pmatrix}, \begin{pmatrix}\frac{6}{5}\\ \frac{-3}{5}\\ 0\\ -1\\ 0\end{pmatrix}, \begin{pmatrix}\frac{-1}{5}\\ \frac{2}{5}\\ 0\\ 0\\ -1\end{pmatrix}\right\}\right).$$

Notice also that the three vectors above are linearly independent and so the dimension of $\ker(A)$ is 3. This is generally the way it works. The number of free variables equals the dimension of the null space while the number of basic variables equals the number of pivot columns which equals the rank. We state this in the following theorem.

Definition 9.3.22 *The dimension of the null space of a matrix is called the* ***nullity***[2] *and written as nullity*(A).

Theorem 9.3.23 *Let A be an $m \times n$ matrix. Then* $\text{rank}(A) + \textit{nullity}(A) = n$.

This implies the following corollary.

Corollary 9.3.24 *Let A be an $n \times n$ matrix. Then A is onto if and only if A is one to one.*

The following theorem is an interesting review of the transpose of a matrix.

Theorem 9.3.25 *Let A be a real $m \times n$ matrix. Then* $\text{rank}\left(A^TA\right) = \text{rank}(A)$ *and A^TA is invertible if and only if* $\text{rank}(A) = n$.

Proof: There are various ways to show this. From Theorem 9.3.23

$$n = \text{rank}(A) + \text{nullity}(A) = \text{rank}\left(A^TA\right) + \text{nullity}\left(A^TA\right).$$

I will show $\text{null}(A) = \text{null}\left(A^TA\right)$ because this implies the two nullitys above are equal. Suppose $A\mathbf{x} = \mathbf{0}$. Then $A^TA\mathbf{x} = \mathbf{0}$ also. Hence $\text{null}(A) \subseteq \text{null}\left(A^TA\right)$. Next suppose $A^TA\mathbf{x} = \mathbf{0}$. Then

$$\mathbf{x}^TA^TA\mathbf{x} = (A\mathbf{x})^TA\mathbf{x} = A\mathbf{x}\cdot A\mathbf{x}$$

and so $A\mathbf{x} = \mathbf{0}$. Hence $\text{null}(A) \supseteq \text{null}\left(A^TA\right)$.

If A^TA is invertible, then this implies A is one to one and so the columns of A are independent. Hence $\text{rank}(A) = n$. If $\text{rank}(A) = n$, then the dimension of the column space is n and so since the column vectors span the column space, they are a basis for it. In particular they are independent. Hence A is one to one. It follows as above that A^TA is also one to one mapping $\mathbb{F}^n$ to $\mathbb{F}^n$. The n columns of A^TA are linearly independent because the matrix is one to one. Therefore, by Corollary 9.3.16 these columns are a basis for $\mathbb{F}^n$ and so A^TA is both onto and one to one. Hence it is invertible.

9.3.5 Rank And Existence Of Solutions To Linear Systems*

Consider the linear system of equations,

$$A\mathbf{x} = \mathbf{b} \tag{9.4}$$

[2] Isn't it amazing how many different words are available for use in linear algebra?

where A is an $m \times n$ matrix, $\mathbf{x}$ is a $n \times 1$ column vector, and $\mathbf{b}$ is an $m \times 1$ column vector. Suppose

$$A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}$$

where the $\mathbf{a}_k$ denote the columns of A. Then if $\mathbf{x} = (x_1, \cdots, x_n)^T$ is a solution of the system 9.4, it follows

$$x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{b}$$

which says that $\mathbf{b}$ is a vector in $\text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_n)$. This shows that there exists a solution to the system, 9.4 if and only if $\mathbf{b}$ is contained in $\text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_n)$. In words, there is a solution to 9.4 if and only if $\mathbf{b}$ is in the column space of A. In terms of rank, the following proposition describes the situation.

Proposition 9.3.26 *Let A be an $m \times n$ matrix and let $\mathbf{b}$ be an $m \times 1$ column vector. Then there exists a solution to 9.4 if and only if*

$$\text{rank}\begin{pmatrix} A & | & \mathbf{b} \end{pmatrix} = \text{rank}(A). \tag{9.5}$$

Proof: Place $\begin{pmatrix} A & | & \mathbf{b} \end{pmatrix}$ and A in row reduced echelon form, respectively B and C. If the above condition on rank is true, then both B and C have the same number of nonzero rows. In particular, you cannot have a row of the form

$$\begin{pmatrix} 0 & \cdots & 0 & \blacksquare \end{pmatrix}$$

where $\blacksquare \neq 0$ in B. Therefore, there will exist a solution to the system.

Conversely, suppose there exists a solution. This means there cannot be such a row in B described above. Therefore, B and C must have the same number of zero rows and so they have the same number of nonzero rows. Therefore, the rank of the two matrices in 9.5 is the same.

9.3.6 Exercises With Answers

1. Find the rank of the following matrices. If the rank is r, identify r columns **in the original matrix** which have the property that every other column may be written as a linear combination of these. Also find a basis for the row and column spaces of the matrices.

   (a) $\begin{pmatrix} 9 & 2 & 0 \\ 3 & 7 & 1 \\ 6 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}$

   From using row operations we obtain the row reduced echelon form which is

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

   Therefore, a basis for the column space of the original matrix is the first three columns of the original matrix. A basis for the row space is just $\begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$, and $\begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$.

(b) $\begin{pmatrix} 3 & 0 & 3 \\ 10 & 9 & 1 \\ 1 & 1 & 0 \\ 2 & 2 & 0 \end{pmatrix}$

In this case the row reduced echelon form is

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

and so a basis for the column space of the original matrix consists of the first two columns of the original matrix and a basis for the row space is $\begin{pmatrix} 1 & 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 & -1 \end{pmatrix}$.

(c) $\begin{pmatrix} 0 & 1 & 7 & 8 & 1 & 9 & 2 \\ 0 & 3 & 2 & 5 & 1 & 6 & 8 \\ 0 & 1 & 1 & 2 & 0 & 2 & 3 \\ 0 & 2 & 1 & 3 & 0 & 3 & 4 \end{pmatrix}$

The row reduced echelon form of this matrix is

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

and so a basis for the column space of the original matrix consists of the second, third, fifth, and seventh columns of the original matrix. A basis for the row space consists of the rows of this last matrix in row reduced echelon form.

2. Let H denote span $\left(\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 4 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right)$. Find the dimension of H and determine a basis.

   Make these the columns of a matrix and ask for the rank of this matrix.

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 4 & 3 & 1 \\ 0 & 5 & 1 & 1 \end{pmatrix}$$

   The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 0 & \frac{8}{7} \\ 0 & 1 & 0 & \frac{2}{7} \\ 0 & 0 & 1 & -\frac{3}{7} \end{pmatrix}$$

   A basis for H is

$$\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 4 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix} \right\}$$

   and so H has dimension 3.

3. Here are three vectors. Determine whether they are linearly independent or linearly dependent.

$$\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}$$

You need to consider the solutions to the equation

$$c_1\begin{pmatrix}1\\0\\1\end{pmatrix}+c_2\begin{pmatrix}2\\0\\1\end{pmatrix}+c_3\begin{pmatrix}3\\0\\0\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}$$

and determine whether there is a solution other than the obvious one, $c_1=c_2=c_3=0$. The augmented matrix for the system of equations is

$$\left(\begin{array}{ccc|c}1&2&3&0\\0&0&0&0\\1&1&0&0\end{array}\right)$$

Taking -1 times the top row and adding to the bottom and then switching the two bottom rows yields

$$\left(\begin{array}{ccc|c}1&2&3&0\\0&-1&-3&0\\0&0&0&0\end{array}\right)$$

Next take 2 times the second row and add to the top. This yields

$$\left(\begin{array}{ccc|c}1&0&-3&0\\0&-1&-3&0\\0&0&0&0\end{array}\right)$$

There are solutions other than the zero solution because c_3 is a free variable. Therefore, these vectors are not linearly independent.

4. Here are four vectors. Determine whether they span $\mathbb{R}^3$. Are these vectors linearly independent?

$$\begin{pmatrix}1\\2\\3\end{pmatrix},\begin{pmatrix}4\\0\\3\end{pmatrix},\begin{pmatrix}3\\2\\0\end{pmatrix},\begin{pmatrix}2\\1\\6\end{pmatrix}$$

The vectors can't possibly be linearly independent. If they were, they would constitute a linearly independent set consisting of four vectors even though there exists a spanning set of only three,

$$\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\1\\0\end{pmatrix},\begin{pmatrix}0\\0\\1\end{pmatrix}$$

However, the four given vectors might still span $\mathbb{R}^3$ even though they are not a basis. What does it take to span $\mathbb{R}^3$? Given a vector $(x,y,z)^T\in\mathbb{R}^3$, do there exist scalars c_1,c_2,c_3, and c_4 such that

$$c_1\begin{pmatrix}1\\2\\3\end{pmatrix}+c_2\begin{pmatrix}4\\0\\3\end{pmatrix}+c_3\begin{pmatrix}3\\2\\0\end{pmatrix}+c_4\begin{pmatrix}2\\1\\6\end{pmatrix}=\begin{pmatrix}x\\y\\z\end{pmatrix}?$$

Consider the augmented matrix of the above,

$$\left(\begin{array}{cccc|c}1&4&3&2&x\\2&0&2&1&y\\3&3&0&6&z\end{array}\right)$$

Doing row operations till an echelon form is obtained leads to

$$\left(\begin{array}{cccc|c} 1 & 0 & 0 & \frac{5}{4} & \frac{1}{4}y+\frac{2}{9}z-\frac{1}{6}x \\ 0 & 1 & 0 & \frac{3}{4} & -\frac{1}{4}y+\frac{1}{6}x+\frac{1}{9}z \\ 0 & 0 & 1 & -\frac{3}{4} & -\frac{2}{9}z+\frac{1}{6}x+\frac{1}{4}y \end{array}\right)$$

and you see there is a solution to the desired system of equations. In fact there are infinitely many because c_4 is a free variable. Therefore, the four vectors do span $\mathbb{R}^3$.

5. Consider the vectors of the form

$$\left\{\left(\begin{array}{c} 2t+6s \\ s-2t \\ 3t+s \end{array}\right) : s,t \in \mathbb{R}\right\}.$$

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.

This is indeed a subspace. You only need to verify the set of vectors is closed with respect to the vector space operations. Let $\left(\begin{array}{c} 2t_1+6s_1 \\ s_1-2t_1 \\ 3t_1+s_1 \end{array}\right)$ and $\left(\begin{array}{c} 2t+6s \\ s-2t \\ 3t+s \end{array}\right)$ be two vectors in the given set of vectors.

$$\begin{aligned} & \alpha\left(\begin{array}{c} 2t+6s \\ s-2t \\ 3t+s \end{array}\right)+\beta\left(\begin{array}{c} 2t_1+6s_1 \\ s_1-2t_1 \\ 3t_1+s_1 \end{array}\right) \\ = & \left(\begin{array}{c} 2\alpha t+6\alpha s+2\beta t_1+6\beta s_1 \\ \alpha s-2\alpha t+\beta s_1-2\beta t_1 \\ 3\alpha t+\alpha s+3\beta t_1+\beta s_1 \end{array}\right) \\ = & \left(\begin{array}{c} 2\left(\alpha t+\beta t_1\right)+6\left(\alpha s+\beta s_1\right) \\ \alpha s+\beta s_1-2\left(\alpha t+\beta t_1\right) \\ 3\left(\alpha t+\beta t_1\right)+\alpha s+\beta s_1 \end{array}\right) \end{aligned}$$

If we let $T \equiv \alpha t+\beta t_1$ and $S \equiv \alpha s+\beta s_1$, this is seen to be of the form

$$\left(\begin{array}{c} 2T+6S \\ S-2T \\ 3T+S \end{array}\right)$$

which is the way the vectors in the given set are described. Another way to see this is to notice that the vectors in the given set are of the form

$$t\left(\begin{array}{c} 2 \\ -2 \\ 3 \end{array}\right)+s\left(\begin{array}{c} 6 \\ 1 \\ 1 \end{array}\right)$$

so it consists of the span of the two vectors,

$$\left(\begin{array}{c} 2 \\ -2 \\ 3 \end{array}\right), \left(\begin{array}{c} 6 \\ 1 \\ 1 \end{array}\right). \tag{9.6}$$

Recall that the span of a set of vectors is always a subspace. You can also verify these vectors in 9.6 form a linearly independent set and so they are a basis.

6. Let $M = \left\{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 \geq u_2\right\}$. Is M a subspace? Explain.

   This is not a subspace because if $\mathbf{u} \in M$, is such that $u_3 > u_2$, then consider $(-1)\,\mathbf{u}$. If this were in M you would need to have $-u_3 > -u_2$ and so $u_3 < u_2$ which cannot be true if $u_3 > u_2$. Thus M is not closed under scalar multiplication so it is not a subspace.

7. Let $\mathbf{w}, \mathbf{w}_1$ be given vectors in $\mathbb{R}^2$ and define

$$M = \left\{\mathbf{u} = (u_1, u_2) \in \mathbb{R}^2 : \mathbf{w} \cdot \mathbf{u} = 0 \text{ and } \mathbf{w}_1 \cdot \mathbf{u} = 0\right\}.$$

   Is M a subspace? Explain.

   Suppose $\mathbf{u}'$ and $\mathbf{u}$ are both in M. What about $\alpha\mathbf{u}' + \beta\mathbf{u}$?

$$\mathbf{w}\cdot(\alpha\mathbf{u}' + \beta\mathbf{u}) = \alpha\mathbf{w}\cdot\mathbf{u}' + \beta\mathbf{w}\cdot\mathbf{u} = \alpha 0 + \beta 0 = 0$$

   Similarly,

$$\mathbf{w}_1\cdot(\alpha\mathbf{u}' + \beta\mathbf{u}) = \alpha\mathbf{w}_1\cdot\mathbf{u}' + \beta\mathbf{w}_1\cdot\mathbf{u} = \alpha 0 + \beta 0 = 0$$

   and so $\alpha\mathbf{u}' + \beta\mathbf{u} \in M$. This has verified that M is a subspace.

Linear Transformations 27 Sept.

Quiz

1. Find the rank of the matrix
$$\begin{pmatrix} 1 & 2 & 1 & 1 \\ 2 & 1 & 1 & 1 \\ 2 & 7 & 3 & 3 \\ 1 & 8 & 3 & 3 \end{pmatrix}$$

2. For A the above matrix, find
$$null\,(A) = \ker(A).$$
That is, find its null space.

3. For A the matrix of Problem 1 find a basis for the column space of this matrix.

4. For A the matrix of Problem 1 find a basis for the row space of this matrix.

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 10.0.27 *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 10.0.28 *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 7.14 on Page 129. 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 10.0.29 *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 10.0.30 *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 $\mathbf{x} \in \mathbb{F}^n$.

10.1 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

$$A\mathbf{x} = T\mathbf{x} \tag{10.1}$$

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 10.1. 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 10.1.1 *Let T be a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$. Then the matrix, A satisfying 10.1 is given by*

$$\begin{pmatrix} | & & | \\ T(\mathbf{e}_1) & \cdots & T(\mathbf{e}_n) \\ | & & | \end{pmatrix}$$

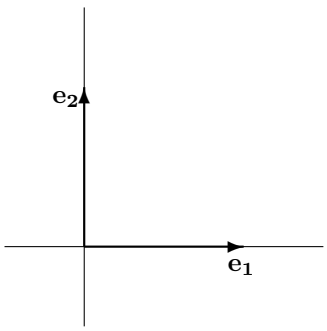
where $T\mathbf{e}_i$ is the i^{th} column of A.

10.1.1 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 10.1.2 *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 10.1.1,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Example 10.1.3 *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.

Example 10.1.4 *Find the matrix of the linear transformation which is obtained by first rotating all vectors through an angle of $\pi/6$ and then reflecting through the x axis.*

As shown in Example 10.1.3, the matrix of the transformation which involves rotating through an angle of $\pi/6$ is

$$\begin{pmatrix} \cos(\pi/6) & -\sin(\pi/6) \\ \sin(\pi/6) & \cos(\pi/6) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\sqrt{3} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}\sqrt{3} \end{pmatrix}$$

The matrix for the transformation which reflects all vectors through the x axis is

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

Therefore, the matrix of the linear transformation which first rotates through $\pi/6$ and then reflects through the x axis is

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{2}\sqrt{3} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}\sqrt{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\sqrt{3} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2}\sqrt{3} \end{pmatrix}.$$

10.1.2 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 10.1.5 *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

$$\frac{1}{14}\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{pmatrix}.$$

10.1.3 Matrices Which Are One To One Or Onto

Lemma 10.1.6 *Let A be an $m \times n$ matrix. Then $A(\mathbb{F}^n) = \text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_n)$ 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 7.1.9 on Page 124.

The following is a theorem of major significance. First here is an interesting observation.

Observation 10.1.7 *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 10.1.8 *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 $\mathbb{F}^n$ by Corollary 9.3.16. 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. This implies the span of the columns of A equals $\mathbb{F}^n$ and by Corollary 9.3.16 this implies the columns of A are independent. If $A\mathbf{x} = \mathbf{0}$, then letting $\mathbf{x} = (x_1, \cdots, x_n)^T$, it follows

$$\sum_{i=1}^{n} x_i \mathbf{a}_i = \mathbf{0}$$

and so each $x_i = 0$. If $A\mathbf{x} = A\mathbf{y}$, then $A(\mathbf{x} - \mathbf{y}) = \mathbf{0}$ and so $\mathbf{x} = \mathbf{y}$. This shows A is one to one.

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

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \; B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & -1 \end{pmatrix} \tag{10.2}$$

Then

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

but

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

10.1.4 The General Solution Of A Linear System

Recall the following definition which was discussed above.

Definition 10.1.9 *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 10.1.10 *Let T be a linear transformation. Define*

$$\ker(T) \equiv \{\mathbf{x} : T\mathbf{x} = \mathbf{0}\}.$$

In words, $\ker(T)$ *is called the* ***kernel*** *of T. As just described,* $\ker(T)$ *consists of the set of all vectors which T sends to* $\mathbf{0}$*. This is also called the* ***null space*** *of T. It is also called the* ***solution space*** *of the equation $T\mathbf{x} = \mathbf{0}$.*

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 10.1.11 *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 10.1.12 *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 10.1.13 *Let*

$$A = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 2 & 1 & 1 & 2 \\ 4 & 5 & 7 & 2 \end{pmatrix}$$

Find $\ker(A)$*. Equivalently, find the solution space to the system of equations $A\mathbf{x} = \mathbf{0}$.*

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

$$\begin{array}{c} x+2y+3z=0 \\ 2x+y+z+2w=0 \\ 4x+5y+7z+2w=0 \end{array}$$

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 the row 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,

$$\left(\begin{array}{c} \frac{1}{3}z - \frac{4}{3}w \\ \frac{2}{3}w - \frac{5}{3}z \\ z \\ w \end{array}\right) = z\left(\begin{array}{c} \frac{1}{3} \\ -\frac{5}{3} \\ 1 \\ 0 \end{array}\right) + w\left(\begin{array}{c} -\frac{4}{3} \\ \frac{2}{3} \\ 0 \\ 1 \end{array}\right).$$

Example 10.1.14 *The **general solution** of a linear system of equations is just the set of all solutions. Find the general solution to the linear system,*

$$\left(\begin{array}{cccc} 1 & 2 & 3 & 0 \\ 2 & 1 & 1 & 2 \\ 4 & 5 & 7 & 2 \end{array}\right)\left(\begin{array}{c} x \\ y \\ z \\ w \end{array}\right) = \left(\begin{array}{c} 9 \\ 7 \\ 25 \end{array}\right)$$

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 10.1.13. Therefore, from Theorem 10.1.12, you will obtain all solutions to the above linear system in the form

$$z\left(\begin{array}{c} \frac{1}{3} \\ -\frac{5}{3} \\ 1 \\ 0 \end{array}\right) + w\left(\begin{array}{c} -\frac{4}{3} \\ \frac{2}{3} \\ 0 \\ 1 \end{array}\right) + \left(\begin{array}{c} 1 \\ 1 \\ 2 \\ 1 \end{array}\right)$$

because $\left(\begin{array}{c} x \\ y \\ z \\ w \end{array}\right) = \left(\begin{array}{c} 1 \\ 1 \\ 2 \\ 1 \end{array}\right)$ is a particular solution to the given system of equations.

10.1.5 Exercises With Answers

1. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $5\pi/12$.

   You note that $5\pi/12 = 2\pi/3 - \pi/4$. Therefore, you can first rotate through $-\pi/4$ and then rotate through $2\pi/3$ to get the rotation through $5\pi/12$. The matrix of the transformation with respect to the usual coordinates which rotates through $-\pi/4$ is

   $$\begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix}$$

   and the matrix of the transformation which rotates through $2\pi/3$ is

   $$\begin{pmatrix} -1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix}.$$

   Multiplying these gives

   $$\begin{pmatrix} -1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix} \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix}$$

   $$= \begin{pmatrix} -\frac{1}{4}\sqrt{2} + \frac{1}{4}\sqrt{3}\sqrt{2} & -\frac{1}{4}\sqrt{2} - \frac{1}{4}\sqrt{3}\sqrt{2} \\ \frac{1}{4}\sqrt{3}\sqrt{2} + \frac{1}{4}\sqrt{2} & -\frac{1}{4}\sqrt{2} + \frac{1}{4}\sqrt{3}\sqrt{2} \end{pmatrix}$$

   and this is the matrix of the desired transformation. Note this shows that

   $$\cos(5\pi/12) = -\frac{1}{4}\sqrt{2} + \frac{1}{4}\sqrt{3}\sqrt{2} \approx .258\,819\,05$$

   $$\sin(5\pi/12) = \frac{1}{4}\sqrt{3}\sqrt{2} + \frac{1}{4}\sqrt{2} \approx .965\,925\,83.$$

2. 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.

   What does it do to $\mathbf{e}_1$? First you rotate $\mathbf{e}_1$ through the given angle to obtain

   $$\begin{pmatrix} -1/2 \\ \sqrt{3}/2 \end{pmatrix}$$

   and then this becomes

   $$\begin{pmatrix} -1/2 \\ -\sqrt{3}/2 \end{pmatrix}.$$

   This is the first column of the desired matrix. Next $\mathbf{e}_2$ first is rotated through the given angle to give

   $$\begin{pmatrix} -\sqrt{3}/2 \\ -1/2 \end{pmatrix}$$

   and then it is reflected across the x axis to give

   $$\begin{pmatrix} -\sqrt{3}/2 \\ 1/2 \end{pmatrix}$$

   and this gives the second column of the desired matrix. Thus the matrix is

   $$\begin{pmatrix} -1/2 & -\sqrt{3}/2 \\ -\sqrt{3}/2 & 1/2 \end{pmatrix}.$$

3. Find the matrix for $\mathrm{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1,-2,3)^T$.

   Recall
   $$\mathrm{proj}_{\mathbf{u}}(\mathbf{v}) = \frac{\mathbf{v}\cdot\mathbf{u}}{||\mathbf{u}||^2}\mathbf{u}$$
   Therefore,
   $$\begin{aligned}\mathrm{proj}_{\mathbf{u}}(\mathbf{e}_1) &= \frac{1}{14}\begin{pmatrix}1\\-2\\3\end{pmatrix},\ \mathrm{proj}_{\mathbf{u}}(\mathbf{e}_2) = \frac{-2}{14}\begin{pmatrix}1\\-2\\3\end{pmatrix},\\ \mathrm{proj}_{\mathbf{u}}(\mathbf{e}_2) &= \frac{3}{14}\begin{pmatrix}1\\-2\\3\end{pmatrix}.\end{aligned}$$
   Hence the desired matrix is
   $$\frac{1}{14}\begin{pmatrix}1&-2&3\\-2&4&-6\\3&-6&9\end{pmatrix}.$$

4. Show that the function $T_{\mathbf{u}}$ defined by $T_{\mathbf{u}}(\mathbf{v}) \equiv \mathbf{v} - \mathrm{proj}_{\mathbf{u}}(\mathbf{v})$ is also a linear transformation.

   $$T_{\mathbf{u}}(\alpha\mathbf{v}+\beta\mathbf{w}) = \alpha\mathbf{v}+\beta\mathbf{w} - \mathrm{proj}_{\mathbf{u}}(\alpha\mathbf{v}+\beta\mathbf{w})$$
   which from 3 equals
   $$\alpha(\mathbf{v}-\mathrm{proj}_{\mathbf{u}}(\mathbf{v})) + \beta(\mathbf{w}-\mathrm{proj}_{\mathbf{u}}(\mathbf{w})) = \alpha T_{\mathbf{u}}\mathbf{v} + \beta T_{\mathbf{u}}\mathbf{w}.$$
   This is what it takes to be a linear transformation.

5. If A, B, and C are each $n\times n$ matrices and ABC is invertible, why are each of A, B, and C invertible.

   $0 \neq \det(ABC) = \det(A)\det(B)\det(C)$ and so none of $\det(A), \det(B)$, or $\det(C)$ can equal zero. Therefore, each is invertible. You should do this another way, showing that each of A, B, and C is one to one and then using a theorem presented earlier.

6. Give an example of a 3×1 matrix with the property that the linear transformation determined by this matrix is one to one but not onto.

   Here is one. $\begin{pmatrix}1\\0\\0\end{pmatrix}$. If $\begin{pmatrix}1\\0\\0\end{pmatrix}x = \begin{pmatrix}0\\0\\0\end{pmatrix}$, then $x = 0$ but this is certainly not onto as a map from $\mathbb{R}^1$ to $\mathbb{R}^3$ because it does not ever yield $\begin{pmatrix}1\\1\\0\end{pmatrix}$.

7. Find the matrix of the linear transformation from $\mathbb{R}^3$ to $\mathbb{R}^3$ which first rotates every vector through an angle of $\pi/4$ about the z axis when viewed from the positive z axis and then rotates every vector through an angle of $\pi/6$ about the x axis when viewed from the positive x axis.

   The matrix of the linear transformation which accomplishes the first rotation is
   $$\begin{pmatrix}\sqrt{2}/2 & -\sqrt{2}/2 & 0\\ \sqrt{2}/2 & \sqrt{2}/2 & 0\\ 0 & 0 & 1\end{pmatrix}$$

and the matrix which accomplishes the second rotation is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{3}/2 & -1/2 \\ 0 & 1/2 & \sqrt{3}/2 \end{pmatrix}$$

Therefore, the matrix of the desired linear transformation is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{3}/2 & -1/2 \\ 0 & 1/2 & \sqrt{3}/2 \end{pmatrix} \begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 & 0 \\ \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} & 0 \\ \frac{1}{4}\sqrt{3}\sqrt{2} & \frac{1}{4}\sqrt{3}\sqrt{2} & -\frac{1}{2} \\ \frac{1}{4}\sqrt{2} & \frac{1}{4}\sqrt{2} & \frac{1}{2}\sqrt{3} \end{pmatrix}$$

This might not be the first thing you would think of.

Part V

Eigenvalues, Eigenvectors, Determinants, Diagonalization

Outcomes

A. Interpret the eigenvalue problem algebraically.

   i. Determine whether a given vector is an eigenvector.
   ii. Verify that a given value is an eigenvalue.

B. Interpret the eigenvalue problem geometrically. Determine eigenvalues and eigenvectors based on:

   i. an understanding of the linear transformation determined by the matrix
   ii. from the graph of the eigenspace.

C. Find the eigenvalues and eigenvectors of a general 2×2 matrix.

Reading: Linear Algebra 4.1

Homework: 4.1:

Outcome Mapping:

A. 1-6,7-12

B. 13-18,19-22

C. 23-26,27-30,31-34,35-38

A. Apply the Laplace Expansion to evaluate determinants of $n \times n$ matrices.

B. Recall and apply the properties of determinants to evaluate determinants, including:

   i. $\det(AB) = \det(A)\det(B)$
   ii. $\det(kA) = k^n \det(A)$
   iii. $\det(A^{-1}) = \dfrac{1}{\det(A)}$
   iv. $\det(A^T) = \det(A)$

C. Recall the effects that row operations have on the determinants of matrices. Relate to the determinants of elementary matrices.

D. Prove theorems involving determinants.

E. Evaluate matrix inverses using the adjoint method. Determine whether or not a matrix has an inverse based on its determinant.

F. Use Cramer's rule to solve a linear system.

Reading: Linear Algebra 4.2

Homework: 4.2:

Outcome Mapping:

A. 1-6,7-15,16-20

B. 35-38,47-52

C. 26-33,35-40

D. 21,41-44,53-56,66

E. 45-46,61-64,65

F. 57-60

A. Given an $n \times n$ matrix, compute

   i. the characteristic polynomial
   ii. the eigenvalues
   iii. a basis for each eigenspace
   iv. the algebraic and geometric multiplicities of each eigenvalue

B. Solve application problems involving eigenvalues and eigenvectors.

C. Recall and prove theorems involving eigenvalues and eigenvectors.

Reading: Linear Algebra 4.3

Homework: 4.3:

Outcome Mapping:

A. 1-12

B. 15-22,26-31,33-38

C. 23-25,32,39-42

A. Define similarity. Determine whether or not two matrices are similar.

B. Determine if a matrix is diagonalizable. Find the diagonalization of a matrix.

C. Find powers of a matrix using the diagonalization of a matrix.

D. Prove theorems involving the similarity and diagonalization of matrices.

Reading: Linear Algebra 4.4

Homework: 4.4:

Outcome Mapping:

A. 1-4,36-39

B. 5-7,8-15,24-29

C. 16-23

D. 30-35,40-50

Determinants 2,3 Oct.

Quiz

1. A linear transformation involves first rotating the vectors in $\mathbb{R}^2$ counterclockwise through an angle of 30 degrees and then reflecting across the x axis. Find the matrix of this linear transformation.

2. A linear transformation involves projecting all vectors on to the span of the vector $(1,1,1)$. Find the matrix of this linear transformation.

11.1 Basic Techniques And Properties

11.1.1 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 11.1.1 *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 11.1.2 *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 11.1.3 *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 11.1.4 *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 11.1.5 *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* $\operatorname{cof}(A)_{ij}$. *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* $\operatorname{cof}(A)_{ij}$. *Thus the cofactor of the* ij^{th} *entry is just the* ij^{th} *cofactor.*

Example 11.1.6 *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

$$\operatorname{cof}(A)_{12} = (-1)^{1+2}\det\begin{pmatrix} 4 & 2 \\ 3 & 1 \end{pmatrix} = (-1)^{1+2}(-2) = 2$$

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,

$$\operatorname{cof}(A)_{23} = (-1)^{2+3}\det\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = (-1)^{2+3}(-4) = 4.$$

Similarly,

$$\operatorname{cof}(A)_{22} = (-1)^{2+2}\det\begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix} = -8.$$

Definition 11.1.7 *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 11.1.8 *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 11.1.9

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 11.1.10

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 11.1.11 *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. $\det(A) =$

$$3(-1)^{1+3}\begin{vmatrix} 5 & 4 & 3 \\ 1 & 3 & 5 \\ 3 & 4 & 2 \end{vmatrix} + 2(-1)^{2+3}\begin{vmatrix} 1 & 2 & 4 \\ 1 & 3 & 5 \\ 3 & 4 & 2 \end{vmatrix} +$$

$$4(-1)^{3+3}\begin{vmatrix} 1 & 2 & 4 \\ 5 & 4 & 3 \\ 3 & 4 & 2 \end{vmatrix} + 3(-1)^{4+3}\begin{vmatrix} 1 & 2 & 4 \\ 5 & 4 & 3 \\ 1 & 3 & 5 \end{vmatrix}.$$

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 11.1.12 *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****, $\operatorname{cof}(A)$ 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 $(-1)^{i+j}$. 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, $\operatorname{cof}(A)_{ij}$ 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 11.1.13 *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*

$$\det(A)=\sum_{j=1}^{n}a_{ij}\operatorname{cof}(A)_{ij}=\sum_{i=1}^{n}a_{ij}\operatorname{cof}(A)_{ij}. \tag{11.1}$$

The first formula consists of expanding the determinant along the i^{th} row and the second expands the determinant along the j^{th} column.

Theorem 11.1.14 *Expanding the $n\times n$ matrix along any row or column always gives the same answer so the above definition is a good definition.*

11.1.2 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 11.1.15 *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 11.1.16 *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 11.1.17 *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.

11.1.3 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 11.1.18 *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 11.1.19 *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 11.1.20 *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 11.1.21 *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 11.1.22 *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 11.1.23 *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 11.1.24 *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 11.1.25 *In Theorems 11.1.19 - 11.1.23 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 11.1.26 *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 11.1.27 *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 $\det(A)\det(B) = 8 \times (-5) = -40$.

11.1.4 Finding Determinants Using Row Operations

Theorems 11.1.23 - 11.1.25 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 11.1.28 *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 11.1.23, 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 11.1.23 this didn't change the value of the determinant. Then the last row was multiplied by (-3). By Theorem 11.1.19 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 11.1.23. 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 11.1.29 *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 11.1.23 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 11.1.25 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 11.1.23 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.

11.1.5 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. Also recall the definition of the cofactor matrix given in Definition 11.1.12 on Page 200. 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 11.1.30 *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 11.1.31 *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 11.1.23 - 11.1.25 on Page 202, 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 11.1.32 *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 11.1.30: 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 11.1.21. 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,

$$\frac{\operatorname{cof}(A)^T}{\det(A)} A = I. \tag{11.2}$$

Using the other formula in Definition 11.1.13, 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,

$$A \frac{\operatorname{cof}(A)^T}{\det(A)} = I, \tag{11.3}$$

and it follows from 11.2 and 11.3 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 11.1.26,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A) \det\left(A^{-1}\right)$$

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 11.1.33 *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

$$\frac{1}{e^t} \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{pmatrix}^T = \begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}.$$

Eigenvalues And Eigenvectors Of A Matrix 4-6 Oct.

Quiz

1. Here is a matrix.
$$\begin{pmatrix} 1 & 0 & 2 \\ 2 & 1 & 3 \\ 3 & 2 & 1 \end{pmatrix}$$
Find its determinant.

2. Use the theory of determinants to find the inverse of the matrix,
$$\begin{pmatrix} 2 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

3. Let $C = F^T F$ where F is an $n \times n$ real matrix. Show $\det(C) \geq 0$.

4. Show that if A^{-1} exists, then $\det\left(A^{-1}\right) = 1/\det(A)$.

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.

12.0.6 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 12.0.34 *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 12.0.35 *Let M be an $n \times n$ matrix and let $\mathbf{x} \in \mathbb{C}^n$ be a* ***nonzero vector*** *for which*

$$M\mathbf{x} = \lambda\mathbf{x} \tag{12.1}$$

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 scalar multiple of $\mathbf{x}$. 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 and eigenvalues be identified?

There is an important characterization of when a matrix is invertible in terms of determinants. This is proved completely in the section on the theory of determinants where a formula is given for the inverse in terms of the determinant and cofactors.

Theorem 12.0.36 *Let M be an $n \times n$ matrix and let T_M denote the linear transformation determined by M. Thus $T_M\mathbf{x} = M\mathbf{x}$. Then the following are equivalent.*

1. *T_M is one to one.*

2. T_M *is onto.*

3. $\det(M) \neq 0$.

Suppose $\mathbf{x}$ satisfies 12.1. 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,

$$\mathbf{x} = \left(\left(M - \lambda I\right)^{-1}\left(M - \lambda I\right)\right)\mathbf{x} = \left(M - \lambda I\right)^{-1}\left(\left(M - \lambda I\right)\mathbf{x}\right) = \left(M - \lambda I\right)^{-1}\mathbf{0} = \mathbf{0},$$

and this would require $\mathbf{x} = \mathbf{0}$, contrary to the requirement that $\mathbf{x} \neq \mathbf{0}$. By Theorem 12.0.36,

$$\det(M - \lambda I) = 0. \tag{12.2}$$

(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, 12.2 has a solution, $\lambda \in \mathbb{C}$ by the fundamental theorem of algebra. Is it actually an eigenvalue? The answer is yes by Theorem 12.0.36. Since $\lambda I - M$ has no inverse due to its determinant equaling zero, it must fail to be one to one and so there must exist a nonzero vector which it maps to zero. This proves the following corollary.

Corollary 12.0.37 *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 $(M - \lambda I)\,\mathbf{x} = \mathbf{0}$.*

12.0.7 Finding Eigenvectors And Eigenvalues

As an example, consider the following.

Example 12.0.38 *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,

$$\left(\left(\begin{array}{ccc} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{array} \right) - 5 \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right) \right) \left(\begin{array}{c} x \\ y \\ z \end{array} \right) = \left(\begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right).$$

That is you need to find the solution to

$$\left(\begin{array}{ccc} 0 & -10 & -5 \\ 2 & 9 & 2 \\ -4 & -8 & 1 \end{array} \right) \left(\begin{array}{c} x \\ y \\ z \end{array} \right) = \left(\begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right)$$

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

$$\left(\begin{array}{ccc|c} 0 & -10 & -5 & 0 \\ 2 & 9 & 2 & 0 \\ -4 & -8 & 1 & 0 \end{array} \right). \tag{12.3}$$

The row 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

$$\left(\begin{array}{c} \frac{5}{4}t \\ \frac{-1}{2}t \\ t \end{array} \right) = t \left(\begin{array}{c} \frac{5}{4} \\ \frac{-1}{2} \\ 1 \end{array} \right)$$

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 12.3 is

$$t \left(\begin{array}{c} 5 \\ -2 \\ 4 \end{array} \right) \tag{12.4}$$

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 12.4 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.

$$\left(\begin{array}{ccc} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{array} \right) \left(\begin{array}{c} 5 \\ -2 \\ 4 \end{array} \right) = \left(\begin{array}{c} 25 \\ -10 \\ 20 \end{array} \right) = 5 \left(\begin{array}{c} 5 \\ -2 \\ 4 \end{array} \right)$$

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 12.4 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,

$$\left(\begin{array}{ccc|c} -5 & -10 & -5 & 0 \\ 2 & 4 & 2 & 0 \\ -4 & -8 & -4 & 0 \end{array}\right)$$

The row 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.

12.0.8 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 12.0.39 *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 row 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 row 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 the row 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 $t \neq 0$.

12.0.9 Defective And Nondefective Matrices

Definition 12.0.40 *By the fundamental theorem of algebra, it is possible to write the characteristic equation in the form*

$$(\lambda - \lambda_1)^{r_1} (\lambda - \lambda_2)^{r_2} \cdots (\lambda - \lambda_m)^{r_m} = 0$$

where r_i is some integer no smaller than 1. Thus the eigenvalues are $\lambda_1, \lambda_2, \cdots, \lambda_m$. The ***algebraic multiplicity*** *of λ_j is defined to be r_j.*

Example 12.0.41 *Consider the matrix,*

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \tag{12.5}$$

What is the algebraic multiplicity of the eigenvalue $\lambda = 1$?

In this case the characteristic equation is

$$\det(A - \lambda I) = (1 - \lambda)^3 = 0$$

or equivalently,

$$\det(\lambda I - A) = (\lambda - 1)^3 = 0.$$

Therefore, λ is of algebraic multiplicity 3.

Definition 12.0.42 *The **geometric multiplicity** of an eigenvalue is the dimension of the eigenspace,*

$$\ker(A - \lambda I).$$

Example 12.0.43 *Find the geometric multiplicity of $\lambda = 1$ for the matrix in 12.5.*

We need to solve

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The augmented matrix which must be row reduced to get this solution is therefore,

$$\left(\begin{array}{ccc|c} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

This requires $z = y = 0$ and x is arbitrary. Thus the eigenspace is

$$t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \ t \in \mathbb{F}.$$

It follows the geometric multiplicity of $\lambda = 1$ is 1.

Definition 12.0.44 *An $n \times n$ matrix is called **defective** if the geometric multiplicity is not equal to the algebraic multiplicity for some eigenvalue. Sometimes such an eigenvalue for which the geometric multiplicity is not equal to the algebraic multiplicity is called a defective eigenvalue. If the geometric multiplicity for an eigenvalue equals the algebraic multiplicity, the eigenvalue is sometimes referred to as nondefective.*

Here is another more interesting example of a defective matrix.

Example 12.0.45 *Let*

$$A = \begin{pmatrix} 2 & -2 & -1 \\ -2 & -1 & -2 \\ 14 & 25 & 14 \end{pmatrix}.$$

Find the eigenvectors and eigenvalues.

In this case the eigenvalues are $3, 6, 6$ where we have listed 6 twice because it is a zero of algebraic multiplicity two, the characteristic equation being

$$(\lambda - 3)(\lambda - 6)^2 = 0.$$

It remains to find the eigenvectors for these eigenvalues. First consider the eigenvectors for $\lambda = 3$. You must solve

$$\left(\left(\begin{array}{ccc} 2 & -2 & -1 \\ -2 & -1 & -2 \\ 14 & 25 & 14 \end{array} \right) - 3 \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right) \right) \left(\begin{array}{c} x \\ y \\ z \end{array} \right) = \left(\begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right).$$

The augmented matrix is

$$\left(\begin{array}{ccc|c} -1 & -2 & -1 & 0 \\ -2 & -4 & -2 & 0 \\ 14 & 25 & 11 & 0 \end{array} \right)$$

and the row reduced echelon form is

$$\left(\begin{array}{cccc} 1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right)$$

so the eigenvectors are nonzero vectors of the form

$$\left(\begin{array}{c} t \\ -t \\ t \end{array} \right) = t \left(\begin{array}{c} 1 \\ -1 \\ 1 \end{array} \right)$$

Next consider the eigenvectors for $\lambda = 6$. This requires you to solve

$$\left(\left(\begin{array}{ccc} 2 & -2 & -1 \\ -2 & -1 & -2 \\ 14 & 25 & 14 \end{array} \right) - 6 \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right) \right) \left(\begin{array}{c} x \\ y \\ z \end{array} \right) = \left(\begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right)$$

and the augmented matrix for this system of equations is

$$\left(\begin{array}{ccc|c} -4 & -2 & -1 & 0 \\ -2 & -7 & -2 & 0 \\ 14 & 25 & 8 & 0 \end{array} \right)$$

The row reduced echelon form is

$$\left(\begin{array}{cccc} 1 & 0 & \frac{1}{8} & 0 \\ 0 & 1 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 \end{array} \right)$$

and so the eigenvectors for $\lambda = 6$ are of the form

$$t \left(\begin{array}{c} -\frac{1}{8} \\ -\frac{1}{4} \\ 1 \end{array} \right)$$

or written more simply,

$$t \left(\begin{array}{c} -1 \\ -2 \\ 8 \end{array} \right)$$

where $t \in \mathbb{F}$.

Note that in this example the eigenspace for the eigenvalue, $\lambda = 6$ is of dimension 1 because there is only one parameter. However, this eigenvalue is of multiplicity two as a root to the characteristic equation. Thus this eigenvalue is a defective eigenvalue. However, the eigenvalue 3 is nondefective. The matrix is defective because it has a defective eigenvalue.

The word, defective, seems to suggest there is something wrong with the matrix. This is in fact the case. Defective matrices are a lot of trouble in applications and we may wish they never occurred. However, they do occur as the above example shows. When you study linear systems of differential equations, you will have to deal with the case of defective matrices and you will see how awful they are. The reason these matrices are so horrible to work with is that it is impossible to obtain a basis of eigenvectors. When you study differential equations, solutions to first order systems are expressed in terms of eigenvectors of a certain matrix times $e^{\lambda t}$ where λ is an eigenvalue. In order to obtain a general solution of this sort, you must have a basis of eigenvectors. For a defective matrix, such a basis does not exist and so you have to go to something called generalized eigenvectors. Unfortunately, it is **never** explained in beginning differential equations courses why there are enough generalized eigenvectors and eigenvectors to represent the general solution. In fact, this reduces to a difficult question in linear algebra equivalent to the existence of something called the Jordan Canonical form which is much more difficult than everything discussed in the entire differential equations course. If you become interested in this, see a good book in linear algebra. The good ones do discuss this topic. There is such a linear algebra book on my web page.

Ultimately, the algebraic issues which will occur in differential equations are a red herring anyway. The real issues relative to existence of solutions to systems of ordinary differential equations are analytical, having much more to do with calculus than with linear algebra although this will likely not be made clear when you take a beginning differential equations class.

In terms of algebra, this lack of a basis of eigenvectors says that it is impossible to obtain a diagonal matrix which is similar to the given matrix.

Although there may be repeated roots to the characteristic equation, 12.2 and it is not known whether the matrix is defective in this case, there is an important theorem which holds when considering eigenvectors which correspond to distinct eigenvalues.

Theorem 12.0.46 *Suppose* $M\mathbf{v}_i = \lambda_i \mathbf{v}_i, i = 1, \cdots, r$, $\mathbf{v}_i \neq 0$, *and that if* $i \neq j$, *then* $\lambda_i \neq \lambda_j$. *Then the set of eigenvectors,* $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ *is linearly independent.*

Proof: If the conclusion of this theorem is not true, then there exist non zero scalars, c_{k_j} such that

$$\sum_{j=1}^{m} c_{k_j} \mathbf{v}_{k_j} = \mathbf{0}. \tag{12.6}$$

Take m to be the smallest number possible for an expression of the form 12.6 to hold. Then solving for $\mathbf{v}_{k_1}$

$$\mathbf{v}_{k_1} = \sum_{k_j \neq k_1} d_{k_j} \mathbf{v}_{k_j} \tag{12.7}$$

where $d_{k_j} = c_{k_j}/c_{k_1} \neq 0$. Multiplying both sides by M,

$$\lambda_{k_1} \mathbf{v}_{k_1} = \sum_{k_j \neq k_1} d_{k_j} \lambda_{k_j} \mathbf{v}_{k_j},$$

which from 12.7 yields

$$\sum_{k_j \neq k_1} d_{k_j} \lambda_{k_1} \mathbf{v}_{k_j} = \sum_{k_j \neq k_1} d_{k_j} \lambda_{k_j} \mathbf{v}_{k_j}$$

and therefore,

$$\mathbf{0} = \sum_{k_j \neq k_1} d_{k_j} \left(\lambda_{k_1} - \lambda_{k_j} \right) \mathbf{v}_{k_j},$$

a sum having fewer than m terms. However, from the assumption that m is as small as possible for 12.6 to hold with all the scalars, c_{k_j} non zero, it follows that for some $j \neq 1$,

$$d_{k_j} \left(\lambda_{k_1} - \lambda_{k_j} \right) = 0$$

which implies $\lambda_{k_1} = \lambda_{k_j}$, a contradiction.

12.0.10 Diagonalization

Definition 12.0.47 *Let A be an $n \times n$ matrix. Then A is **diagonalizable** if there exists an invertible matrix, S such that*

$$S^{-1}AS = D$$

where D is a diagonal matrix. This means D has a zero as every entry except for the main diagonal.

Theorem 12.0.48 *An $n \times n$ matrix is diagonalizable if and only if $\mathbb{F}^n$ has a basis of eigenvectors of A. Furthermore, you can take the matrix, S described above to be given as*

$$S = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{pmatrix}$$

where here the v_k are the eigenvectors in the basis for $\mathbb{F}^n$. If A is diagonalizable, the eigenvalues of A are the diagonal entries of the diagonal matrix.

Proof: Suppose there exists a basis of eigenvectors, $\{\mathbf{v}_k\}$ where $A\mathbf{v}_k = \lambda_k \mathbf{v}_k$. Then let S be given as above. It follows S^{-1} exists and is of the form

$$S^{-1} = \begin{pmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_n^T \end{pmatrix}$$

where $\mathbf{w}_k^T \mathbf{v}_j = \delta_{kj}$. Then

$$\begin{aligned} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} &= \begin{pmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_n^T \end{pmatrix} \begin{pmatrix} \lambda_1 \mathbf{v}_1 & \lambda_2 \mathbf{v}_2 & \cdots & \lambda_n \mathbf{v}_n \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_n^T \end{pmatrix} \begin{pmatrix} A\mathbf{v}_1 & A\mathbf{v}_2 & \cdots & A\mathbf{v}_n \end{pmatrix} \\ &= S^{-1}AS \end{aligned}$$

Next suppose A is diagonalizable so $S^{-1}AS = D$. Let $S = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{pmatrix}$ where the columns are the $\mathbf{v}_k$ and

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

Then

$$AS = SD = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

and so

$$\begin{pmatrix} A\mathbf{v}_1 & A\mathbf{v}_2 & \cdots & A\mathbf{v}_n \end{pmatrix} = \begin{pmatrix} \lambda_1\mathbf{v}_1 & \lambda_2\mathbf{v}_2 & \cdots & \lambda_n\mathbf{v}_n \end{pmatrix}$$

showing the $\mathbf{v}_i$ are eigenvectors of A and the λ_k are eigenvectors. Now the $\mathbf{v}_k$ form a basis for $\mathbb{F}^n$ because the matrix, S having these vectors as columns is given to be invertible. This proves the theorem.

Definition 12.0.49 *Let A, B be two diagonal matrices. Then A is said to be similar to B if there exists an invertible matrix, S such that $B = S^{-1}AS$.*

Example 12.0.50 *Let $A = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 4 & -1 \\ -2 & -4 & 4 \end{pmatrix}$. Find a matrix, S such that $S^{-1}AS = D$, a diagonal matrix.*

Solving $\det(\lambda I - A) = 0$ yields the eigenvalues are 2 and 6 with 2 an eigenvalue of multiplicity two. Solving $(2I - A)\mathbf{x} = \mathbf{0}$ to find the eigenvectors, you find that the eigenvectors are

$$a\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + b\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

where a, b are scalars. An eigenvector for $\lambda = 6$ is $\begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}$. Let the matrix S be

$$S = \begin{pmatrix} -2 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -2 \end{pmatrix}$$

That is, the columns are the eigenvectors. Then

$$S^{-1} = \begin{pmatrix} -\frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \end{pmatrix}.$$

$$\begin{aligned} S^{-1}AS &= \begin{pmatrix} -\frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 1 & 4 & -1 \\ -2 & -4 & 4 \end{pmatrix} \begin{pmatrix} -2 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -2 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 6 \end{pmatrix}. \end{aligned}$$

Example 12.0.51 *Here is a matrix. $A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix}$ Find A^{50}.*

Sometimes this sort of problem can be made easy by using diagonalization. In this case there are eigenvectors,

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix},$$

the first two corresponding to $\lambda = 1$ and the last corresponding to $\lambda = 2$. Then let the eigenvectors be the columns of the matrix, S. Thus

$$S = \begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

Then also

$$S^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix}$$

and

$$\begin{aligned} S^{-1}AS &= \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} = D \end{aligned}$$

Now it follows

$$A = SDS^{-1} = \begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix}.$$

Now note that $\left(SDS^{-1}\right)^2 = SDS^{-1}SDS^{-1} = SD^2S^{-1}$ and

$$\left(SDS^{-1}\right)^3 = SDS^{-1}SDS^{-1}SDS^{-1} = SD^3S^{-1},$$

etc. In general, you can see that

$$\left(SDS^{-1}\right)^n = SD^nS^{-1}$$

In other words, $A^n = SD^nS^{-1}$. Therefore,

$$\begin{aligned} A^{50} &= SD^{50}S^{-1} \\ &= \begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}^{50} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix}. \end{aligned}$$

Now

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}^{50} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2^{50} \end{pmatrix}.$$

It follows

$$\begin{aligned} A^{50} &= \begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2^{50} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 2^{50} & -1+2^{50} & 0 \\ 0 & 1 & 0 \\ 1-2^{50} & 1-2^{50} & 1 \end{pmatrix}. \end{aligned}$$

That isn't too hard. However, this would have been horrendous if you had tried to multiply A^{50} by hand.

This technique of diagonalization is also important in solving the differential equations resulting from vibrations. Sometimes you have systems of differential equation and when you diagonalize an appropriate matrix, you "decouple" the equations. This is very nice. It makes hard problems trivial.

The above example is entirely typical. If $A = SDS^{-1}$ then $A^m = SD^mS^{-1}$ and it is easy to compute D^m. More generally, you can define functions of the matrix using power series in this way. However, the real interesting case is when A is defective. This is much more interesting. You can always speak of things like $\sin(A)$ for A an $n \times n$ matrix. However, more interesting functions have no power series and you have to work harder for these. This is enough on this. One can go on and on.

12.0.11 Migration Matrices

There are applications of the eigenvalue problem which are of great importance and feature only one eigenvalue.

Consider the following table.

	A	B
A	1/4	2/3
B	3/4	1/3

In this table, 1/4 is the probability that someone in location A ends up in A after a single unit of time. 2/3 is the probability that a person in location B ends up in location A after a single unit of time. 3/4 is the probability that a person in location A ends up in location B after a single unit of time and 1/3 is the probability that a person in location B ends up in location B. Instead of the word probability, you could use the word "proportion" and the numbers would then represent the proportion of people in the various locations who end up in the other location after one unit of time. Thue 1/4 is the proportion of people in A who end up in A,etc. Then this matrix is called a stochastic matrix, a Markov matrix or a Migration matrix. In the case the numbers are interpreted as probabilities, it is called a Markov or Stochastic matrix. In the case where they are proportions it is called a migration matrix.

Consider it as a migration matrix and suppose that initially there are 200 people in location A and 120 in location B. You might wonder how many there would be in the two locations after one unit of time. This is easy to figure out. Those in A after one unit of time consist of those in A who were in A to begin with added to those in A who started off in B. Thus

$$\#\text{ in } A = \frac{1}{4}(200) + \frac{2}{3}(120) = 130$$

$$\#\text{ in } B = \frac{3}{4}(200) + \frac{1}{3}(120) = 190.$$

You can see that this amounts to nothing more than matrix multiplication. Thus, letting $(a_1, b_1)^T$ be defined by

$$\begin{pmatrix} a_1 \\ b_1 \end{pmatrix} = \begin{pmatrix} \#\text{ in } A \text{ after one unit of time} \\ \#\text{ in } B \text{ after one unit of time} \end{pmatrix}$$

It follows

$$\begin{pmatrix} a_1 \\ b_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 200 \\ 120 \end{pmatrix} = \begin{pmatrix} 130 \\ 190 \end{pmatrix}$$

Now with this vector as new input, you can determine how many are in the two locations after another unit of time using the same procedure. Thus letting a_n denote the numbers in location A after n units of time and b_n the number in B after n units of time,

$$\begin{aligned} \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} &= \begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} a_1 \\ b_1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 200 \\ 120 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix}^2 \begin{pmatrix} 200 \\ 120 \end{pmatrix} = \begin{pmatrix} 159.166\,667 \\ 160.833\,333 \end{pmatrix}. \end{aligned}$$

Obviously you need to round off if you are considering people doing the migrating. Then by analogy,

$$\begin{aligned} \begin{pmatrix} a_n \\ b_n \end{pmatrix} &= \begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} a_{n-1} \\ b_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix}^{n-1} \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix}^n \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} \end{aligned}$$

After 50 units of time you would have

$$\begin{aligned} \begin{pmatrix} a_{50} \\ b_{50} \end{pmatrix} &= \begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix}^{50} \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} \\ &= \begin{pmatrix} .470\,588\,235 & .470\,588\,235 \\ .529\,411\,765 & .529\,411\,765 \end{pmatrix} \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} \end{aligned}$$

After 100 units of time, you would have

$$\begin{aligned} \begin{pmatrix} a_{100} \\ b_{100} \end{pmatrix} &= \begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix}^{100} \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} \\ &= \begin{pmatrix} .470\,588\,235 & .470\,588\,235 \\ .529\,411\,765 & .529\,411\,765 \end{pmatrix} \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} \end{aligned}$$

You can't detect any difference between these two answers. In general, if you wanted to know about how many would be in the two locations, you would need to take a limit. However, there is a better way.

More generally here is a definition.

Definition 12.0.52 *Let n locations be denoted by the numbers $1, 2, \cdots, n$. Also suppose it is the case that each year a_{ij} denotes the proportion of residents in location j*

which move to location i. Also suppose no one escapes or emigrates from without these n locations. This last assumption requires $\sum_i a_{ij} = 1$. *Such matrices in which the columns are nonnegative numbers which sum to one are called* ***Markov matrices****. In this context describing migration, they are also called* ***migration matrices****.*

Example 12.0.53 *Here is an example of one of these matrices.*

$$\begin{pmatrix} .4 & .2 \\ .6 & .8 \end{pmatrix}$$

Thus if it is considered as a migration matrix, .4 *is the proportion of residents in location 1 which stay in location one in a given time period while* .6 *is the proportion of residents in location 1 which move to location 2 and* .2 *is the proportion of residents in location 2 which move to location 1. Considered as a Markov matrix, these numbers are usually identified with probabilities.*

If $\mathbf{v} = (x_1, \cdots, x_n)^T$ where x_i is the population of location i at a given instant, you obtain the population of location i one year later by computing $\sum_j a_{ij}x_j = (A\mathbf{v})_i$. Therefore, the population of location i after k years is $\left(A^k\mathbf{v}\right)_i$. An obvious application of this would be to a situation in which you rent trailers which can go to various parts of a city and you observe through experiments the proportion of trailers which go from point i to point j in a single day. Then you might want to find how many trailers would be in all the locations after 8 days.

Proposition 12.0.54 *Let* $A = (a_{ij})$ *be a migration matrix. Then* 1 *is always an eigenvalue for A.*

Proof: Remember that $\det\left(B^T\right) = \det(B)$. Therefore,

$$\det(A - \lambda I) = \det\left((A - \lambda I)^T\right) = \det\left(A^T - \lambda I\right)$$

because $I^T = I$. Thus the characteristic equation for A is the same as the characteristic equation for A^T and so A and A^T have the same eigenvalues. We will show that 1 is an eigenvalue for A^T and then it will follow that 1 is an eigenvalue for A.

Remember that for a migration matrix, $\sum_i a_{ij} = 1$. Therefore, if $A^T = (b_{ij})$ so $b_{ij} = a_{ji}$, it follows that

$$\sum_j b_{ij} = \sum_j a_{ji} = 1.$$

Therefore, from matrix multiplication,

$$A^T \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \sum_j b_{ij} \\ \vdots \\ \sum_j b_{ij} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

which shows that $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ is an eigenvector for A^T corresponding to the eigenvalue, $\lambda = 1$. As explained above, this shows that $\lambda = 1$ is an eigenvalue for A because A and A^T have the same eigenvalues.

Example 12.0.55 *Consider the migration matrix,* $\begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix}$ *for locations* $1, 2,$ *and* 3. *Suppose initially there are 100 residents in location 1, 200 in location 2 and 400 in location 4. Find the population in the three locations after 10 units of time.*

From the above, it suffices to consider

$$\begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix}^{10} \begin{pmatrix} 100 \\ 200 \\ 400 \end{pmatrix} = \begin{pmatrix} 115.085\,829\,22 \\ 120.130\,672\,44 \\ 464.783\,498\,34 \end{pmatrix}$$

Of course you would need to round these numbers off.

A related problem asks for how many there will be in the various locations after a long time. It turns out that if some power of the migration matrix has all positive entries, then there is a limiting vector, $\mathbf{x} = \lim_{k\to\infty} A^k \mathbf{x}_0$ where $\mathbf{x}_0$ is the initial vector describing the number of inhabitants in the various locations initially. This vector will be an eigenvector for the eigenvalue 1 because

$$\mathbf{x} = \lim_{k\to\infty} A^k \mathbf{x}_0 = \lim_{k\to\infty} A^{k+1} \mathbf{x}_0 = A \lim_{k\to\infty} A^k \mathbf{x} = A\mathbf{x},$$

and the sum of its entries will equal the sum of the entries of the initial vector, $\mathbf{x}_0$ because this sum is preserved for every multiplication by A since

$$\sum_i \sum_j a_{ij} x_j = \sum_j x_j \left(\sum_i a_{ij} \right) = \sum_j x_j.$$

Here is an example. It is the same example as the one above but here it will involve the long time limit.

Example 12.0.56 *Consider the migration matrix,* $\begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix}$ *for locations* $1, 2,$ *and* 3. *Suppose initially there are 100 residents in location 1, 200 in location 2 and 400 in location 4. Find the population in the three locations after a long time.*

You just need to find the eigenvector which goes with the eigenvalue 1 and then normalize it so the sum of its entries equals the sum of the entries of the initial vector. Thus you need to find a solution to

$$\left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix} \right) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

The augmented matrix is

$$\left(\begin{array}{ccc|c} .4 & 0 & -.1 & 0 \\ -.2 & .2 & 0 & 0 \\ -.2 & -.2 & .1 & 0 \end{array} \right)$$

and its row reduced echelon form is

$$\left(\begin{array}{cccc} 1 & 0 & -.25 & 0 \\ 0 & 1 & -.25 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right)$$

Therefore, the eigenvectors are

$$s\begin{pmatrix} (1/4) \\ (1/4) \\ 1 \end{pmatrix}$$

and all that remains is to choose the value of s such that

$$\frac{1}{4}s + \frac{1}{4}s + s = 100 + 200 + 400$$

This yields $s = \frac{1400}{3}$ and so the long time limit would equal

$$\frac{1400}{3}\begin{pmatrix} (1/4) \\ (1/4) \\ 1 \end{pmatrix} = \begin{pmatrix} 116.\,666\,666\,666\,666\,7 \\ 116.\,666\,666\,666\,666\,7 \\ 466.\,666\,666\,666\,666\,7 \end{pmatrix}.$$

You would of course need to round these numbers off. You see that you are not far off after just 10 units of time. Therefore, you might consider this as a useful procedure because it is probably easier to solve a simple system of equations than it is to raise a matrix to a large power.

Example 12.0.57 *Suppose a migration matrix is* $\begin{pmatrix} \frac{1}{5} & \frac{1}{2} & \frac{1}{5} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{11}{20} & \frac{1}{4} & \frac{3}{10} \end{pmatrix}$. *Find the comparison between the populations in the three locations after a long time.*

This amounts to nothing more than finding the eigenvector for $\lambda = 1$. Solve

$$\left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{5} & \frac{1}{2} & \frac{1}{5} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{11}{20} & \frac{1}{4} & \frac{3}{10} \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The augmented matrix is

$$\begin{pmatrix} \frac{4}{5} & -\frac{1}{2} & -\frac{1}{5} & | & 0 \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{2} & | & 0 \\ -\frac{11}{20} & -\frac{1}{4} & \frac{7}{10} & | & 0 \end{pmatrix}$$

The row echelon form is

$$\begin{pmatrix} 1 & 0 & -\frac{16}{19} & 0 \\ 0 & 1 & -\frac{18}{19} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and so an eigenvector is

$$\begin{pmatrix} 16 \\ 18 \\ 19 \end{pmatrix}.$$

Thus there will be $\frac{18}{16}^{th}$ more in location 2 than in location 1. There will be $\frac{19}{18}^{th}$ more in location 3 than in location 2.

You see the eigenvalue problem makes these sorts of determinations fairly simple.

There are many other things which can be said about these sorts of **migration problems**. They include things like the gambler's ruin problem which asks for the probability that a compulsive gambler will eventually lose all his money. However those problems are not so easy although they still involve eigenvalues and eigenvectors.

12.0.12 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 12.0.58 *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

$$\det\left(\left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)\right) = 0.$$

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.

12.0.13 The Estimation Of Eigenvalues

There are ways to estimate the eigenvalues for matrices. The most famous is known as Gerschgorin's theorem. This theorem gives a rough idea where the eigenvalues are just from looking at the matrix.

Theorem 12.0.59 *Let A be an $n \times n$ matrix. Consider the n Gerschgorin discs defined as*

$$D_i \equiv \left\{ \lambda \in \mathbb{C} : |\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}.$$

Then every eigenvalue is contained in some Gerschgorin disc.

This theorem says to add up the absolute values of the entries of the i^{th} row which are off the main diagonal and form the disc centered at a_{ii} having this radius. The union of these discs contains $\sigma(A)$.

Proof: Suppose $A\mathbf{x} = \lambda\mathbf{x}$ where $\mathbf{x} \neq \mathbf{0}$. Then for $A = (a_{ij})$

$$\sum_{j \neq i} a_{ij}x_j = (\lambda - a_{ii})\, x_i.$$

Therefore, picking k such that $|x_k| \geq |x_j|$ for all x_j, it follows that $|x_k| \neq 0$ since $|\mathbf{x}| \neq 0$ and

$$|x_k| \sum_{j \neq i} |a_{kj}| \geq \sum_{j \neq i} |a_{kj}|\, |x_j| \geq |\lambda - a_{ii}|\, |x_k|.$$

Now dividing by $|x_k|$, it follows λ is contained in the k^{th} Gerschgorin disc.

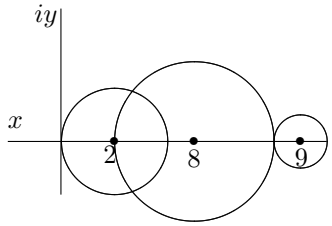
Example 12.0.60 *Here is a matrix. Estimate its eigenvalues.*

$$\begin{pmatrix} 2 & 1 & 1 \\ 3 & 5 & 0 \\ 0 & 1 & 9 \end{pmatrix}$$

According to Gerschgorin's theorem the eigenvalues are contained in the disks

$$\begin{aligned} D_1 &= \{\lambda \in \mathbb{C} : |\lambda - 2| \leq 2\}, \\ D_2 &= \{\lambda \in \mathbb{C} : |\lambda - 5| \leq 3\}, \\ D_3 &= \{\lambda \in \mathbb{C} : |\lambda - 9| \leq 1\} \end{aligned}$$

It is important to observe that these disks are in the complex plane. In general this is the case. If you want to find eigenvalues they will be complex numbers.



So what are the values of the eigenvalues? In this case they are real. You can compute them by graphing the characteristic polynomial, $\lambda^3 - 16\lambda^2 + 70\lambda - 66$ and then zooming in on the zeros. If you do this you find the solution is $\{\lambda = 1.\,295\,3\}, \{\lambda = 5.\,590\,5\}, \{\lambda = 9.\,114\,2\}$. Of course these are only approximations and so this information is useless for finding eigenvectors. However, in many applications, it is the size of the eigenvalues which is important and so these numerical values would be helpful for such applications. Because of this example, you might think there is no real reason for Gerschgorin's theorem. Why not just compute the characteristic equation and graph and zoom? This is fine up to a point, but what if the matrix was huge? Then it might be hard to find the characteristic polynomial. Remember the difficulties in expanding a big matrix along a row or column. You would need a better way to come up with the characteristic polynomial. Also, what if the eigenvalue were complex? You don't see these by following this procedure. However, Gerschgorin's theorem will at least estimate them.

There are also more advanced versions of this theorem which depend on the theory of functions of a complex variable covering the case where the Gerschgorin disks are disjoint. In this case, you can assert each disk contains an eigenvalue. In fact, if k of the Gerschgorin disks are disjoint from the other disks then they contain k eigenvalues. To see this proved, see the linear algebra book on my web page. Don't bother to look at it if you have not had a substantial course on complex analysis because it won't make any sense. Math is not like comparative literature, history, or humanities. You can't read the advanced topics until you have mastered the basic topics even if you are real smart.

12.1 The Mathematical Theory Of Determinants*

This material is definitely not for the faint of heart. It is only for people who want

to see everything proved. It is a fairly complete and unusually elementary treatment of the subject. 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 [21]. A recent book which also has a good introduction is Baker [4]. Most linear algebra books do not do an honest job presenting this topic.

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.

The following Lemma will be essential in the definition of the determinant.

Lemma 12.1.1 *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.*

$$\operatorname{sgn}_n(1, \cdots, n) = 1 \tag{12.8}$$

$$\operatorname{sgn}_n(i_1, \cdots, p, \cdots, q, \cdots, i_n) = -\operatorname{sgn}_n(i_1, \cdots, q, \cdots, p, \cdots, i_n) \tag{12.9}$$

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,* $(i_1, \cdots, i_n)$,

$$\operatorname{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) \equiv$$

$$(-1)^{n-\theta} \operatorname{sgn}_{n-1}(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n) \tag{12.10}$$

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 12.10 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 12.8 and 12.9 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 12.9 holds because both sides would equal zero from the above definition. It remains to verify 12.9 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 12.9 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.

$$\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{n+1}, \cdots, \overset{s}{q}, \cdots, i_{n+1}\right) =$$

$$(-1)^{n+1-r} \operatorname{sgn}_n\left(i_1, \cdots, \overset{s-1}{q}, \cdots, i_{n+1}\right) \tag{12.11}$$

while

$$\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{q}, \cdots, \overset{s}{n+1}, \cdots, i_{n+1}\right) =$$

$$(-1)^{n+1-s} \operatorname{sgn}_n\left(i_1, \cdots, \overset{r}{q}, \cdots, i_{n+1}\right). \tag{12.12}$$

By making $s-1-r$ switches, move the q which is in the $s-1^{th}$ position in 12.11 to the r^{th} position in 12.12. 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,

$$\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{n+1}, \cdots, \overset{s}{q}, \cdots, i_{n+1}\right) = (-1)^{n+1-r} \operatorname{sgn}_n\left(i_1, \cdots, \overset{s-1}{q}, \cdots, i_{n+1}\right)$$

$$= (-1)^{n+1-r}(-1)^{s-1-r} \operatorname{sgn}_n\left(i_1, \cdots, \overset{r}{q}, \cdots, i_{n+1}\right)$$

$$= (-1)^{n+s} \operatorname{sgn}_n\left(i_1, \cdots, \overset{r}{q}, \cdots, i_{n+1}\right) = (-1)^{2s-1}(-1)^{n+1-s} \operatorname{sgn}_n\left(i_1, \cdots, \overset{r}{q}, \cdots, i_{n+1}\right)$$

$$= -\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{q}, \cdots, \overset{s}{n+1}, \cdots, i_{n+1}\right).$$

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 12.8 and 12.9, 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 12.9 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 12.1.2 *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 12.1.3 *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

$$\det(A(r_1,\cdots,r_n)) = \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n} \tag{12.13}$$

and

$$A(1,\cdots,n) = A.$$

Proposition 12.1.4 *Let*

$$(r_1,\cdots,r_n)$$

be an ordered list of numbers from $\{1,\cdots,n\}$. Then

$$\operatorname{sgn}(r_1,\cdots,r_n)\det(A)$$

$$= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n} \tag{12.14}$$

$$= \det(A(r_1,\cdots,r_n)). \tag{12.15}$$

Proof: Let $(1,\cdots,n) = (1,\cdots,r,\cdots s,\cdots,n)$ so $r<s$.

$$\det(A(1,\cdots,r,\cdots,s,\cdots,n)) = \tag{12.16}$$

$$\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_r,\cdots,k_s,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_r}\cdots a_{sk_s}\cdots a_{nk_n},$$

and renaming the variables, calling k_s,k_r and k_r, k_s, this equals

$$= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_s,\cdots,k_r,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_s}\cdots a_{sk_r}\cdots a_{nk_n}$$

$$= \sum_{(k_1,\cdots,k_n)} -\operatorname{sgn}\left(k_1,\cdots,\overbrace{k_r,\cdots,k_s}^{\text{These got switched}},\cdots,k_n\right) a_{1k_1}\cdots a_{sk_r}\cdots a_{rk_s}\cdots a_{nk_n}$$

$$= -\det\left(A\left(1,\cdots,s,\cdots,r,\cdots,n\right)\right). \tag{12.17}$$

Consequently,

$$\det\left(A\left(1,\cdots,s,\cdots,r,\cdots,n\right)\right) =$$

$$-\det\left(A\left(1,\cdots,r,\cdots,s,\cdots,n\right)\right) = -\det\left(A\right)$$

Now letting $A\left(1,\cdots,s,\cdots,r,\cdots,n\right)$ play the role of A, and continuing in this way, switching pairs of numbers,

$$\det\left(A\left(r_1,\cdots,r_n\right)\right) = (-1)^p \det\left(A\right)$$

where it took p switches to obtain$(r_1,\cdots,r_n)$ from $(1,\cdots,n)$. By Lemma 12.1.1, this implies

$$\det\left(A\left(r_1,\cdots,r_n\right)\right) = (-1)^p \det\left(A\right) = \operatorname{sgn}\left(r_1,\cdots,r_n\right)\det\left(A\right)$$

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 12.16 -12.17 shows that $A\left(r_1,\cdots,r_n\right) = 0$ and also $\operatorname{sgn}\left(r_1,\cdots,r_n\right) = 0$ so the formula holds in this case also.

Observation 12.1.5 *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 $n\left(n-1\right)$ 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 $\det\left(A\right) = \det\left(A^T\right)$.

Corollary 12.1.6 *The following formula for $\det\left(A\right)$ is valid.*

$$\det\left(A\right) = \frac{1}{n!}\cdot$$

$$\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}\left(r_1,\cdots,r_n\right)\operatorname{sgn}\left(k_1,\cdots,k_n\right) a_{r_1k_1}\cdots a_{r_nk_n}. \tag{12.18}$$

And also $\det\left(A^T\right) = \det\left(A\right)$ 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 12.1.4, if the r_i are distinct,

$$\det\left(A\right) = \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}\left(r_1,\cdots,r_n\right)\operatorname{sgn}\left(k_1,\cdots,k_n\right) a_{r_1k_1}\cdots a_{r_nk_n}.$$

Summing over all ordered lists, $(r_1,\cdots,r_n)$ where the r_i are distinct, (If the r_i are not distinct, $\operatorname{sgn}\left(r_1,\cdots,r_n\right) = 0$ and so there is no contribution to the sum.)

$$n!\det\left(A\right) =$$

$$\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}\left(r_1,\cdots,r_n\right)\operatorname{sgn}\left(k_1,\cdots,k_n\right) a_{r_1k_1}\cdots a_{r_nk_n}.$$

This proves the corollary since the formula gives the same number for A as it does for A^T.

Corollary 12.1.7 *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 12.1.4 when two rows are switched, the determinant of the resulting matrix is (-1) times the determinant of the original matrix. By Corollary 12.1.6 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 12.1.8 *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 $\mathbf{w} \in$* span $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$.

The following corollary is also of great use.

Corollary 12.1.9 *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 12.1.7 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 12.1.7

$$\det(A) = \sum_{k=1}^r c_k \det\begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \mathbf{a}_k \end{pmatrix} = 0.$$

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 12.1.10 *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 12.1.11 *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 12.1.4,

$$\begin{aligned}
\det(AB) = & \\
& \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, c_{1k_1} \cdots c_{nk_n} \\
= & \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n) \left(\sum_{r_1} a_{1r_1} b_{r_1k_1}\right) \cdots \left(\sum_{r_n} a_{nr_n} b_{r_nk_n}\right) \\
= & \sum_{(r_1\cdots,r_n)} \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, b_{r_1k_1} \cdots b_{r_nk_n} \left(a_{1r_1} \cdots a_{nr_n}\right) \\
= & \sum_{(r_1\cdots,r_n)} \operatorname{sgn}(r_1 \cdots r_n)\, a_{1r_1} \cdots a_{nr_n} \det(B) = \det(A) \det(B).
\end{aligned}$$

This proves the theorem.

Lemma 12.1.12 *Suppose a matrix is of the form*

$$M = \begin{pmatrix} A & * \\ \mathbf{0} & a \end{pmatrix} \tag{12.19}$$

or

$$M = \begin{pmatrix} A & \mathbf{0} \\ * & a \end{pmatrix} \tag{12.20}$$

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 12.1.1, $\det(M)$ 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 12.20. 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 12.19 use Corollary 12.1.6 and 12.20 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 12.1.13 *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 12.1.14 *Let A be an $n \times n$ matrix where $n \geq 2$. Then*

$$\det(A) = \sum_{j=1}^{n} a_{ij}\operatorname{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij}\operatorname{cof}(A)_{ij}. \tag{12.21}$$

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 12.1.7,

$$\det(A) = \sum_{j=1}^{n} \det(B_j)$$

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 12.1.4, 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 12.1.12,

$$\begin{aligned} \det(B_j) &= (-1)^{n-j}(-1)^{n-i}\det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij} \end{pmatrix}\right) \\ &= (-1)^{i+j}\det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij} \end{pmatrix}\right) = a_{ij}\operatorname{cof}(A)_{ij}. \end{aligned}$$

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.

Theorem 12.1.15 *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 12.1.14 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 12.1.7. 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 12.1.14, 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 12.1.11,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A) \det\left(A^{-1}\right)$$

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 12.1.16 *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 12.1.11 implies

$$\det B \det A = 1$$

and so $\det A \neq 0$. Therefore from Theorem 12.1.15, 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 12.1.15 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,

$$\mathbf{x} = \left(A^{-1}A\right)\mathbf{x} = A^{-1}(A\mathbf{x}) = A^{-1}\mathbf{y}$$

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.

Definition 12.1.17 *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 12.1.14.

Corollary 12.1.18 *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 12.1.19 *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 12.1.20 *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 12.1.21 *The determinant rank equals the row rank.*

Proof: From Theorem 12.1.20, 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 12.1.22 *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 12.1.21, 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 12.1.23 *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 12.1.20. 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 12.1.15 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 12.1.24 *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.

12.1.1 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 $A\mathbf{x} = \mathbf{0}$.

12.2 The Cayley Hamilton Theorem*

Definition 12.2.1 *Let A be an $n \times n$ matrix. The characteristic polynomial is defined as*

$$p_A(t) \equiv \det(tI - A)$$

and the solutions to $p_A(t) = 0$ are called eigenvalues. 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 12.2.2 *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 12.2.3 *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 12.2.4 *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 12.1.15

$$C(\lambda) = p(\lambda)(\lambda I - A)^{-1}.$$

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 12.2.3 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.

12.2.1 Exercises With Answers

1. Find the following determinant by expanding along the second column.

$$\begin{vmatrix} 1 & 3 & 1 \\ 2 & 1 & 5 \\ 2 & 1 & 1 \end{vmatrix}$$

   This is

$$3(-1)^{2+1}\begin{vmatrix} 2 & 5 \\ 2 & 1 \end{vmatrix} + 1(-1)^{1+1}\begin{vmatrix} 1 & 1 \\ 2 & 1 \end{vmatrix} + 1(-1)^{3+2}\begin{vmatrix} 1 & 1 \\ 2 & 5 \end{vmatrix} = 20.$$

2. Compute the determinant by cofactor expansion. Pick the easiest row or column to use.

$$\begin{vmatrix} 2 & 0 & 0 & 1 \\ 2 & 1 & 1 & 0 \\ 0 & 0 & 0 & 3 \\ 2 & 3 & 3 & 1 \end{vmatrix}$$

   You ought to use the third row. This yields

$$3\begin{vmatrix} 2 & 0 & 0 \\ 2 & 1 & 1 \\ 2 & 3 & 3 \end{vmatrix} = (3)(2)\begin{vmatrix} 1 & 1 \\ 3 & 3 \end{vmatrix} = 0.$$

3. Find the determinant using row and column operations.

$$\begin{vmatrix} 5 & 4 & 3 & 2 \\ 3 & 2 & 4 & 3 \\ -1 & 2 & 3 & 3 \\ 2 & 1 & 2 & -2 \end{vmatrix}$$

Replace the first row by 5 times the third added to it and then replace the second by 3 times the third added to it and then the last by 2 times the third added to it. This yields

$$\begin{vmatrix} 0 & 14 & 18 & 17 \\ 0 & 8 & 13 & 12 \\ -1 & 2 & 3 & 3 \\ 0 & 5 & 8 & 4 \end{vmatrix}$$

Now lets replace the third column by -1 times the last column added to it.

$$\begin{vmatrix} 0 & 14 & 1 & 17 \\ 0 & 8 & 1 & 12 \\ -1 & 2 & 0 & 3 \\ 0 & 5 & 4 & 4 \end{vmatrix}$$

Now replace the top row by -1 times the second added to it and the bottom row by -4 times the second added to it. This yields

$$\begin{vmatrix} 0 & 6 & 0 & 5 \\ 0 & 8 & 1 & 12 \\ -1 & 2 & 0 & 3 \\ 0 & -27 & 0 & -44 \end{vmatrix}. \tag{12.22}$$

This looks pretty good because it has a lot of zeros. Expand along the first column and next along the second,

$$(-1)\begin{vmatrix} 6 & 0 & 5 \\ 8 & 1 & 12 \\ -27 & 0 & -44 \end{vmatrix} = (-1)(1)\begin{vmatrix} 6 & 5 \\ -27 & -44 \end{vmatrix} = 129.$$

Alternatively, you could continue doing row and column operations. Switch the third and first row in 12.22 to obtain

$$-\begin{vmatrix} -1 & 2 & 0 & 3 \\ 0 & 8 & 1 & 12 \\ 0 & 6 & 0 & 5 \\ 0 & -27 & 0 & -44 \end{vmatrix}$$

Next take $9/2$ times the third row and add to the bottom.

$$-\begin{vmatrix} -1 & 2 & 0 & 3 \\ 0 & 8 & 1 & 12 \\ 0 & 6 & 0 & 5 \\ 0 & 0 & 0 & -44+(9/2)5 \end{vmatrix}.$$

Finally, take $-6/8$ times the second row and add to the third.

$$-\begin{vmatrix} -1 & 2 & 0 & 3 \\ 0 & 8 & 1 & 12 \\ 0 & 0 & -6/8 & 5+(-6/8)(12) \\ 0 & 0 & 0 & -44+(9/2)5 \end{vmatrix}.$$

Therefore, since the matrix is now upper triangular, the determinant is

$$-\left((-1)\left(8\right)\left(-6/8\right)\left(-44+\left(9/2\right)5\right)\right)=129.$$

4. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

This involved taking the transpose so the determinant of the new matrix is the same as the determinant of the first matrix.

5. Show that for A a 2×2 matrix $\det\left(aA\right)=a^2\det\left(A\right)$ where a is a scalar.

$a^2\det\left(A\right)=a\det\left(A_1\right)$ where the first row of A is replaced by a times it to get A_1. Then $a\det\left(A_1\right)=A_2$ where A_2 is obtained from A by multiplying both rows by a. In other words, $A_2=aA$. Thus the conclusion is established.

6. Use Cramer's rule to find y in

$$\begin{array}{c} 2x+2y+z=3 \\ 2x-y-z=2 \\ x+2z=1 \end{array}$$

From Cramer's rule,

$$y=\frac{\begin{vmatrix} 2 & 3 & 1 \\ 2 & 2 & -1 \\ 1 & 1 & 2 \end{vmatrix}}{\begin{vmatrix} 2 & 2 & 1 \\ 2 & -1 & -1 \\ 1 & 0 & 2 \end{vmatrix}}=\frac{5}{13}.$$

7. Here is a matrix,

$$\begin{pmatrix} e^t & e^{-t}\cos t & e^{-t}\sin t \\ e^t & -e^{-t}\cos t-e^{-t}\sin t & -e^{-t}\sin t+e^{-t}\cos t \\ e^t & 2e^{-t}\sin t & -2e^{-t}\cos t \end{pmatrix}$$

Does there exist a value of t for which this matrix fails to have an inverse? Explain.

$\det\begin{pmatrix} e^t & e^{-t}\cos t & e^{-t}\sin t \\ e^t & -e^{-t}\cos t-e^{-t}\sin t & -e^{-t}\sin t+e^{-t}\cos t \\ e^t & 2e^{-t}\sin t & -2e^{-t}\cos t \end{pmatrix}=5e^te^{2(-t)}\cos^2 t+5e^te^{2(-t)}\sin^2 t$
$=5e^{-t}$ which is never equal to zero for any value of t and so there is no value of t for which the matrix has no inverse.

8. Use the formula for the inverse in terms of the cofactor matrix to find if possible the inverse of the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 6 & 1 \\ 4 & 1 & 1 \end{pmatrix}.$$

First you need to take the determinant

$$\det\begin{pmatrix} 1 & 2 & 3 \\ 0 & 6 & 1 \\ 4 & 1 & 1 \end{pmatrix}=-59$$

and so the matrix has an inverse. Now you need to find the cofactor matrix.

$$\begin{pmatrix} \begin{vmatrix} 6 & 1 \\ 1 & 1 \end{vmatrix} & -\begin{vmatrix} 0 & 1 \\ 4 & 1 \end{vmatrix} & \begin{vmatrix} 0 & 6 \\ 4 & 1 \end{vmatrix} \\ -\begin{vmatrix} 2 & 3 \\ 1 & 1 \end{vmatrix} & \begin{vmatrix} 1 & 3 \\ 4 & 1 \end{vmatrix} & -\begin{vmatrix} 1 & 2 \\ 4 & 1 \end{vmatrix} \\ \begin{vmatrix} 2 & 3 \\ 6 & 1 \end{vmatrix} & -\begin{vmatrix} 1 & 3 \\ 0 & 1 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ 0 & 6 \end{vmatrix} \end{pmatrix}$$

$$= \begin{pmatrix} 5 & 4 & -24 \\ 1 & -11 & 7 \\ -16 & -1 & 6 \end{pmatrix}.$$

Thus the inverse is

$$\frac{1}{-59}\begin{pmatrix} 5 & 4 & -24 \\ 1 & -11 & 7 \\ -16 & -1 & 6 \end{pmatrix}^T$$

$$= \frac{1}{-59}\begin{pmatrix} 5 & 1 & -16 \\ 4 & -11 & -1 \\ -24 & 7 & 6 \end{pmatrix}.$$

If you check this, it does work.

9. Find the eigenvectors and eigenvalues of the matrix, $A = \begin{pmatrix} 8 & -3 & 1 \\ -2 & 7 & 1 \\ 0 & 0 & 10 \end{pmatrix}$. Determine whether the matrix is defective. If nondefective, diagonalize the matrix with an appropriate similarity transformation.

First you need to write the characteristic equation.

$$\det\left(\lambda\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 8 & -3 & 1 \\ -2 & 7 & 1 \\ 0 & 0 & 10 \end{pmatrix}\right) = \det\begin{pmatrix} \lambda-8 & 3 & -1 \\ 2 & \lambda-7 & -1 \\ 0 & 0 & \lambda-10 \end{pmatrix} \tag{12.23}$$

$$= \lambda^3 - 25\lambda^2 + 200\lambda - 500 = 0$$

Next you need to find the solutions to this equation. Of course this is a real joy. If there are any rational zeros they are

$$\pm\frac{\text{factor of } 500}{\text{factor of } 1}$$

I hope to find a rational zero. If there are none, then I don't know what to do at this point. This is a really lousy method for finding eigenvalues and eigenvectors. It only works if things work out well. Lets try 10. You can plug it in and see if it works or you can use synthetic division.

$$\begin{array}{cccccc} 0 & 1 & -25 & 200 & -500 \\ 10 & & 10 & -150 & 500 \\ \hline & 1 & -15 & 50 & 0 \end{array}$$

Yes, it appears 10 works and you can factor the polynomial as $(\lambda-10)\left(\lambda^2-15\lambda+50\right)$ which factors further to $(\lambda-10)(\lambda-5)(\lambda-10)$ so you find the eigenvalues are $5, 10,$

and 10. It remains to find the eigenvectors. First find an eigenvector for $\lambda = 5$. To do this, you find a vector which is sent to 0 by the matrix on the right in 12.23 in which you let $\lambda = 5$. Thus the augmented matrix of the system of equations you need to solve to get the eigenvector is

$$\left(\begin{array}{ccc|c} 5-8 & 3 & -1 & 0 \\ 2 & 5-7 & -1 & 0 \\ 0 & 0 & 5-10 & 0 \end{array}\right)$$

Now the row reduced echelon form is

$$\left(\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

and so you need $x = y$ and $z = 0$. An eigenvector is $(1,1,0)^T$. Now you have the glorious opportunity to solve for the eigenvectors associated with $\lambda = 10$. You do it the same way. The augmented matrix for the system of equations you solve to find the eigenvectors is

$$\left(\begin{array}{ccc|c} 10-8 & 3 & -1 & 0 \\ 2 & 10-7 & -1 & 0 \\ 0 & 0 & 10-10 & 0 \end{array}\right)$$

The row reduced echelon form is

$$\left(\begin{array}{ccc|c} 1 & \frac{3}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

and so you need $x = -\frac{3}{2}y + \frac{1}{2}z$. It follows the eigenvectors for $\lambda = 10$ are

$$\left(-\frac{3}{2}y + \frac{1}{2}z, y, z\right)^T$$

where $x, y \in \mathbb{R}$, not both equal to zero. Why? Let $y = 2$ and $z = 0$. This gives the vector,

$$(-3,2,0)^T$$

as one of the eigenvectors. You could also let $y = 0$ and $z = 2$ to obtain another eigenvector,

$$(1,0,2)^T.$$

If there exists a basis of eigenvectors, then the matrix is nondefective and as discussed above, the matrix can be diagonalized by considering $S^{-1}AS$ where the columns of S are the eigenvectors. In this case, I have found three eigenvectors and so it remains to determine whether these form a basis. Remember how to do this. You let them be the columns of a matrix and then find the rank of this matrix. If it is three, then they are a basis because they are linearly independent and the vectors are in $\mathbb{R}^3$. This is equivalent to the following matrix has an inverse.

$$\left(\begin{array}{ccc} 1 & -3 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{array}\right)$$

$$\left(\begin{array}{ccc} 1 & -3 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{array}\right)^{-1} = \left(\begin{array}{ccc} \frac{2}{5} & \frac{3}{5} & -\frac{1}{5} \\ -\frac{1}{5} & \frac{1}{5} & \frac{1}{10} \\ 0 & 0 & \frac{1}{2} \end{array}\right)$$

Then to diagonalize

$$\begin{pmatrix} \frac{2}{5} & \frac{3}{5} & -\frac{1}{5} \\ -\frac{1}{5} & \frac{1}{5} & \frac{1}{10} \\ 0 & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 8 & -3 & 1 \\ -2 & 7 & 1 \\ 0 & 0 & 10 \end{pmatrix} \begin{pmatrix} 1 & -3 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} =$$

$$\begin{pmatrix} 5 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}$$

Isn't this stuff marvelous! You can know this matrix is nondefective at the point when you find the eigenvectors for the repeated eigenvalue. This eigenvalue was repeated with multiplicity 2 and there were two parameters, y and z in the description of the eigenvectors. Therefore, the matrix is nondefective. Also note that there is no uniqueness for the similarity transformation.

10. Now consider the matrix, $\begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$. Find its eigenvectors and eigenvalues and determine whether it is defective.

The characteristic equation is

$$\det\left(\lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}\right) = 0$$

thus the characteristic equation is

$$(\lambda - 2)(\lambda - 1)^2 = 0.$$

The zeros are $1, 1, 2$. Lets find the eigenvectors for $\lambda = 1$. The augmented matrix for the system you need to solve is

$$\left(\begin{array}{ccc|c} -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{array}\right)$$

The row reduced echelon form is

$$\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Then you find $x = y = 0$ and there is no restriction on z. Thus the eigenvectors are of the form

$$(0, 0, z)^T, \ z \in \mathbb{R}.$$

The eigenvalue had multiplicity 2 but the eigenvectors depend on only one parameter. Therefore, the matrix is defective and cannot be diagonalized. The other eigenvector comes from row reducing the following

$$2\left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right) - \left(\begin{array}{ccc|c} 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \end{array}\right) = \left(\begin{array}{ccc|c} 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{array}\right)$$

The row reduced echelon form is

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Therefore the eigenvectors are of the form $(x, 0, -x)^T$. One such eigenvector is $(1, 0, -1)^T$.

11. Let M be an $n \times n$ matrix. Then define the adjoint of M,denoted by M^* to be the transpose of the conjugate of M. For example,

$$\begin{pmatrix} 2 & i \\ 1+i & 3 \end{pmatrix}^* = \begin{pmatrix} 2 & 1-i \\ -i & 3 \end{pmatrix}.$$

A matrix, M, is self adjoint if $M^* = M$. Show the eigenvalues of a self adjoint matrix are all real. If the self adjoint matrix has all real entries, it is called symmetric. Show that the eigenvalues and eigenvectors of a symmetric matrix occur in conjugate pairs.

First note that for $\mathbf{x}$ a vector, $\mathbf{x}^*\mathbf{x} = |\mathbf{x}|^2$. This is because

$$\mathbf{x}^*\mathbf{x} = \sum_k \overline{x_k} x_k = \sum_k |x_k|^2 \equiv |\mathbf{x}|^2.$$

Also note that $(AB)^* = B^*A^*$ because this holds for transposes.This implies that for A an $n \times m$ matrix,

$$\mathbf{x}^* A^* \mathbf{x} = (A\mathbf{x})^* \mathbf{x}$$

Then if $M\mathbf{x} = \lambda\mathbf{x}$

$$\begin{aligned} \overline{\lambda}\mathbf{x}^*\mathbf{x} &= (\lambda\mathbf{x})^*\mathbf{x} = (M\mathbf{x})^*\mathbf{x} = \mathbf{x}^*M^*\mathbf{x} \\ &= \mathbf{x}^*M\mathbf{x} = \mathbf{x}^*\lambda\mathbf{x} = \lambda\mathbf{x}^*\mathbf{x} \end{aligned}$$

and so $\lambda = \overline{\lambda}$ showing that λ must be real.

12. Suppose A is an $n \times n$ matrix consisting entirely of real entries but $a+ib$ is a complex eigenvalue having the eigenvector, $\mathbf{x} + i\mathbf{y}$. Here $\mathbf{x}$ and $\mathbf{y}$ are real vectors. Show that then $a - ib$ is also an eigenvalue with the eigenvector, $\mathbf{x} - i\mathbf{y}$. **Hint:** You should remember that the conjugate of a product of complex numbers equals the product of the conjugates. Here $a + ib$ is a complex number whose conjugate equals $a - ib$.

If A is real then the characteristic equation has all real coefficients. Therefore, letting $p(\lambda)$ be the characteristic polynomial,

$$0 = p(\lambda) = \overline{p(\lambda)} = p\left(\overline{\lambda}\right)$$

showing that $\overline{\lambda}$ is also an eigenvalue.

13. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} -10 & -2 & 11 \\ -18 & 6 & -9 \\ 10 & -10 & -2 \end{pmatrix}.$$

Determine whether the matrix is defective.

The matrix has eigenvalues -12 and 18. Of these, -12 is repeated with multiplicity two. Therefore, you need to see whether the eigenspace has dimension two. If it does,

then the matrix is non defective. If it does not, then the matrix is defective. The row reduced echelon form for the system you need to solve is

$$\left(\begin{array}{ccc|c} 2 & -2 & 11 & 0 \\ -18 & 18 & -9 & 0 \\ 10 & -10 & 10 & 0 \end{array}\right)$$

and its row reduced echelon form is

$$\left(\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Therefore, the eigenspace is of the form

$$\begin{pmatrix} t \\ t \\ 0 \end{pmatrix}$$

This is only one dimensional and so the matrix is defective.

14. Here is a matrix. $A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & -1 & 0 \\ 0 & -2 & 1 \end{pmatrix}$. Find a formula for A^n where n is an integer.

First you find the eigenvectors and eigenvalues. $\begin{pmatrix} 1 & 2 & 0 \\ 0 & -1 & 0 \\ 0 & -2 & 1 \end{pmatrix}$, eigenvectors:

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \leftrightarrow 1, \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} \leftrightarrow -1.$$

The matrix, S used to diagonalize the matrix is obtained by letting these vectors be the columns of S. Then S^{-1} is given by

$$S^{-1} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Then $S^{-1}AS$ equals

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 0 & -1 & 0 \\ 0 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \equiv D$$

Then $A = SDS^{-1}$ and $A^n = SD^nS^{-1}$. Now it is easy to find D^n.

$$D^n = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & (-1)^n \end{pmatrix}$$

Therefore,

$$\begin{aligned} A^n &= \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & (-1)^n \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1-(-1)^n & 0 \\ 0 & (-1)^n & 0 \\ 0 & -1+(-1)^n & 1 \end{pmatrix}. \end{aligned}$$

15. Suppose the eigenvalues of A are $\lambda_1, \cdots, \lambda_n$ and that A is nondefective. Show that $e^{At} = S \begin{pmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{pmatrix} S^{-1}$ where S is the matrix which satisfies $S^{-1}AS = D$.

The diagonal matrix, D has the same characteristic equation as A why? and so it has the same eigenvalues. However the eigenvalues of D are the diagonal entries and so the diagonal entries of D are the eigenvalues of A. Now

$$S^{-1}tAS = tD$$

and

$$(tD)^n = \begin{pmatrix} (\lambda_1 t)^n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (\lambda_n t)^n \end{pmatrix}$$

Therefore,

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{1}{n!} (tD)^n &= \sum_{n=0}^{\infty} \frac{\left(S^{-1}tAS\right)^n}{n!} \\ &= S^{-1} \sum_{n=0}^{\infty} \frac{(tA)^n}{n!} S. \end{aligned}$$

Now the left side equals

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{1}{n!} (tD)^n &= \sum_{n=0}^{\infty} \frac{1}{n!} \begin{pmatrix} (\lambda_1 t)^n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (\lambda_n t)^n \end{pmatrix} \\ &= \begin{pmatrix} \sum_{n=0}^{\infty} \frac{(\lambda_1 t)^n}{n!} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{n=0}^{\infty} \frac{(\lambda_n t)^n}{n!} \end{pmatrix} \\ &= \begin{pmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{pmatrix}. \end{aligned}$$

Therefore,

$$e^{tA} \equiv \sum_{n=0}^{\infty} \frac{(tA)^n}{n!} = S \begin{pmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{pmatrix} S^{-1}.$$

Do you think you understand this? If so, think again. What exactly do you mean by an infinite sum? Actually there is no problem here. You can do this just fine and the sums converge in the sense that the ij^{th} entries converge in the partial sums. Think about this. You know what you need from calculus to see this.

16. Show that if A is similar to B then A^T is similar to B^T.

    This is easy. $A = S^{-1}BS$ and so $A^T = S^T B^T \left(S^{-1}\right)^T = S^T B^T \left(S^T\right)^{-1}$.

17. Suppose $A^m = 0$ for some m a positive integer. Show that if A is diagonalizable, then $A = 0$.

    Since $A^m = 0$ suppose $S^{-1}AS = D$. Then raising to the m^{th} power, $D^m = S^{-1}A^m S = 0$. Therefore, $D = 0$. But then $A = S0S^{-1} = 0$.

18. Find the complex eigenvalues and eigenvectors of the matrix $\begin{pmatrix} 1 & 1 & -6 \\ 7 & -5 & -6 \\ -1 & 7 & 2 \end{pmatrix}$.

    Determine whether the matrix is defective.

    After wading through much affliction you find the eigenvalues are $-6, 2+6i, 2-6i$. Since these are distinct, the matrix cannot be defective. We must find the eigenvectors for these eigenvalues. The augmented matrix for the system of equations which must be solved to find the eigenvectors associated with $2-6i$ is

$$\left(\begin{array}{ccc|c} -1+6i & 1 & -6 & 0 \\ 7 & -7+6i & -6 & 0 \\ -1 & 7 & 6i & 0 \end{array}\right).$$

    The row reduced echelon form is

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

    and so the eigenvectors are of the form

$$t\begin{pmatrix} -i \\ -i \\ 1 \end{pmatrix}.$$

    You can check this as follows

$$\begin{pmatrix} 1 & 1 & -6 \\ 7 & -5 & -6 \\ -1 & 7 & 2 \end{pmatrix}\begin{pmatrix} -i \\ -i \\ 1 \end{pmatrix} = \begin{pmatrix} -6-2i \\ -6-2i \\ 2-6i \end{pmatrix}$$

    and

$$(2-6i)\begin{pmatrix} -i \\ -i \\ 1 \end{pmatrix} = \begin{pmatrix} -6-2i \\ -6-2i \\ 2-6i \end{pmatrix}.$$

    It follows that the eigenvectors for $\lambda = 2+6i$ are

$$t\begin{pmatrix} i \\ i \\ 1 \end{pmatrix}.$$

This is because A is real. If $A\mathbf{v} = \lambda\mathbf{v}$, then taking the conjugate,

$$A\overline{\mathbf{v}} = \overline{A\mathbf{v}} = \overline{\lambda}\overline{\mathbf{v}}.$$

It only remains to find the eigenvector for $\lambda = -6$. The augmented matrix to row reduce is

$$\begin{pmatrix} 7 & 1 & -6 & | & 0 \\ 7 & 1 & -6 & | & 0 \\ -1 & 7 & 8 & | & 0 \end{pmatrix}$$

The row reduced echelon form is

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

Then an eigenvector is

$$\begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}.$$

Part VI

Curves, Curvilinear Motion, Surfaces

Outcomes

Curves in Space

A. Identify the domain of a vector function.

B. Identify a curve given its parameterization.

C. Determine combinations of vector functions such as sums, vector products and scalar products.

D. Define limit, derivative and integral for vector functions.

E. Evaluate limits, derivatives and integrals of vector functions.

F. Find the line tangent to a curve at a given point.

G. Describe what is meant by arc length.

H. Evaluate the arc length of a curve.

I. Recall, derive and apply rules to combinations of vector functions for the following:

   (a) limits
   (b) differentiation
   (c) integration

Reading: Multivariable Calculus 1.6
Outcome Mapping:

A. 1

B. 2

C. 3,4

D. C1

E. C2,C3,C4

F. 5,16

G. C5

H. 6

I. 7,8,14

Curvilinear Motion

A. Sketch the curve determined by a vector function in 2-space or 3-space.

B. Parameterize a curve in 2-space or 3-space.

C. Given the position vector function of a moving object, calculate the velocity, speed, and acceleration of the object.

D. Model and analyze curvilinear motion in applications.

Reading: Multivariable Calculus 1.7
Outcome Mapping:

A. D1

B. D2

C. 1

D. 3,5,6,10,18

Curvature

A. Recall the definitions of unit tangent, unit normal, binormal and osculating plane for a space curve. Illustrate each graphically.

B. Calculate the curvature, the radius of curvature, the center of curvature and the osculating plane for a space curve.

C. Derive formulas for the curvature of a parameterized curve and the curvature of a plane curve given as a function.

D. Determine the tangential and normal components of acceleration for a given path.

Reading: Multivariable Calculus 1.8
Learning Module: Moving Trihedron
Outcome Mapping:

A. E1,E2

B. 1,5

C. 4

D. 2

Surfaces

A. Identify standard quadratic surfaces given their functions or graphs.

B. Sketch the graph of a quadratic surface by identifying the intercepts, traces, sections, symmetry and boundedness or unboundedness of the surface.

Reading: Multivariable Calculus 1.4
Outcome Mapping:

A. 2,3

B. 3,4,6

Quadric Surfaces 9 Oct.

Quiz

1. Find the eigenvectors of the matrix,

$$\begin{pmatrix} 4 & 1 & 0 \\ 2 & 8 & 3 \\ -2 & -2 & 3 \end{pmatrix}$$

   given the eigenvalues are 6 and 3. Also tell whether the matrix is defective or non defective.

2. Here is a matrix.

$$A = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 \\ -\frac{1}{4} & -\frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

   Find A^{50}. The eigenvalues of this matrix are 1 and $1/2$ and eigenvectors for these eigenvalues are $\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$ for $\lambda = 1$ and $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$ for $\lambda = 1/2$.

3. Here is a Markov matrix. This is also called a migration matrix or a stochastic matrix.

$$A = \begin{pmatrix} 1/2 & 1/3 & 1/4 \\ 0 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/2 \end{pmatrix}$$

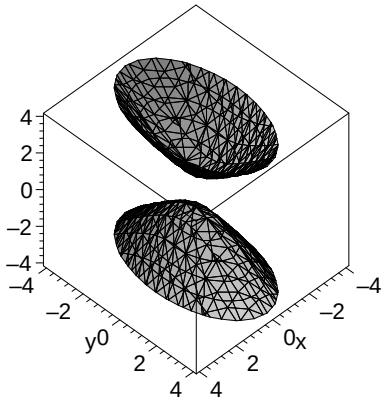
   Find

$$\lim_{n\to\infty} A^n \begin{pmatrix} 16 \\ 30 \\ 5 \end{pmatrix}.$$

Recall the equation of an arbitrary plane is an equation of the form $ax + by + cz = d$. More generally, a set of points of the following form

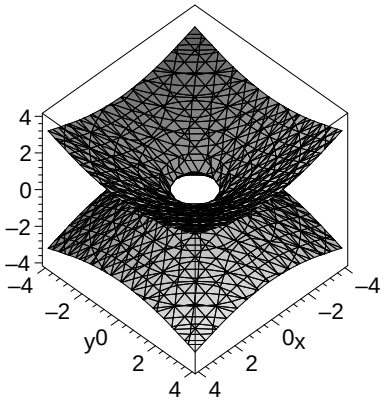
$$\{(x, y, z) : f(x, y, z) = c\}$$

is called a level surface. 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.



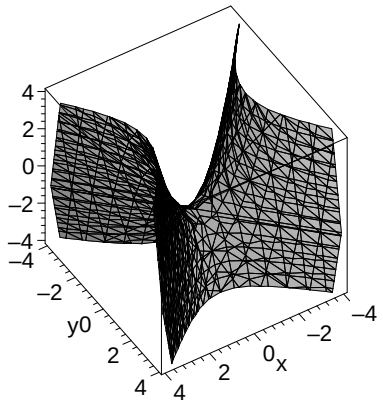
$z^2/a^2 - x^2/b^2 - y^2/c^2 = 1$

hyperboloid of two sheets



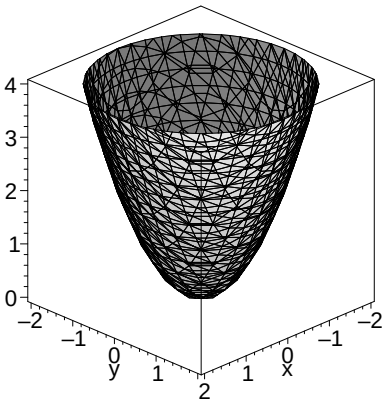
$x^2/b^2 + y^2/c^2 - z^2/a^2 = 1$

hyperboloid of one sheet



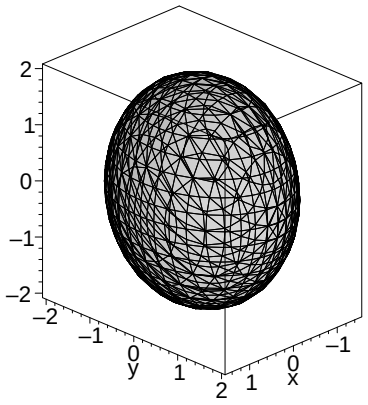
$z = x^2/a^2 - y^2/b^2$

hyperbolic paraboloid



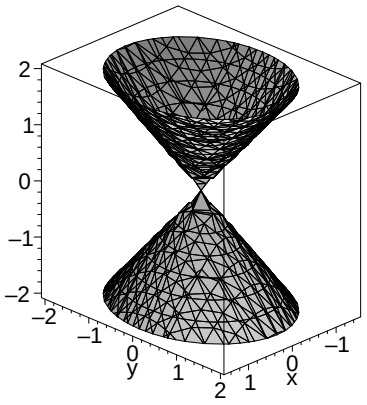
$x^2/b^2 + y^2/c^2 = z$

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 13.0.5 *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 13.0.6 *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 13.0.7 *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 13.0.8 *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

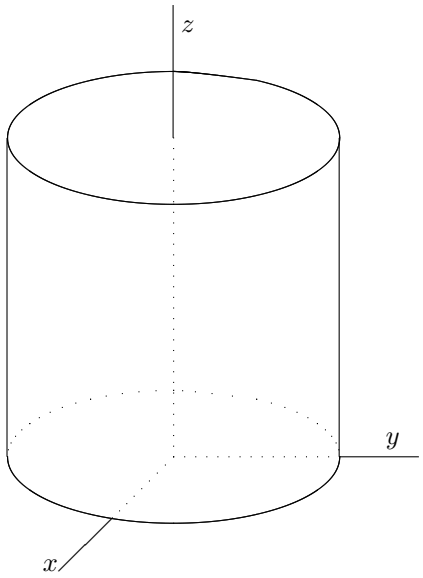
$$(1,1,1)+\left(-\frac{1}{2}+\frac{1}{10}\sqrt{5}\right)(1,2,1), \text{ and } (1,1,1)+\left(-\frac{1}{2}-\frac{1}{10}\sqrt{5}\right)(1,2,1)$$

That is

$$\left(\frac{1}{2}+\frac{1}{10}\sqrt{5},\frac{1}{5}\sqrt{5},\frac{1}{2}+\frac{1}{10}\sqrt{5}\right),\left(\frac{1}{2}-\frac{1}{10}\sqrt{5},-\frac{1}{5}\sqrt{5},\frac{1}{2}-\frac{1}{10}\sqrt{5}\right).$$

A cylinder generated by a curve, C is the surface generated by moving the curve C through space along a straight line. If you are given a level surface of the form $f(x,y)=c$ this will yield a cylinder parallel to the z axis. Here is why: If $z=0$, then $f(x,y)=c$ is a curve in the plane, $z=0$. If $z=1$, then you get exactly the same curve but just shifted up to a height of 1. Similarly, $f(y,z)=c$ gives a cylinder parallel to the x axis and $f(x,z)=c$ gives one which is parallel to the y axis.

Example 13.0.9 *Consider the cylinder $x^2+y^2=1$. Sketch its graph.*



You see that at every height above or below the $z=0$ plane if you slice it at that level, you will just see the graph of $x^2+y^2=1$ at that level. This is a circle of radius 1. Since the equation describing the surface does not depend on z, this is why it looks the same at every level.

Curves In Space 10,11 Oct.

14.1 Limits Of A Vector Valued Function Of One Variable

Quiz

1. Find the determinant of the matrix,

$$\begin{pmatrix} 1 & 2 & 1 \\ 2 & 1 & -1 \\ -1 & -3 & 1 \end{pmatrix}$$

2. Find all eigenspaces and eigenvalues for the matrix,

$$\begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

3. Find the eigenspace for the eigenvalue $\lambda = 2$ for the matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ -1 & 1 & 2 \end{pmatrix}$$

A vector valued function is just one which has vector values. For example, consider

$$\left(\cos t, t^2, t+1\right)$$

where $t \in [0,2]$. Each value of t corresponds to a point in $\mathbb{R}^3$ whose coordinates are as given. Thus when $t = 0$, the point in $\mathbb{R}^3$ is $(1,0,1)$ and when $t = \pi/2$ the point is $\left(0, \left(\frac{\pi}{2}\right)^2, \frac{\pi}{2}+1\right)$, etc. Often t will be considered as time. Thus, in this case, the vector valued function gives the coordinates of a point which is moving in three dimensions as a function of time. Imagine a fly buzzing around the room for example. Let the origin be a corner of the room and consider the position vector of the fly. This position vector could be described by a vector valued function of the form $(x(t), y(t), z(t))$ where t is in some interval. Here $x(t)$ is the x coordinate of the fly, $y(t)$, the y coordinate, and $z(t)$, the z coordinate corresponding to a given time. Later the physical significance of all this will be discussed more. For right now, t will just be in some interval and general vector valued functions will be considered.

Definition 14.1.1 *Let $\mathbf{x}(t) = (x_1(t), \cdots, x_n(t))$ for $t \in [a,b]$ be a vector valued function. The curve **parameterized** by this vector valued function is the set of points in $\mathbb{R}^n$ which are obtained by letting t vary over the interval, $[a,b]$. The vector valued function is also called a **parameterization** of this curve. The variable, t is called a **parameter**. More generally, if $\mathbf{x}(t) = (x_1(t), \cdots, x_n(t))$ is given where each $x_i(t)$ is a formula the **domain** of $\mathbf{x}$ is defined to be the set where each of the $x_i(t)$ is defined. It is denoted by $D(\mathbf{x})$.*

Example 14.1.2 *Let $\mathbf{x}(t) = \left(\frac{1}{t}, \sqrt{1-t^2}, \sin(t)\right)$ find the domain of $\mathbf{x}$.*

You need each function to make sense. Thus you must have $-1 \leq t \leq 1$ and $t \neq 0$. The domain is $[-1,0) \cup (0,1]$.

In useful situations the domain will typically be an interval.

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

$$\lim_{s \to -\infty} \mathbf{f}(s).$$

Definition 14.1.3 *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 $t > l$,

$$|\mathbf{f}(t) - \mathbf{L}| < \varepsilon \tag{14.1}$$

and

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

if for every $\varepsilon > 0$ there exists l such that whenever $t < l$, 14.1 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 $p > 1$.

Observation 14.1.4 *Let* $\mathbf{f}(t) = (f_1(t), \cdots, f_n(t))$ *and let* $\mathbf{L} = (L_1, \cdots, L_n)$. *Then* $\lim_{t\to a} \mathbf{f}(t) = \mathbf{L}$ *if and only if* $\lim_{t\to a} f_k(t) = L_k$ *for each* k.

Example 14.1.5 *Let* $\mathbf{f}(t) = \left(\cos t, \sin t, t^2+1, \ln(t)\right)$. *Find* $\lim_{t\to \pi/2} \mathbf{f}(t)$.

Using the above observation, this limit equals

$$\begin{aligned} & \left(\lim_{t\to\pi/2} \cos t, \lim_{t\to\pi/2} (\sin t), \lim_{t\to\pi/2} \left(t^2+1\right), \lim_{t\to\pi/2} \ln(t)\right) \\ = & \left(0, 1, \left(\frac{\pi^2}{4}+1\right), \ln\left(\frac{\pi}{2}\right)\right). \end{aligned}$$

Example 14.1.6 *Let* $\mathbf{f}(t) = \left(\frac{\sin t}{t}, t^2, t+1\right)$. *Find* $\lim_{t\to 0} \mathbf{f}(t)$.

Recall that $\lim_{t\to 0} \frac{\sin t}{t} = 1$. Then using the above observation, $\lim_{t\to 0} \mathbf{f}(t) = (1, 0, 1)$.

14.2 The Derivative And Integral

The following definition is on the derivative and integral of a vector valued function of one variable.

Definition 14.2.1 *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 $\mathbf{f} \in R([a,b])$. *In other words,* f *is Riemann integrable. That is you can take the integral.*

It is easier to write $\mathbf{f} \in R([a,b])$ than to write $\mathbf{f}$ is Riemann integrable. Thus, if you see $\mathbf{f} \in R([a,b])$, think: $\int_a^b \mathbf{f}(x)\, dx$ exists.

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 14.2.2 *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 14.2.3 *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 14.2.4 *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 14.2.5 *Let $\mathbf{f}(t) = (at, bt)$ where a, b are constants. Find $\mathbf{f}'(t)$.*

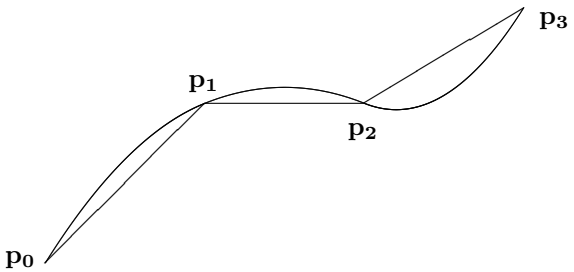
From the above discussion this derivative is just the vector valued functions whose components consist of the derivatives of the components of $\mathbf{f}$. Thus $\mathbf{f}'(t) = (a, b)$.

14.2.1 Arc Length

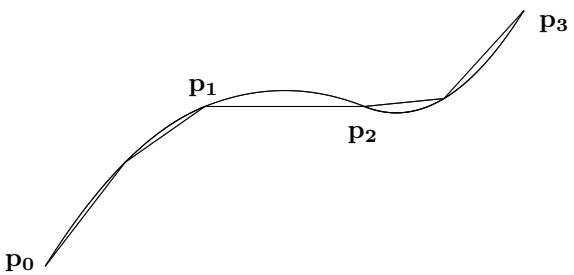
C is a **smooth curve** in $\mathbb{R}^n$ if there exists an interval, $[a,b] \subseteq \mathbb{R}$ and functions $x_i : [a,b] \to \mathbb{R}$ such that the following conditions hold

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. It is proved in Section 16.2.2 which starts on Page 295.

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.

Definition 14.2.6 *A curve C is piecewise smooth if there exist points on this curve, $\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_n$ such that, denoting $C_{\mathbf{p}_{k-1}\mathbf{p}_k}$ the part of the curve joining $\mathbf{p}_{k-1}$ and $\mathbf{p}_k$, it follows $C_{\mathbf{p}_{k-1}\mathbf{p}_k}$ is a smooth curve and $\cup_{k=1}^{n} C_{\mathbf{p}_{k-1}\mathbf{p}_k} = C$. In other words, it is piecewise smooth if it consists of a finite number of smooth curves linked together.*

To find the length of a piecewise smooth curve, just sum the lengths of the smooth pieces described above.

Example 14.2.7 *The parameterization for a smooth curve is $\mathbf{r}(t) = \left(t, 2t^2, t^2\right)$ for $t \in [0,1]$. Find the length of this curve.*

From the above, the length is

$$\int_0^1 |\mathbf{r}'(t)| \, dt = \int_0^1 \sqrt{1 + (4t)^2 + (2t)^2} dt = \frac{1}{2}\sqrt{21} + \frac{1}{20}\sqrt{5} \ln\left(2\sqrt{5} + \sqrt{21}\right).$$

You need to use a trig substitution of some sort to do this integral but it is routine. Of course, if you can't find an antiderivative, then you solve it numerically.

Example 14.2.8 *The parameterization for a smooth curve is* $\mathbf{r}(t) = \left(t, t^3, t^2\right)$ *for* $t \in [0,1]$. *Find the length of this curve.*

The length is

$$\int_0^1 \sqrt{1 + (3t^2)^2 + (2t)^2}dt = \int_0^1 \sqrt{1 + 9t^4 + 4t^2}dt$$

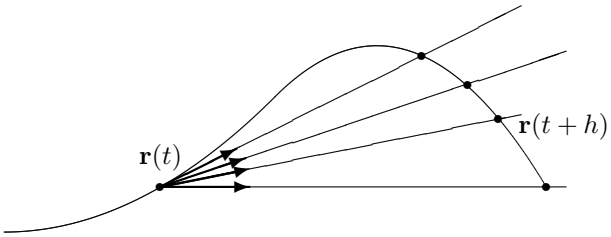
and I have no clue how to find an antiderivative for this. Therefore, I find the integral numerically.

$$\int_0^1 \sqrt{1 + 9t^4 + 4t^2}dt = 1.863$$

This is all right to do. Numerical methods are allowed and somtimes that is all you can get.

14.2.2 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 $\mathbf{r}$.



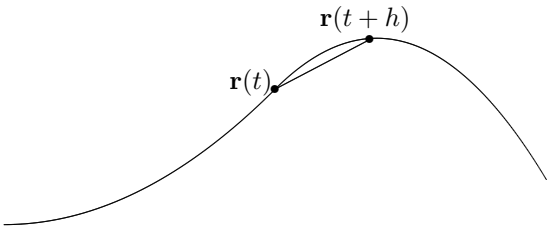
In this picture there are unit vectors in the direction of the vector from $\mathbf{r}(t)$ to $\mathbf{r}(t+h)$. You can see that it is reasonable to suppose these unit vectors, if they converge, converge to a unit vector, $\mathbf{T}$ which is tangent to the curve at the point $\mathbf{r}(t)$. Now each 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 travelled 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

$$\frac{\mathbf{r}(t_0+h)-\mathbf{r}(t_0)}{h} \tag{14.2}$$

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, 14.2 converges to

$$\mathbf{v}\equiv(v_1,\cdots,v_q)$$

where $v_k=r_k'(t)$. This by Observation 14.1.4, which says that the term in 14.2 gets close to a vector, $\mathbf{v}$ if and only if all the coordinate functions of the term in 14.2 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 14.2.9 *Let* $\mathbf{r}(t)=\left(\sin t,t^2,t+1\right)$ *for* $t\in[0,5]$. *Find a tangent line to the curve parameterized by* $\mathbf{r}$ *at the point* $\mathbf{r}(2)$.

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 14.2.10 *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 $\mathbf{r}'(1)=(\cos 1,2,1)$.

14.2.3 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 14.2.11 *Let $a, b \in \mathbb{R}$ and suppose $\mathbf{f}'(t)$ and $\mathbf{g}'(t)$ exist. Then the following formulas are obtained.*

$$(a\mathbf{f} + b\mathbf{g})'(t) = a\mathbf{f}'(t) + b\mathbf{g}'(t). \tag{14.3}$$

$$(\mathbf{f} \cdot \mathbf{g})'(t) = \mathbf{f}'(t) \cdot \mathbf{g}(t) + \mathbf{f}(t) \cdot \mathbf{g}'(t) \tag{14.4}$$

If $\mathbf{f}, \mathbf{g}$ have values in $\mathbb{R}^3$, then

$$(\mathbf{f} \times \mathbf{g})'(t) = \mathbf{f}(t) \times \mathbf{g}'(t) + \mathbf{f}'(t) \times \mathbf{g}(t) \tag{14.5}$$

The formulas, 14.4, and 14.5 are referred to as the product rule.

Proof: The first formula is left for you to prove. Consider the second, 14.4.

$$\lim_{h\to 0} \frac{\mathbf{f} \cdot \mathbf{g}(t+h) - \mathbf{fg}(t)}{h}$$

$$\begin{aligned}
&= \lim_{h\to 0} \frac{\mathbf{f}(t+h) \cdot \mathbf{g}(t+h) - \mathbf{f}(t+h) \cdot \mathbf{g}(t)}{h} + \frac{\mathbf{f}(t+h) \cdot \mathbf{g}(t) - \mathbf{f}(t) \cdot \mathbf{g}(t)}{h} \\
&= \lim_{h\to 0} \left(\mathbf{f}(t+h) \cdot \frac{(\mathbf{g}(t+h) - \mathbf{g}(t))}{h} + \frac{(\mathbf{f}(t+h) - \mathbf{f}(t))}{h} \cdot \mathbf{g}(t) \right) \\
&= \lim_{h\to 0} \sum_{k=1}^{n} f_k(t+h) \frac{(g_k(t+h) - g_k(t))}{h} + \sum_{k=1}^{n} \frac{(f_k(t+h) - f_k(t))}{h} g_k(t) \\
&= \sum_{k=1}^{n} f_k(t) g_k'(t) + \sum_{k=1}^{n} f_k'(t) g_k(t) \\
&= \mathbf{f}'(t) \cdot \mathbf{g}(t) + \mathbf{f}(t) \cdot \mathbf{g}'(t).
\end{aligned}$$

Formula 14.5 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 50. You can also see this is true by using the distributive law of the cross product.

$$\begin{aligned}
&\mathbf{f}(t+h) \times \mathbf{g}(t+h) - \mathbf{f}(t) \times \mathbf{g}(t) \\
= \;& \mathbf{f}(t+h) \times \mathbf{g}(t+h) - \mathbf{f}(t+h) \times \mathbf{g}(t) + \mathbf{f}(t+h) \times \mathbf{g}(t) - \mathbf{f}(t) \times \mathbf{g}(t)
\end{aligned}$$

and so

$$\begin{aligned}
&\frac{1}{h} (\mathbf{f}(t+h) \times \mathbf{g}(t+h) - \mathbf{f}(t) \times \mathbf{g}(t)) \\
= \;& \mathbf{f}(t+h) \times \left(\frac{\mathbf{g}(t+h) - \mathbf{g}(t)}{h} \right) + \left(\frac{\mathbf{f}(t+h) - \mathbf{f}(t)}{h} \right) \times \mathbf{g}(t)
\end{aligned}$$

Now assuming the cross product is continuous, (This is obvious from either the component or the geometric description of the cross product.) you can take a limit in the above as $h \to 0$ and obtain

$$\mathbf{f}(t) \times \mathbf{g}'(t) + \mathbf{f}'(t) \times \mathbf{g}(t).$$

It is exactly like the product rule for scalar valued functions except you need to be very careful about the order in which things are multiplied becasue the cross product is not commutative.

Example 14.2.12 *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 14.5 this equals$(2t, \cos t, -\sin t) \times (t, \ln(t+1), 2t) + \left(t^2, \sin t, \cos t\right) \times \left(1, \frac{1}{t+1}, 2\right)$

$$\begin{aligned} &= \left(2(\cos t)t + \sin t \ln(t+1), -(\sin t)t - 4t^2, 2t\ln(t+1) - (\cos t)t\right) \\ &\quad + \left(2\sin t - \frac{\cos t}{t+1}, \cos t - 2t^2, \frac{t^2}{t+1} - \sin t\right) \end{aligned}$$

$$\begin{aligned} &= (2(\cos t)t + \sin t \ln(t+1) + 2\sin t - \tfrac{\cos t}{t+1}, -(\sin t)t - 6t^2 + \cos t, \\ &\qquad 2t\ln(t+1) - (\cos t)t + \tfrac{t^2}{t+1} - \sin t) \end{aligned}$$

Example 14.2.13 *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 14.2.14 *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 14.2.15 *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 14.2.16 *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

$$\frac{(\mathbf{r}(t) - \mathbf{P}_0) \cdot (-\mathbf{r}'(t))}{|\mathbf{r}(t) - \mathbf{P}_0|} = \frac{(\mathbf{r}(t) - \mathbf{P}_1) \cdot (\mathbf{r}'(t))}{|\mathbf{r}(t) - \mathbf{P}_1|} \tag{14.6}$$

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,

$$\begin{aligned}\frac{d}{dt}|\mathbf{r}(t) - \mathbf{P}_1| &= \frac{1}{2}\left((\mathbf{r}(t) - \mathbf{P}_1) \cdot (\mathbf{r}(t) - \mathbf{P}_1)\right)^{-1/2} 2\left((\mathbf{r}(t) - \mathbf{P}_1) \cdot \mathbf{r}'(t)\right) \\ &= \frac{(\mathbf{r}(t) - \mathbf{P}_1) \cdot (\mathbf{r}'(t))}{|\mathbf{r}(t) - \mathbf{P}_1|}.\end{aligned}$$

Therefore, from 14.6,

$$\frac{d}{dt}(|\mathbf{r}(t) - \mathbf{P}_1|) + \frac{d}{dt}(|\mathbf{r}(t) - \mathbf{P}_0|) = 0$$

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.

14.2.4 Leibniz's Notation

Leibniz's notation also generalizes routinely. For example, $\frac{d\mathbf{y}}{dt} = \mathbf{y}'(t)$ with other similar notations holding.

14.2.5 Exercises With Answers

1. Find the following limits if possible

   (a) $\lim_{x\to 0+}\left(\frac{|x|}{x}, \sin 2x/x, \frac{\tan x}{x}\right) = (1, 2, 1)$

   (b) $\lim_{x\to 0+}\left(\frac{x}{|x|}, \cos x, e^{2x}\right) = (1, 1, 1)$

   (c) $\lim_{x\to 4}\left(\frac{x^2-16}{x+4}, x-7, \frac{\tan 7x}{5x}\right) = \left(0, -3, \frac{7}{5}\right)$

2. 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.

   You need to differentiate this. $\mathbf{r}'(t) = \left(2(t-1), (t-1)^2 \frac{4t^2-t+3}{\sqrt{(t^2+1)}}, -(t-1)^2 \frac{2t-5}{t^6}\right)$.
   Now you need to find the value(s) of t where $\mathbf{r}'(t) = \mathbf{0}$.

3. Let $\mathbf{r}(t) = \left(\sin t, t^2, 2t+1\right)$ for $t \in [0,4]$. Find a tangent line to the curve parameterized by $\mathbf{r}$ at the point $\mathbf{r}(2)$.

   $\mathbf{r}'(t) = (\cos t, 2t, 2)$. When $t = 2$, the point on the curve is $(\sin 2, 4, 5)$. A direction vector is $\mathbf{r}'(2)$ and so a tangent line is $\mathbf{r}(t) = (\sin 2, 4, 5) + t(\cos 2, 4, 2)$.

4. Let $\mathbf{r}(t) = \left(\sin t, \cos\left(t^2\right), t+1\right)$ for $t \in [0,5]$. Find the velocity when $t = 3$.

   $\mathbf{r}'(t) = \left(\cos t, -2t\sin\left(t^2\right), 1\right)$. The velocity when $t = 3$ is just $\mathbf{r}'(3) = (\cos 3, -6\sin(9), 1)$.

5. 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 $(\mathbf{r}(t) \times \mathbf{s}(t) \cdot \mathbf{p}(t))'$.

   From the product rules for the cross and dot product, this equals

   $$(\mathbf{r}(t) \times \mathbf{s}(t))' \cdot \mathbf{p}(t) + \mathbf{r}(t) \times \mathbf{s}(t) \cdot \mathbf{p}'(t) = \mathbf{r}'(t) \times \mathbf{s}(t) \cdot \mathbf{p}(t) + \mathbf{r}(t) \times \mathbf{s}'(t) \cdot \mathbf{p}(t) + \mathbf{r}(t) \times \mathbf{s}(t) \cdot \mathbf{p}'(t)$$

6. 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 $t \in (a,b)$.

   Do this by considering standard one variable calculus and on the components of $\mathbf{r}(t)$.

7. If $\mathbf{F}'(t) = \mathbf{f}(t)$ for all $t \in (a,b)$ and $\mathbf{F}$ is continuous on $[a,b]$, show $\int_a^b \mathbf{f}(t)\,dt = \mathbf{F}(b) - \mathbf{F}(a)$.

   Do this by considering standard one variable calculus and on the components of $\mathbf{r}(t)$.

8. Verify that if $\mathbf{\Omega} \times \mathbf{u} = \mathbf{0}$ for all $\mathbf{u}$, then $\mathbf{\Omega} = \mathbf{0}$.

   Geometrically this says that if $\mathbf{\Omega}$ is not equal to zero then it is parallel to every vector. Why does this make it obvious that $\mathbf{\Omega}$ must equal zero?

Newton's Laws Of Motion*

I assume you have seen basic mechanics as found in introductory physics course. However, if you need a review, the following section is offered. Read it if you need to. Otherwise, skip it. Calculus was invented to solve problems in physics and engineering, not to do cute geometry. The material which follows on physics of motion on a space curve will make more sense to you if you know Newton's laws.

Definition 15.0.17 *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[1] 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:

$$\mathbf{F} = m\mathbf{a} = m\mathbf{r}''(t) \tag{15.1}$$

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 15.0.18 *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* $\mathbf{r}(t)$.

[1] 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.

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

$$\mathbf{r}'(t) = \left(\frac{t^2}{4}, \frac{t - t^3/3}{2}, -e^{-t}\right) + \mathbf{c}$$

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 15.0.19 *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 15.0.20 *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 $\mathbf{r}(t) = (r_1(t), r_2(t))$.

$$2\,(r_1''(t), r_2''(t)) = (-.1)\,(r_1'(t), r_2'(t)) + (0, -64),\ (r_1(0), r_2(0)) = (0, 30000),$$
$$(r_1'(0), r_2'(0)) = (880, 0)$$

Therefore,

$$\begin{array}{c} 2r_1''(t) + (.1)\, r_1'(t) = 0 \\ 2r_2''(t) + (.1)\, r_2'(t) = -64 \\ r_1(0) = 0 \\ r_2(0) = 30000 \\ r_1'(0) = 880 \\ r_2'(0) = 0 \end{array} \tag{15.2}$$

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

$$r(t) = (c/k)\, t - \frac{1}{k} m e^{-\frac{k}{m}t}\left(v - \frac{c}{k}\right) + D \tag{15.3}$$

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

$$r(t) = (c/k)\, t - \frac{1}{k} m e^{-\frac{k}{m}t}\left(v - \frac{c}{k}\right) + \left(u + \frac{m}{k}\left(v - \frac{c}{k}\right)\right).$$

Now apply this to the system 15.2 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 15.3 in the same way to obtain the velocity,

$$\begin{aligned} r_1'(t) &= 880.0 \exp(-.0\,5t), \\ r_2'(t) &= -640.0 + 640.0 \exp(-.0\,5t). \end{aligned} \tag{15.4}$$

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 15.4 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.

15.0.6 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

$$dW = \mathbf{F}(t) \cdot \mathbf{v}(t)\, dt. \tag{15.5}$$

If no work has been done at time $t = 0$,then 15.5 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.

15.0.7 Impulse And Momentum*

Implulse

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 15.0.21 *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*

$$\text{Linear momentum } = m\mathbf{v}.$$

The notion of impulse and momentum are related in the following theorem.

Theorem 15.0.22 *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 $\mathbf{F}$.

$$\int_a^b \mathbf{F}(t)\,dt = \int_a^b m\frac{d\mathbf{v}}{dt}dt = m\mathbf{v}(b) - m\mathbf{v}(a) \tag{15.6}$$

15.0.8 Conservation Of Momentum*

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

$$\begin{aligned} m_B\mathbf{v}_B(b) - m_B\mathbf{v}_B(a) &= \int_a^b (\text{Force of } A \text{ on } B)\,dt \\ &= -\int_a^b (\text{Force of } B \text{ on } A)\,dt \\ &= -\left(m_A\mathbf{v}_A(b) - m_A\mathbf{v}_A(a)\right) \end{aligned}$$

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. This law is why rockets work. Think about it.

15.0.9 Exercises With Answers

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?

   Multiply both sides of the differential equation by e^{rt}. Then the left side becomes $\frac{d}{dt}\left(e^{rt}\mathbf{v}\right) = e^{rt}\mathbf{c}$. Now integrate both sides. This gives $e^{rt}\mathbf{v}(t) = \mathbf{C} + \frac{e^{rt}}{r}\mathbf{c}$. You finish the rest.

2. 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.

   This is done by using Newton's law. Thus $5\frac{d^2\mathbf{r}}{dt^2} = e^{-t}\mathbf{i} + \cos(t)\,\mathbf{j} + t^2\mathbf{k}$ and so $5\frac{d\mathbf{r}}{dt} = -e^{-t}\mathbf{i} + \sin(t)\,\mathbf{j} + \left(t^3/3\right)\mathbf{k} + \mathbf{C}$. Find the constant, $\mathbf{C}$ by using the given initial velocity. Next do another integration obtaining another constant vector which will be determined by using the given initial position of the object.

3. 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 $dW = \mathbf{F}(t) \cdot \mathbf{v}(t)\,dt$?

   Remember $|\mathbf{v}|^2 = \mathbf{v} \cdot \mathbf{v}$. Now use the product rule.

4. 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.

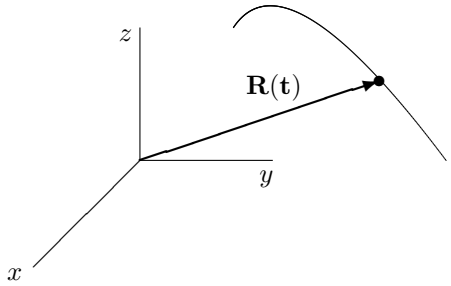
   $0 = \mathbf{F} \cdot \mathbf{v} = m\mathbf{v}' \cdot \mathbf{v}$. Explain why this is $\frac{d}{dt}\left(m\frac{1}{2}|\mathbf{v}|^2\right)$, the derivative of the kinetic energy.

Physics Of Curvilinear Motion 12 Oct.

16.0.10 The Acceleration In Terms Of The Unit Tangent And Normal

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 position vector of 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 16.0.23 *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,* $\mathbf{N}(t)$ *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.

Definition 16.0.24 *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 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.*

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 16.0.25 *For* $\mathbf{R}(t)$ *the position vector of a space curve, the acceleration is given by the formula*

$$\begin{aligned} \mathbf{a} &= \frac{d|\mathbf{v}|}{dt}\mathbf{T} + \kappa|\mathbf{v}|^2\mathbf{N} \\ &\equiv a_T\mathbf{T} + a_N\mathbf{N}. \end{aligned} \tag{16.1}$$

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 parameterization 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 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

$$\mathbf{R}(t) = \left(r\cos\left(\frac{v}{r}t\right), r\sin\left(\frac{v}{r}t\right)\right).$$

[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.

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,

$$\kappa = \left|\mathbf{T}'(t)\right| / v = \frac{1}{r}.$$

I hope it is not surprising that the curvature of a circle of radius r is $1/r$. 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 16.1 also yields an easy way to find the curvature. Take the cross product of both sides with $\mathbf{v}$, the velocity. Then

$$\begin{aligned}\mathbf{a}\times\mathbf{v} &= \frac{d\left|\mathbf{v}\right|}{dt}\mathbf{T}\times\mathbf{v} + \left|\mathbf{v}\right|^2\kappa\mathbf{N}\times\mathbf{v} \\ &= \frac{d\left|\mathbf{v}\right|}{dt}\mathbf{T}\times\mathbf{v} + \left|\mathbf{v}\right|^3\kappa\mathbf{N}\times\mathbf{T}\end{aligned}$$

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,

$$\left|\mathbf{a}\times\mathbf{v}\right| = \left|\mathbf{v}\right|^3\kappa$$

and so

$$\kappa = \frac{\left|\mathbf{a}\times\mathbf{v}\right|}{\left|\mathbf{v}\right|^3}. \tag{16.2}$$

Example 16.0.26 *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

$$\left|\mathbf{v}\right| = \sqrt{\sin^2(t) + 1 + 4t^2}.$$

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

$$\begin{aligned}\kappa &= \frac{\left|(-\cos t, 0, 2)\times(-\sin t, 1, 2t)\right|}{\left(\sqrt{\sin^2(t) + 1 + 4t^2}\right)^3} \\ &= \frac{\sqrt{4 + \left(-2\sin(t) + 2(\cos(t))\,t\right)^2 + \cos^2(t)}}{\left(\sqrt{\sin^2(t) + 1 + 4t^2}\right)^3}\end{aligned}$$

Then

$$a_N = \kappa\left|\mathbf{v}\right|^2$$

$$= \frac{\sqrt{4+\left(-2\sin(t)+2(\cos(t))\,t\right)^2+\cos^2(t)}}{\left(\sqrt{\sin^2(t)+1+4t^2}\right)^3}\left(\sin^2(t)+1+4t^2\right)$$
$$= \frac{\sqrt{4+\left(-2\sin(t)+2(\cos(t))\,t\right)^2+\cos^2(t)}}{\sqrt{\sin^2(t)+1+4t^2}}.$$

You can observe the formula $a_N^2+a_T^2=|\mathbf{a}|^2$ holds. Indeed $a_N^2+a_T^2=$

$$\left(\frac{\sqrt{4+\left(-2\sin(t)+2(\cos(t))\,t\right)^2+\cos^2(t)}}{\sqrt{\sin^2(t)+1+4t^2}}\right)^2+\left(\frac{\sin(t)\cos(t)+4t}{\sqrt{(2+4t^2-\cos^2 t)}}\right)^2$$
$$= \frac{4+\left(-2\sin t+2(\cos t)\,t\right)^2+\cos^2 t}{\sin^2 t+1+4t^2}+\frac{(\sin t\cos t+4t)^2}{2+4t^2-\cos^2 t}=\cos^2 t+4=|\mathbf{a}|^2$$

Some Simple Techniques

Recall the formula for acceleration is

$$\mathbf{a}=a_T\mathbf{T}+a_N\mathbf{N} \tag{16.3}$$

where $a_T=\frac{d|\mathbf{v}|}{dt}$ and $a_N=\kappa|\mathbf{v}|^2$. Of course one way to find a_T and a_N is to just find $|\mathbf{v}|$, $\frac{d|\mathbf{v}|}{dt}$ and κ and plug in. However, there is another way which might be easier. Take the dot product of both sides with $\mathbf{T}$. This gives,

$$\mathbf{a}\cdot\mathbf{T}=a_T\mathbf{T}\cdot\mathbf{T}+a_N\mathbf{N}\cdot\mathbf{T}=a_T.$$

Thus

$$\mathbf{a}=(\mathbf{a}\cdot\mathbf{T})\,\mathbf{T}+a_N\mathbf{N}$$

and so

$$\mathbf{a}-(\mathbf{a}\cdot\mathbf{T})\,\mathbf{T}=a_N\mathbf{N} \tag{16.4}$$

and taking norms of both sides,

$$|\mathbf{a}-(\mathbf{a}\cdot\mathbf{T})\,\mathbf{T}|=a_N.$$

Also from 16.4,

$$\frac{\mathbf{a}-(\mathbf{a}\cdot\mathbf{T})\,\mathbf{T}}{|\mathbf{a}-(\mathbf{a}\cdot\mathbf{T})\,\mathbf{T}|}=\frac{a_N\mathbf{N}}{a_N|\mathbf{N}|}=\mathbf{N}.$$

Also recall

$$\kappa=\frac{|\mathbf{a}\times\mathbf{v}|}{|\mathbf{v}|^3},\ a_T^2+a_N^2=|\mathbf{a}|^2$$

This is usually easier than computing $\mathbf{T}'/|\mathbf{T}'|$. To illustrate the use of these simple observations, consider the example worked above which was fairly messy. I will make it easier by selecting a value of t.

Example 16.0.27 *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 when* $t=0$. *Also find* $\mathbf{N}$ *at the point where* $t=0$.

First I need to find the velocity and acceleration. Thus

$$\mathbf{v} = (-\sin t, 1, 2t), \ \mathbf{a} = (-\cos t, 0, 2)$$

and consequently,

$$\mathbf{T} = \frac{(-\sin t, 1, 2t)}{\sqrt{\sin^2(t) + 1 + 4t^2}}.$$

When $t = 0$, this reduces to

$$\mathbf{v}(0) = (0, 1, 0), \ \mathbf{a} = (-1, 0, 2), \ |\mathbf{v}(0)| = 1, \ \mathbf{T} = (0, 1, 0),$$

and consequently,

$$\mathbf{T} = (0, 1, 0).$$

Then the tangential component of acceleration when $t = 0$ is

$$a_T = (-1, 0, 2) \cdot (0, 1, 0) = 0$$

Now $|\mathbf{a}|^2 = 5$ and so $a_N = \sqrt{5}$ because $a_T^2 + a_N^2 = |\mathbf{a}|^2$. Thus $\sqrt{5} = \kappa |\mathbf{v}(0)|^2 = \kappa \cdot 1 = \kappa$. Next lets find $\mathbf{N}$. From $\mathbf{a} = a_T\mathbf{T} + a_N\mathbf{N}$ it follows

$$(-1, 0, 2) = 0 \cdot \mathbf{T} + \sqrt{5}\mathbf{N}$$

and so

$$\mathbf{N} = \frac{1}{\sqrt{5}}(-1, 0, 2).$$

This was pretty easy.

Example 16.0.28 *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 16.0.29 *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 16.0.30 *Find a parameterization 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.

16.0.11 The Curvature Vector

The main item of interest for us is the scalar curvature defined above. Recall this was given by

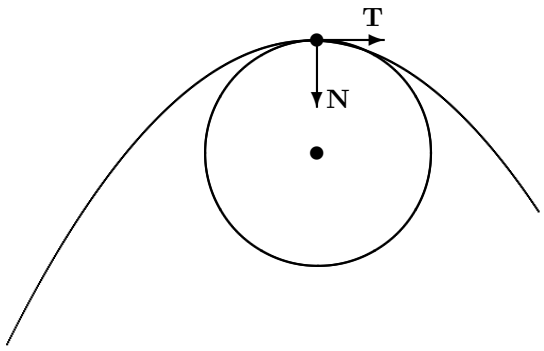
$$\kappa = \frac{|\mathbf{T}'|}{|\mathbf{v}|}.$$

The curvature vector is nothing more than

$$\boldsymbol{\kappa} \equiv \frac{\mathbf{T}'}{|\mathbf{v}|}.$$

16.0.12 The Circle Of Curvature*

In addition to the osculating plane, you can consider something called the circle of curvature. The idea is that near a point on the space curve, the space curve is like a circle. This circle has radius equal to $1/\kappa$, the radius of curvature, lies in the osculating plane, and its center is located by moving a distance of $1/\kappa$ (radiius of curvature) along the line determined by the point on the curve and the principle normal in the direction of the principle normal. It is an attempt to find the circle which best resembles the curve locally.



Here is an example to illustrate this fussy concept.

Example 16.0.31 *Consider the curve having a parameterization,* $(\cos(t), \sin(t), e^t)$. *Find the circle of curvature at the point where* $t = \pi/4$.

First find the curvature and the two vectors, $\mathbf{T}, \mathbf{N}$. The vector, $\mathbf{T}(t) = \frac{(-\sin t, \cos t, e^t)}{\sqrt{1+e^{2t}}}$. At the point of interest this is

$$\mathbf{T} = \frac{\left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, e^{\pi/4}\right)}{\sqrt{1+e^{\pi/2}}}$$

To find $\mathbf{N}$ next appears to be painful. Therefore, I will first find the acceleration and then use the formula for acceleration to dredge up $\mathbf{N}$.

$$\mathbf{a} = \left(-\cos t, -\sin t, e^t\right)$$

Thus the curvature is easy to find.

$$\begin{aligned}
\kappa &= \frac{|(-\cos t, -\sin t, e^t) \times (-\sin t, \cos t, e^t)|}{\left(\sqrt{1+e^{2t}}\right)^3} \\
&= \frac{\left|\left(-(\sin t)\, e^t - e^t \cos t, -(\sin t)\, e^t + e^t \cos t, -\cos^2 t - \sin^2 t\right)\right|}{\left(\sqrt{1+e^{2t}}\right)^3} \\
&= \frac{\left(2e^{2t}+1\right)^{1/2}}{\left(\sqrt{1+e^{2t}}\right)^3}
\end{aligned}$$

It follows that at the point of interest,

$$\left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, e^{\pi/4}\right) = \left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, e^{\pi/4}\right) + \frac{\left(2e^{\pi/2}+1\right)^{1/2}}{\left(\sqrt{1+e^{\pi/2}}\right)^3}\left(1+e^{\pi/2}\right)\mathbf{N}$$

From this, you can find the components of $\mathbf{N}$ without too much trouble.

$$-\frac{\sqrt{2}}{2} = -\frac{\sqrt{2}}{2} + \frac{\left(2e^{\pi/2}+1\right)^{1/2}}{\left(\sqrt{1+e^{\pi/2}}\right)^3}\left(1+e^{\pi/2}\right)N_1$$

and so $N_1 = 0$. Next,

$$-\frac{\sqrt{2}}{2} = \frac{\sqrt{2}}{2} + \frac{\left(2e^{\pi/2}+1\right)^{1/2}}{\left(\sqrt{1+e^{\pi/2}}\right)^3}\left(1+e^{\pi/2}\right)N_2$$

and so

$$N_2 = -\sqrt{2}\frac{\sqrt{\left(1+e^{\pi/2}\right)}}{\sqrt{\left(2e^{\pi/2}+1\right)}}.$$

Finally,

$$e^{\pi/4} = e^{\pi/4} + \frac{\left(2e^{\pi/2}+1\right)^{1/2}}{\left(\sqrt{1+e^{\pi/2}}\right)^3}\left(1+e^{\pi/2}\right)N_3$$

and so $N_3 = 0$ also.

From this you can find the location of the center of curvature. It is at the point

$$\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, e^{\pi/4}\right) + \left(\frac{\left(2e^{\pi/2}+1\right)^{1/2}}{\left(\sqrt{1+e^{\pi/2}}\right)^3}\right)^{-1}\left(0, -\sqrt{2}\frac{\sqrt{\left(1+e^{\pi/2}\right)}}{\sqrt{\left(2e^{\pi/2}+1\right)}}, 0\right).$$

Simplifying this yields

$$\left(\frac{1}{2}\sqrt{2},-\frac{1}{2}\sqrt{2}\frac{2e^{\frac{1}{2}\pi}+1+2e^{\pi}}{2e^{\frac{1}{2}\pi}+1},e^{\frac{1}{4}\pi}\right)$$

for the center of curvature. The circle of curvature is the circle in the osculating plane which has the above point as the center and radius equal to

$$\left(\frac{\left(2e^{\pi/2}+1\right)^{1/2}}{\left(\sqrt{1+e^{\pi/2}}\right)^{3}}\right)^{-1}=\frac{1}{\sqrt{\left(2e^{\frac{1}{2}\pi}+1\right)}}\left(\sqrt{\left(1+e^{\frac{1}{2}\pi}\right)}\right)^{3}.$$

If you try the same problem for $(\cos(t),\sin(t),t)$, you may find the computations much simpler.

One can of course go on and on fussing about geometrical dodads of this sort. The evolute is the locus of centers of curvatures. Imagine finding such a center of curvature as above for each t and considering the resulting curve. This is the evolute. Then if you really like to do this sort of thing, you could think about the evolute of the evolute. There are sure to be some wonderful conclusions hidden in this procedure.

The significant geometrical concepts are discussed in Section 16.1. These lead to the Serrat Frenet formulas which cause some people who like this sort of thing to wax ecstatic over their virtues. These formulas are indeed interesting, unlike the fussy stuff above about the circle of curvature. However, it is not required reading so skip it if you are not interested. It is a system of differential equations which completely describes the geometry of the space curve.

16.1 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 16.1.1 *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 16.1.2 *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 16.1.3 *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. 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 16.1.4 *(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 16.1.5 *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:

$$\begin{aligned}\frac{d\mathbf{T}}{ds} &= \frac{d\mathbf{T}}{dt}\frac{dt}{ds}\\ &= \frac{1}{a^2+b^2}\left((-a\cos t)\,\mathbf{i}+(-a\sin t)\,\mathbf{j}+0\mathbf{k}\right)\end{aligned}$$

and so

$$\mathbf{N}=\frac{d\mathbf{T}}{ds}\Big/\left|\frac{d\mathbf{T}}{ds}\right|=-\left((\cos t)\,\mathbf{i}+(\sin t)\,\mathbf{j}\right)$$

The binormal:

$$\begin{aligned}\mathbf{B} &= \frac{1}{\sqrt{a^2+b^2}}\begin{vmatrix}\mathbf{i} & \mathbf{j} & \mathbf{k}\\ -a\sin t & a\cos t & b\\ -\cos t & -\sin t & 0\end{vmatrix}\\ &= \frac{1}{\sqrt{a^2+b^2}}\left((b\sin t)\,\mathbf{i}-b\cos t\mathbf{j}+a\mathbf{k}\right)\end{aligned}$$

Now the curvature, $\kappa(t)=\left|\frac{d\mathbf{T}}{ds}\right|=\sqrt{\left(\frac{a\cos t}{a^2+b^2}\right)^2+\left(\frac{a\sin t}{a^2+b^2}\right)^2}=\frac{a}{a^2+b^2}$. 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,

$$\begin{aligned}\mathbf{B}'(s) &= \frac{1}{\sqrt{a^2+b^2}}\left((b\cos t)\,\mathbf{i}+(b\sin t)\,\mathbf{j}\right)\frac{dt}{ds}\\ &= \frac{1}{a^2+b^2}\left((b\cos t)\,\mathbf{i}+(b\sin t)\,\mathbf{j}\right)\\ &= \tau\left(-(\cos t)\,\mathbf{i}-(\sin t)\,\mathbf{j}\right)=\tau\mathbf{N}\end{aligned}$$

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 16.1.6 *Let* $\mathbf{R}(t)$ *be a space curve and denote by* $\mathbf{v}(t)$ *the velocity,* $\mathbf{v}(t)=\mathbf{R}'(t)$ *and let* $v(t)\equiv|\mathbf{v}(t)|$ *denote the speed and let* $\mathbf{a}(t)$ *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, $s=\int_0^t v(r)\,dr$ 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 $\mathbf{N}=\frac{\mathbf{T}'(t)}{\left|\mathbf{T}'\right|}$

16.2 Independence Of Parameterization*

This section is for those who want to really understand what is going on. If you are content, do not read this section. It may upset you. However, if you do decide to read it, you might learn something so there is some benefit for the anguish you might endure in the attempt.

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

$$\int_a^b\left|\mathbf{p}'(t)\right|\,dt$$

If some other parameterization were used to trace out C, would the same answer be obtained? The answer is yes. This is indeed fortunate because the length of a curve should only depend on the curve itself, not on some parameterization. To answer this question in a satisfactory manner requires some hard calculus. To answer it even more satisfactorily, you need to consider some very advanced mathematics involving something called Hausdorff measure.

16.2.1 Hard Calculus*

Definition 16.2.1 *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. The limit of a sequence, if it exists, is unique.

Theorem 16.2.2 *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 16.2.3 *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 16.2.4 *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 16.2.5 *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 16.2.6 *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 16.2.7 *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 16.2.8 *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 $x_t \equiv t x_1 + (1-t)\, x_2$ and $y_t \equiv t y_1 + (1-t)\, y_2$. Then $x_t < y_t$ for all $t \in [0,1]$ because

$$t x_1 \le t y_1 \text{ and } (1-t)\, x_2 \le (1-t)\, y_2$$

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 16.2.9 *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 16.2.10 *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 16.2.8 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 16.2.11 *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 in remembering formulas such as the one given in the theorem.

16.2.2 Independence Of Parameterization*

Here is the precise definition of what is meant by a smooth curve.

Definition 16.2.12 *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 $(x_1,\cdots,x_n)$.

Theorem 16.2.13 *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 16.2.8, 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 but easier. This proves the theorem.

Lemma 16.2.14 *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 16.2.8. 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 16.2.7 on Page 293.) 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 16.2.15 *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 $s \in I$,

$$h_i'(g_i(s)) = \frac{1}{g_i'(s)}. \tag{16.5}$$

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 $t \in J_1$,

$$\phi'(t) = h_i'(f_i(t)) f_i'(t) = h_i'(g_i(s)) f_i'(t) = \frac{f_i'(t)}{g_i'(\phi(t))} \tag{16.6}$$

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 16.2.13,

$$\begin{aligned} \int_{\phi(a+\delta)}^{\phi(b-\delta)} |\mathbf{g}'(s)| \, ds &= \int_{a+\delta}^{b-\delta} |\mathbf{g}'(\phi(t))| \left|\phi'(t)\right| dt \\ &= \int_{a+\delta}^{b-\delta} |\mathbf{f}'(t)| \, dt. \end{aligned}$$

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.

16.3 Product Rule For Matrices*

Another kind of multiplication is matrix multiplication. Here is the concept of the product rule extended to matrix multiplication.

Definition 16.3.1 *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 16.3.2 *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.

16.4 Moving Coordinate Systems*

Let $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ be a right handed[3] 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,

$$\mathbf{u}(t) = u_j \mathbf{e}_j(t) = u_j \mathbf{e}_j(t) \cdot \mathbf{e}_i(t_0)\, \mathbf{e}_i(t_0)$$

[3] Recall that right handed implies $\mathbf{i} \times \mathbf{j} = \mathbf{k}$.

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.

Lemma 16.4.1 *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 16.3.2 that

$$Q'(t) Q(t)^T + Q(t) Q'(t)^T = 0$$

and so

$$Q'(t) Q(t)^T = -\left(Q'(t) Q(t)^T\right)^T. \tag{16.7}$$

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 16.7 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,

$$\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t) = Q'(t) Q(t)^T \mathbf{u}(t) \tag{16.8}$$

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

$$\mathbf{\Omega}(t) \times \mathbf{u}(t) \equiv \begin{vmatrix} \mathbf{i}(t_0) & \mathbf{j}(t_0) & \mathbf{k}(t_0) \\ w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \end{vmatrix} \equiv$$

$$\mathbf{i}(t_0)(w_2u_3 - w_3u_2) + \mathbf{j}(t_0)(w_3u_1 - w_1u_3) + \mathbf{k}(t_0)(w_1u_2 - w_2u_1).$$

This proves the lemma and yields the existence part of the following theorem.

Theorem 16.4.2 *Let $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ be as described. Then there exists a unique vector $\mathbf{\Omega}(t)$ 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*

$$\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t).$$

Proof: It only remains to prove uniqueness. Suppose $\mathbf{\Omega}_1$ also works. Then $\mathbf{u}(t) = Q(t)\mathbf{u}$ and so $\mathbf{u}'(t) = Q'(t)\mathbf{u}$ and

$$Q'(t)\mathbf{u} = \mathbf{\Omega} \times Q(t)\mathbf{u} = \mathbf{\Omega}_1 \times Q(t)\mathbf{u}$$

for all $\mathbf{u}$. Therefore,

$$(\mathbf{\Omega} - \mathbf{\Omega}_1) \times Q(t)\mathbf{u} = \mathbf{0}$$

for all $\mathbf{u}$ and since $Q(t)$ is one to one and onto, this implies $(\mathbf{\Omega} - \mathbf{\Omega}_1) \times \mathbf{w} = \mathbf{0}$ for all $\mathbf{w}$ and thus $\mathbf{\Omega} - \mathbf{\Omega}_1 = \mathbf{0}$. This proves the theorem.

Definition 16.4.3 *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 16.4.4 *Let $B(t)$ be the set of points in three dimensions occupied by a rigid body. Then there exists a vector $\mathbf{\Omega}(t)$ such that whenever $\mathbf{u}(t)$ is fixed with the rigid body,*

$$\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t).$$

Part VII

Functions Of Many Variables

Outcomes

Functions of Several variables

A. Identify the domain and range of a function of several variables.

B. Represent a function of two variables by level curves or a function of three variables by level surfaces.

C. Identify the characteristics of a function from its graph or from a graph of its level curves (or level surfaces).

D. Represent combinations of multivariable functions algebraically.

Reading: Multivariable Calculus 2.1
Outcome Mapping:

A. 1

B. 2,7,8

C. 3,5,6

D. 9

Limits and Continuity

A. Describe a delta neighborhood of a point in 2- or 3-space.

B. Evaluate the limit of a function of several variables for a given approach or show that it does not exists.

C. Determine whether a function is continuous at a given point. Interpret the definition of continuity of a function of several variables graphically.

D. Determine whether a set in 2- or 3-space is open, closed or neither. Determine whether a set is compact.

E. Recall and apply the Extreme Value Theorem.

Reading: Multivariable Calculus 2.2
Outcome Mapping:

A. F1

B. 1,2,3

C. 8,F4

D. F3,11,12,13

E. F2

Partial Derivatives

A. Interpret the definition of a partial derivative of a function of two variables graphically.

B. Evaluate the partial derivatives of a function of several variables.

D. Evaluate the higher order partial derivatives of a function of several variables.

E. State the conditions under which mixed partial derivatives are equal.

F. Verify equations involving partial derivatives.

G. Evaluate the gradient of a function.

H. Prove identities involving the gradient.

Reading: Multivariable Calculus 2.3
Outcome Mapping:

A. G1,4,18

B. 3,5,6

D. 5,7

E. G3

F. 12,15

G. 9

H. 10

Functions Of Many Variables 16 Oct.

Quiz

1. Let $\mathbf{r}(t) = (\cos(t), \sin(t), 2t)$. Find $\mathbf{a}$,a_T, and a_N. Also find κ and write the acceleration as the sum of two terms, one in the direction of the unit tangent vector and the other in the direction of the principle normal. Find $\vec{\kappa}$, the curvature vector which is a completely useless concept.

2. Here is a matrix which happens to have -1 as an eigenvalue. Find the eigenspace corresponding to this eigenvalue.

$$\begin{pmatrix} 1 & 2 & 2 \\ -2 & -3 & -2 \\ 2 & 2 & 1 \end{pmatrix}$$

   Is the matrix defective or nondefective?

3. In Problem 2 find the determinant of the matrix.

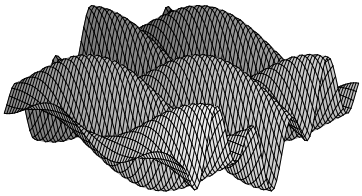
4. Raise the matrix of Problem 2 to the 15^{th} power exactly.

17.1 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.

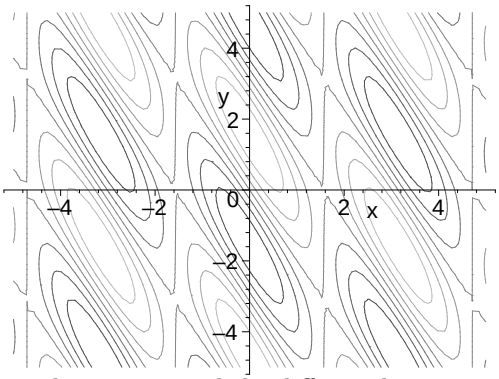
Let $f(x,y)$ denote a scalar valued function of two variables evaluated at the point (x,y). Its graph consists of the set of points, (x,y,z) such that $z = f(x,y)$. How does one go about depicting such a graph? The usual way is to fix one of the variables, say x and consider the function $z = f(x,y)$ 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.



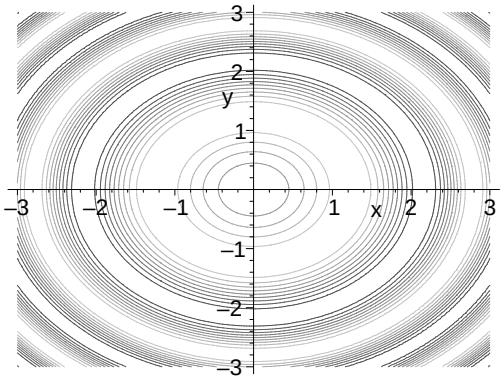
This is in two dimensions and the different lines in two dimensions correspond to points on the three dimensional graph which have the same z value. If you have looked at a weather map, these lines are called isotherms or isobars depending on whether the function involved is temperature or pressure. 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 17.1.1 *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.

17.2 The Domain Of A Function

As usual the domain of a function is either specified or if it is unspecified, it is the set of all points for which the function makes sense. If $\mathbf{f}$ is the name of the function its domain is denoted as $D(\mathbf{f})$.

Example 17.2.1 *Find the domain of the function,* $f(x,y) = \sqrt{1-(x^2+y^2)}$.

You need to have $1 \geq x^2+y^2$ and so the domain of this function is $\left\{(x,y) \in \mathbb{R}^2 : x^2+y^2 \leq 1\right\}$. This is just the inside of the unit circle centered at $(0,0)$. It also includes the edge of this unit circle.

Sometimes the domain is given to you in a very artificial way.

Example 17.2.2 *Let* $D(f) = \left\{(x,y) \in \mathbb{R}^2 : x^2+y^2 \leq 1\right\} \cup (3,7)$. *Let* $f(x,y) = x+2y$ *for* $(x,y) \in \left\{(x,y) \in \mathbb{R}^2 : x^2+y^2 \leq 1\right\}$ *and let* $f(3,7) = 33$.

In this case, the domain of the function is as given above and the function is given the definition just described.

Now remember from calculus of functions of one variable some of the things you did. One of the most important was to consider the derivative of a function. Recall the definition of the derivative, $f'(x)$.

$$\lim_{y \to x} \frac{f(y)-f(x)}{y-x} \equiv f'(x).$$

In order to write this definition you need to have f defined for all values of y near x, That is, you need to have f defined on an open interval containing x of the form $(x-\delta, x+\delta)$ for some $\delta > 0$. Otherwise, you can't consider $f(y)$. This is one reason for the importance of the concepts in the next section.

17.3 Open And Closed Sets

We are going to consider functions 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 $\mathbf{x}$. It is essential you understand a few kinds of sets.

Definition 17.3.1 *Let $\mathbf{x} \in \mathbb{R}^n$. Then $B(\mathbf{x}, r)$, called the ball centered at $\mathbf{x}$ having radius r is defined to be the set of all points of $\mathbb{R}^n, \mathbf{y}$ which have the property that these points are closer than r to $\mathbf{x}$. Thus $\mathbf{y} \in B(\mathbf{x}, r)$ means $|\mathbf{y} - \mathbf{x}| < r$. Written formally, this is*

$$B(\mathbf{x}, r) \equiv \{\mathbf{y} \in \mathbb{R}^n : |\mathbf{y} - \mathbf{x}| < r\}.$$

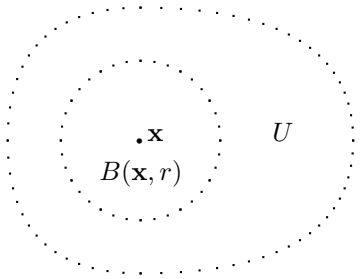
To say that $B(\mathbf{x}, r) \subseteq D(f)$ means that whenever $\mathbf{y}$ is closer to $\mathbf{x}$ than r, it follows $\mathbf{y} \in D(f)$. Now recall this is the sort of thing which you must start with, even in one dimension, to consider the concept of the derivative of a function. Therefore, it is not surprising that such an idea would be important in $\mathbb{R}^n$.

Definition 17.3.2 *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 17.3.3 *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 $\mathbb{R}^n \setminus S$. A **bounded** set is one which is contained in a large enough ball. In $\mathbb{R}^n$ a set which is both closed and bounded is **compact**.* [2]

To illustrate this definition, consider the following picture.

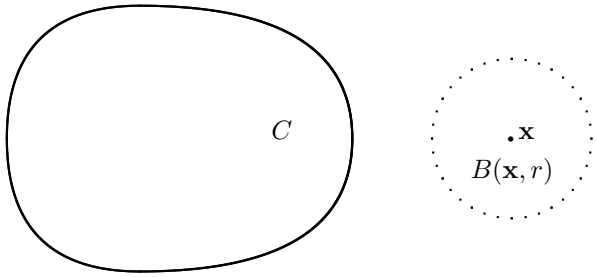


You see in this picture how the edges are dotted. This is because an open set, can't include the edges or the set would fail to be open. For example, consider what would happen if you picked a point out on the edge of U in the above picture. Every open ball centered at that point would have in it some 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.

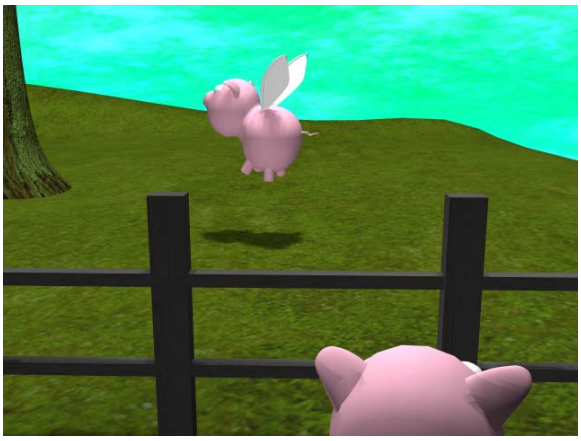
It is roughly the case that open sets don't have their skins while closed sets do. So why might it be important to consider closed sets? Remember from one variable calculus the theorem which says that a continuous function achieves its maximum and minimum on a closed interval. The closed interval contains its "skin", the end points of the interval.

[2] Actually the term compact has independent meaning and there is a theorem called the Heine Borel theorem which states that in $\mathbb{R}^n$ closed and bounded sets are compact. See the section on theory for more on this. This is not just useless jargon and gratuitous terminology.

Similar theorems will end up holding for functions of n variables. 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[3]. 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 17.3.4 *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.

[3]To a mathematician, the statement: 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.

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

$$\begin{aligned} |\mathbf{z}-\mathbf{x}| &= |\mathbf{z}-\mathbf{y}+\mathbf{y}-\mathbf{x}| \\ &\leq |\mathbf{z}-\mathbf{y}| + |\mathbf{y}-\mathbf{x}| \\ &< r_1 + |\mathbf{y}-\mathbf{x}| = r - |\mathbf{x}-\mathbf{y}| + |\mathbf{y}-\mathbf{x}| = r. \end{aligned}$$

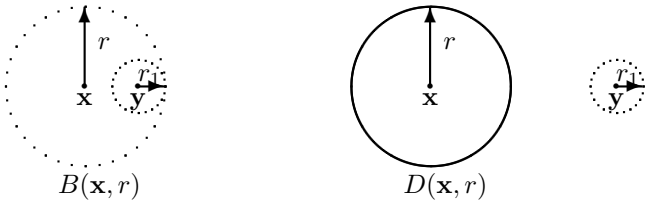
Note that if $r = 0$ then $B(\mathbf{x},r) = \emptyset$, the empty set. This is because if $\mathbf{y} \in \mathbb{R}^n$, $|\mathbf{x}-\mathbf{y}| \geq 0$ and so $\mathbf{y} \notin B(\mathbf{x},0)$. Since $\emptyset$ has no points in it, it must be open because every point in it, (There are none.) satisfies 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.

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 17.3.5 *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 17.3.6 *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 17.3.7 *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.

17.4 Continuous Functions

What was done in beginning calculus for scalar functions is generalized here to include the case of a vector valued function of possibly many variables. What follows is the **correct definition** of continuity. The one you are used to seeing in terms of the value of the function corresponding to the value of its limit is not correct in general. This one you are used to seeing is only correct if the point of the domain of the function is a limit point of the domain, discussed briefly later (Don't worry about it too much. Just use the correct definition and you will be fine.). It isn't a big deal for functions of one variables because you usually are dealing with functions defined on intervals and it happens that all the points are limit points. In multiple dimensions, however, the earlier definition is woefully inadequate and will lead you to profound confusion, confusion which is so severe you will have to relearn everything you thought you understood. I know this from bitter personal experience.

Definition 17.4.1 *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.

17.5 Sufficient Conditions For Continuity

The next theorem is a fundamental result which allows less worry about the ε δ definition of continuity.

Theorem 17.5.1 *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.*

The proof of this theorem is given later. 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 17.5.2 *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.

17.6 Properties Of Continuous Functions

Functions of many 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 17.6.1 *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 $\mathbf{x} \in C$,*

$$f(\mathbf{x}_1) \leq f(\mathbf{x}) \leq f(\mathbf{x}_2).$$

The above theorems are proved in an optional section.

Limits Of A Function 17-23 Oct.

Quiz

1. The position vector of an object is $\mathbf{r}(t) = \left(e^t, \sin(t), t^2 - 1\right)$. Find the unit tangent vector when $t = 0$.
2. Show that for $\mathbf{v}(t)$ a vector valued function $\frac{d}{dt}|\mathbf{v}(t)| = \frac{\mathbf{v}' \cdot \mathbf{v}}{|\mathbf{v}|}$. (Note that in the case where $\mathbf{v}$ is velocity, this implies $\mathbf{a} \cdot \mathbf{T} = \frac{d}{dt}|\mathbf{v}|$.)
3. Suppose $\mathbf{r}(t) = \left(t^2, \cos(t), \sin(t)\right)$. Find the curvature when $t = 0$.
4. Suppose $\mathbf{r}(t) = \left(2t^{1/2}, \frac{2}{3}t^{3/2}, \sqrt{2}t\right)$ for $t \in [1, 2]$. Find the length of this curve.
5. Find the matrix of the linear transformation which projects all vectors onto the line $y = x$.

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. It is a harder concept than the concept of continuity.

Definition 18.0.2 *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 18.0.3 *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 18.0.4 *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 18.0.5 *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 18.0.6 *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, $b \in \mathbb{R}$,*

$$\lim_{\mathbf{y}\to\mathbf{x}} (a\mathbf{f}(\mathbf{y}) + b\mathbf{g}(\mathbf{y})) = a\mathbf{L} + b\mathbf{K}, \tag{18.1}$$

$$\lim_{y\to x} \mathbf{f}\cdot\mathbf{g}(y) = \mathbf{L}\cdot\mathbf{K} \tag{18.2}$$

and if g is scalar valued with $\lim_{\mathbf{y}\to\mathbf{x}} g(\mathbf{y}) = K \neq 0$,

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y})\, g(\mathbf{y}) = \mathbf{L}K. \tag{18.3}$$

Also, if $\mathbf{h}$ is a continuous function defined near $\mathbf{L}$, then

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{h}\circ\mathbf{f}(\mathbf{y}) = \mathbf{h}(\mathbf{L}). \tag{18.4}$$

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 18.1 is left for you. It is like a corresponding theorem for continuous functions. Now 18.2is 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 $0 < |\mathbf{y}-\mathbf{x}| < \delta_1$,

$$|\mathbf{f}\cdot\mathbf{g}(\mathbf{y}) - \mathbf{L}\cdot\mathbf{K}| \leq (1 + |\mathbf{K}| + |\mathbf{L}|)\,[|\mathbf{g}(\mathbf{y}) - \mathbf{K}| + |\mathbf{f}(\mathbf{y}) - \mathbf{L}|]. \tag{18.5}$$

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 18.5 that

$$|\mathbf{f}\cdot\mathbf{g}(\mathbf{y}) - \mathbf{L}\cdot\mathbf{K}| < \varepsilon$$

and this proves 18.2.

The proof of 18.3 is left to you.

Consider 18.4. 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 $|\mathbf{b} - \mathbf{f}(\mathbf{y})| \leq r$.

The next theorem gives the correct relation between continuity and the limit.

Theorem 18.0.7 *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 18.0.8 *Suppose $\mathbf{f} : D(\mathbf{f}) \to \mathbb{R}^q$. Then for $\mathbf{x}$ a limit point of $D(\mathbf{f})$,*

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{L} \tag{18.6}$$

if and only if

$$\lim_{\mathbf{y}\to\mathbf{x}} f_k(\mathbf{y}) = L_k \tag{18.7}$$

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

$$\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) \times \mathbf{g}(\mathbf{y}) = \mathbf{L} \times \mathbf{K}. \tag{18.8}$$

Proof: Suppose 18.6. 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 18.7.

Now suppose 18.7 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 18.8. But from the first part of this theorem and the description of the cross product presented earlier in terms of the permutation symbol,

$$\begin{aligned} \lim_{\mathbf{y}\to\mathbf{x}} (\mathbf{f}(\mathbf{y}) \times \mathbf{g}(\mathbf{y}))_i &= \lim_{\mathbf{y}\to\mathbf{x}} \varepsilon_{ijk} f_j(\mathbf{y}) g_k(\mathbf{y}) \\ &= \varepsilon_{ijk} L_j K_k = (\mathbf{L} \times \mathbf{K})_i. \end{aligned}$$

If you did not read about the permutation symbol, you can simply write out the cross product and observe that the desired limit holds for each component. Therefore, from the first part of this theorem, this establishes 18.8. This completes the proof.

Example 18.0.9 *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 18.0.10 *Find* $\lim_{(x,y)\to(0,0)} \frac{xy}{x^2+y^2}$.

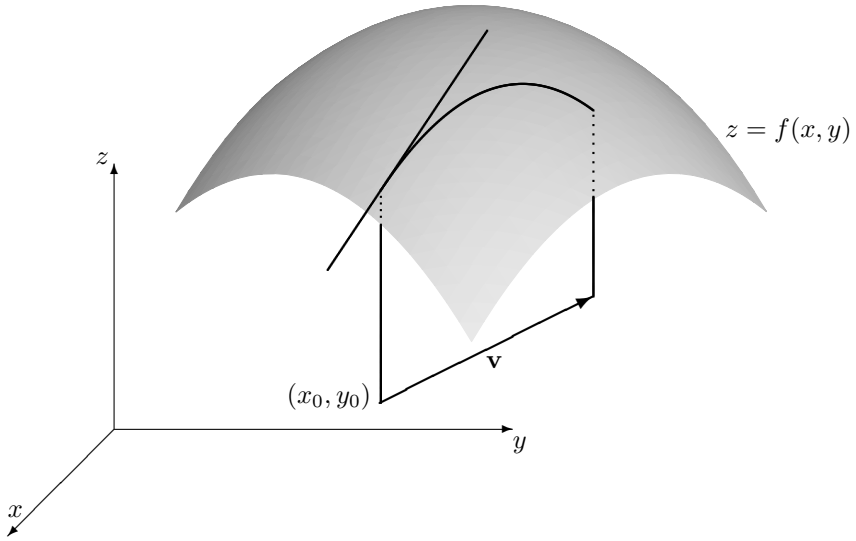
First of all observe the domain of the function is $\mathbb{R}^2 \setminus \{(0,0)\}$, every point in $\mathbb{R}^2$ except the origin. Therefore, $(0,0)$ is a limit point of the domain of the function so it might make sense to take a limit. However, just as in the case of a function of one variable, 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.

18.1 The Directional Derivative And Partial Derivatives

18.1.1 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 18.1.1 *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 18.1.2 *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[1].

18.1.2 Partial Derivatives

Quiz

1. Let $\mathbf{r}(t) = \left(t^2, \cosh(t) + t, \sin(t)\right)$. Find κ when $t = 0$. Remember κ is the curvature. Also find the normal and tangential components of acceleration and the oscullating plane at the point where $t = 0$.

2. Suppose $|\mathbf{r}(t)| = 33$ for all t. Show that $\mathbf{r}'(t) \cdot \mathbf{r}(t) = 0$. Does it follow that $\mathbf{r}'(t) = 0$?

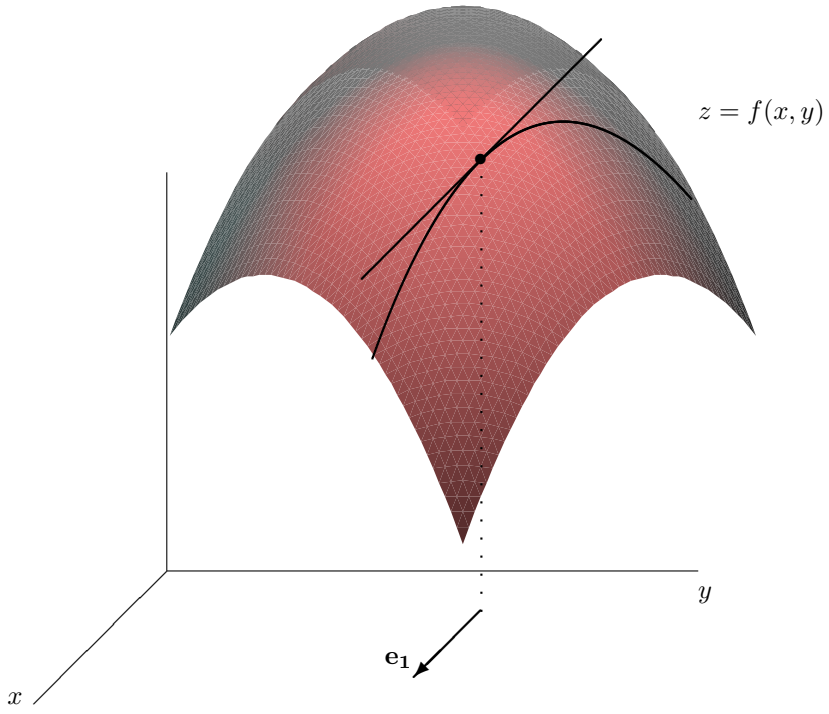
3. Suppose $\mathbf{r}(t) = \left(2t^{1/2}, \frac{2}{3}t^{3/2}, \sqrt{2}t\right)$ for $t \in [1, 2]$. Find the length of this curve.

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.

Thus in case of a function of two variables, the directional derivative in the direction $\mathbf{i} = \mathbf{e_1}$ is the slope of the indicated straight line in the following picture.



As in the case of a general directional derivative, you fix y and take the derivative of the function, $x \to f(x, y)$. More generally, even in situations which cannot be drawn, the definition of a partial derivative is as follows.

[1] Actually, there is a more general formulation of the notion of directional derivative known as the Gateaux derivative in which the length of $\mathbf{v}$ is not one but it is not considered here.

Definition 18.1.3 *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 18.1.4 *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.

Definition 18.1.5 *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 18.1.6 *In the situation of Definition 18.1.5 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 18.1.7 *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 18.1.8 *In the situation of Definition 18.1.5, 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 18.1.6,

$$D_{\mathbf{v}} f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v} = |\nabla f(\mathbf{x})|\,|\mathbf{v}| \cos\theta = |\nabla f(\mathbf{x})| \cos\theta$$

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 18.1.9 *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 18.1.10 *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 18.1.11 *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, $\mathbf{f}_x(x,y,z) = D_1\mathbf{f}(x,y,z) = \left(y^2, zy\cos(xy), z^3\right)^T$.

Example 18.1.12 *Let f, g be two functions defined on an open subset of $\mathbb{R}^3$ which have partial derivatives. Find a formula for $\nabla(fg)$.*

This equals

$$\begin{aligned}\left((fg)_x, (fg)_y, (fg)_z\right) &= (f_xg + fg_x, f_yg + fg_y, f_zg + fg_z)\\ &= g(f_x, f_y, f_z) + f(g_x, g_y, g_z) = g\nabla f + f\nabla g\end{aligned}$$

Example 18.1.13 *Let f, g be functions and a, b be scalars, you should verify that $\nabla(af + bg) = a\nabla f + b\nabla g$.*

Example 18.1.14 *Let* $h(x,y) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases}$. *Find $\frac{\partial h}{\partial x}$ and $\frac{\partial h}{\partial y}$.*

If $x > 0$ or if $x < 0$, both partial derivatives exist and equal 0. What of points like $(0,y)$? $\frac{\partial h}{\partial x}(0,y)$ does not exist but

$$\frac{\partial h}{\partial y}(0,y) \equiv \lim_{t\to 0}\frac{h(0,y+t) - h(0,0)}{t} = \lim_{t\to 0}\frac{1-1}{t} = 0.$$

Do not be afraid to use the definition of the partial derivatives. Sometimes it is the only way to find the partial derivative.

Example 18.1.15 *Let* $u(x,y) = \ln\left(x^2 + y^2\right)$. *Find $u_{xx} + u_{yy}$.*

First find u_x. This equals $2\frac{x}{x^2+y^2}$. Next find u_{xx}. This involves taking the partial derivative of u_x. Thus it equals

$$2\frac{y^2 - x^2}{\left(x^2+y^2\right)^2}$$

Similarly $u_{yy} = 2\frac{x^2-y^2}{(x^2+y^2)^2}$ and so $u_{xx} + u_{yy} = 0$. Of course this assumes $(x,y) \neq (0,0)$.

18.1.3 Mixed Partial Derivatives

Under certain conditions the **mixed partial derivatives** will always be equal. The simple condition is that if they exist and are continuous, then they are equal. This astonishing fact is due to Euler in 1734. For reasons I cannot understand, calculus books hardly ever include a proof of this important result. It is not all that hard. Here it is.

Theorem 18.1.16 *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

$$\Delta(s,t) \equiv \frac{1}{st}\{\overbrace{f(x+t,y+s)-f(x+t,y)}^{h(t)} - \overbrace{(f(x,y+s)-f(x,y))}^{h(0)}\}. \tag{18.9}$$

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,

$$\begin{aligned} \Delta(s,t) &= \frac{1}{st}(h(t)-h(0)) = \frac{1}{st}h'(\alpha t)\, t \\ &= \frac{1}{s}(f_x(x+\alpha t, y+s) - f_x(x+\alpha t, y)) \end{aligned}$$

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 18.9, $\Delta(s,t)$ 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 18.1.17 *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 18.1.18 *Let*

$$f(x,y) = \begin{cases} \frac{xy\left(x^2-y^2\right)}{x^2+y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$$

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 $(x,y) \neq (0,0)$,

$$f_x = y\frac{x^4 - y^4 + 4x^2y^2}{\left(x^2+y^2\right)^2},\ f_y = x\frac{x^4 - y^4 - 4x^2y^2}{\left(x^2+y^2\right)^2}$$

Now

$$\begin{aligned} f_{xy}(0,0) &\equiv \lim_{y\to 0}\frac{f_x(0,y) - f_x(0,0)}{y} \\ &= \lim_{y\to 0}\frac{-y^4}{\left(y^2\right)^2} = -1 \end{aligned}$$

while

$$\begin{aligned} f_{yx}(0,0) &\equiv \lim_{x\to 0}\frac{f_y(x,0) - f_y(0,0)}{x} \\ &= \lim_{x\to 0}\frac{x^4}{\left(x^2\right)^2} = 1 \end{aligned}$$

showing that although the mixed partial derivatives do exist at $(0,0)$, they are not equal there.

18.2 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. They are here to provide depth. If you want something more than a superficial knowledge, you should read this section. However, if you don't want to deal with challenging topics, don't read this stuff. Don't even look at it.

Theorem 18.2.1 *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

$$|a\mathbf{f}(\mathbf{x}) + b\mathbf{f}(\mathbf{x}) - (a\mathbf{g}(\mathbf{y}) + b\mathbf{g}(\mathbf{y}))|$$

$$\leq |a| |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| + |b| |\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})|$$
$$< |a| \left(\frac{\varepsilon}{2(|a| + |b| + 1)} \right) + |b| \left(\frac{\varepsilon}{2(|a| + |b| + 1)} \right) < \varepsilon.$$

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 $\mathbf{y}$,

$$|fg(\mathbf{x}) - fg(\mathbf{y})| \leq |f(\mathbf{x}) g(\mathbf{x}) - g(\mathbf{x}) f(\mathbf{y})| + |g(\mathbf{x}) f(\mathbf{y}) - f(\mathbf{y}) g(\mathbf{y})|$$

$$\leq |g(\mathbf{x})| |f(\mathbf{x}) - f(\mathbf{y})| + |f(\mathbf{y})| |g(\mathbf{x}) - g(\mathbf{y})|$$
$$\leq (1 + |g(\mathbf{x})| + |f(\mathbf{y})|) [|g(\mathbf{x}) - g(\mathbf{y})| + |f(\mathbf{x}) - f(\mathbf{y})|].$$

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

$$|fg(\mathbf{x}) - fg(\mathbf{y})| \leq$$

$$(1 + |g(\mathbf{x})| + |f(\mathbf{y})|) [|g(\mathbf{x}) - g(\mathbf{y})| + |f(\mathbf{x}) - f(\mathbf{y})|]$$
$$< (1 + |g(\mathbf{x})| + |f(\mathbf{y})|) \left(\frac{\varepsilon}{2(1 + |g(\mathbf{x})| + |f(\mathbf{y})|)} + \frac{\varepsilon}{2(1 + |g(\mathbf{x})| + |f(\mathbf{y})|)} \right) = \varepsilon.$$

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 $|\mathbf{x} - \mathbf{y}| < \min(\delta_0, \delta_1)$,

$$\left| \frac{f(\mathbf{x})}{g(\mathbf{x})} - \frac{f(\mathbf{y})}{g(\mathbf{y})} \right| = \left| \frac{f(\mathbf{x}) g(\mathbf{y}) - f(\mathbf{y}) g(\mathbf{x})}{g(\mathbf{x}) g(\mathbf{y})} \right|$$
$$\leq \frac{|f(\mathbf{x}) g(\mathbf{y}) - f(\mathbf{y}) g(\mathbf{x})|}{\left(\frac{|g(\mathbf{x})|^2}{2} \right)}$$
$$= \frac{2|f(\mathbf{x}) g(\mathbf{y}) - f(\mathbf{y}) g(\mathbf{x})|}{|g(\mathbf{x})|^2}$$

$$\begin{aligned}
&\leq \frac{2}{|g(\mathbf{x})|^2}\left[|f(\mathbf{x})\,g(\mathbf{y}) - f(\mathbf{y})\,g(\mathbf{y}) + f(\mathbf{y})\,g(\mathbf{y}) - f(\mathbf{y})\,g(\mathbf{x})|\right]\\
&\leq \frac{2}{|g(\mathbf{x})|^2}\left[|g(\mathbf{y})|\,|f(\mathbf{x}) - f(\mathbf{y})| + |f(\mathbf{y})|\,|g(\mathbf{y}) - g(\mathbf{x})|\right]\\
&\leq \frac{2}{|g(\mathbf{x})|^2}\left[\frac{3}{2}|\mathbf{g}(\mathbf{x})|\,|f(\mathbf{x}) - f(\mathbf{y})| + (1 + |f(\mathbf{x})|)\,|g(\mathbf{y}) - g(\mathbf{x})|\right]\\
&\leq \frac{2}{|g(\mathbf{x})|^2}\left(1 + 2|f(\mathbf{x})| + 2|g(\mathbf{x})|\right)\left[|f(\mathbf{x}) - f(\mathbf{y})| + |g(\mathbf{y}) - g(\mathbf{x})|\right]\\
&\equiv M\left[|f(\mathbf{x}) - f(\mathbf{y})| + |g(\mathbf{y}) - g(\mathbf{x})|\right]
\end{aligned}$$

where

$$M \equiv \frac{2}{|g(\mathbf{x})|^2}\left(1 + 2|f(\mathbf{x})| + 2|g(\mathbf{x})|\right)$$

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

$$\begin{aligned}
|f_k(\mathbf{x}) - f_k(\mathbf{y})| &\leq |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})|\\
&\equiv \left(\sum_{i=1}^{q}|f_i(\mathbf{x}) - f_i(\mathbf{y})|^2\right)^{1/2}\\
&\leq \sum_{i=1}^{q}|f_i(\mathbf{x}) - f_i(\mathbf{y})|. \qquad (18.10)
\end{aligned}$$

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 18.10 implies

$$\begin{aligned} |\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| &\leq \sum_{i=1}^{q} |f_i(\mathbf{x}) - f_i(\mathbf{y})| \\ &< \sum_{i=1}^{q} \frac{\varepsilon}{q} = \varepsilon. \end{aligned}$$

This proves part 4.)

To verify part 5.), let $\varepsilon > 0$ be given and let $\delta = \varepsilon$. Then if $|\mathbf{x}-\mathbf{y}| < \delta$, the triangle inequality implies

$$\begin{aligned} |f(\mathbf{x}) - f(\mathbf{y})| &= ||\mathbf{x}| - |\mathbf{y}|| \\ &\leq |\mathbf{x}-\mathbf{y}| < \delta = \varepsilon. \end{aligned}$$

This proves part 5.) and completes the proof of the theorem.

18.2.1 The Nested Interval Lemma*

Here is a multidimensional version of the nested interval lemma.

Lemma 18.2.2 *Let* $I_k = \prod_{i=1}^{p} \left[a_i^k, b_i^k\right] \equiv \left\{\mathbf{x} \in \mathbb{R}^p : x_i \in \left[a_i^k, b_i^k\right]\right\}$ *and suppose that for all* $k = 1, 2, \cdots$,

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

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

Proof: Since $I_k \supseteq I_{k+1}$, it follows that for each $i = 1, \cdots, p$, $\left[a_i^k, b_i^k\right] \supseteq \left[a_i^{k+1}, b_i^{k+1}\right]$. This implies that for each i,

$$a_i^k \leq a_i^{k+1},\ b_i^k \geq b_i^{k+1}. \tag{18.11}$$

Consequently, if $k \leq l$,

$$a_i^l \leq b_i^l \leq b_i^k. \tag{18.12}$$

Now define

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

By the first inequality in 18.11,

$$c_i = \sup\left\{a_i^l : l = k, k+1, \cdots\right\} \tag{18.13}$$

for each $k = 1, 2 \cdots$. Therefore, picking any k,18.12 shows that b_i^k is an upper bound for the set, $\left\{a_i^l : l = k, k+1, \cdots\right\}$ and so it is at least as large as the least upper bound of this set which is the definition of c_i given in 18.13. Thus, for each i and each k,

$$a_i^k \leq c_i \leq b_i^k.$$

Defining $\mathbf{c} \equiv (c_1, \cdots, c_p)$, $\mathbf{c} \in I_k$ for all k. This proves the lemma.

If you don't like the proof,you could prove the lemma for the one variable case first and then do the following.

Lemma 18.2.3 *Let* $I_k = \prod_{i=1}^p \left[a_i^k, b_i^k\right] \equiv \left\{\mathbf{x} \in \mathbb{R}^p : x_i \in \left[a_i^k, b_i^k\right]\right\}$ *and suppose that for all* $k = 1, 2, \cdots,$

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

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

Proof: For each $i = 1, \cdots, p$, $\left[a_i^k, b_i^k\right] \supseteq \left[a_i^{k+1}, b_i^{k+1}\right]$ and so by the nested interval theorem for one dimensional problems, there exists a point $c_i \in \left[a_i^k, b_i^k\right]$ for all k. Then letting $\mathbf{c} \equiv (c_1, \cdots, c_p)$ it follows $\mathbf{c} \in I_k$ for all k. This proves the lemma.

18.2.2 The Extreme Value Theorem*

Definition 18.2.4 *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 18.2.5 *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 18.2.6 *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 18.2.5, 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 18.2.5. 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.

18.2.3 Sequences And Completeness*

Definition 18.2.7 *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 18.2.8 *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 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 18.2.9 *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 18.2.10 *Let* $\mathbf{a}_k = (k, \sin(k))$. *Thus this sequence has values in* $\mathbb{R}^2$.

Definition 18.2.11 *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

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

Definition 18.2.12 *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 18.2.13 *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 18.2.14 $\{\mathbf{a}_n\}$ *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 18.2.15 *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 18.2.16 *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 18.2.17 *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 18.2.18 *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,

$$|x_i - y_i| \leq 2^{-1} |b_i - a_i|.$$

$\text{diam}(I_0) = \left(\sum_{i=1}^p |b_i - a_i|^2\right)^{1/2}$,

$$|\mathbf{x}-\mathbf{y}| = \left(\sum_{i=1}^p |x_i - y_i|^2\right)^{1/2}$$
$$\leq 2^{-1}\left(\sum_{i=1}^p |b_i - a_i|^2\right)^{1/2} \equiv 2^{-1}\,\text{diam}(I_0).$$

In particular, since $\mathbf{d} \equiv (d_1, \cdots, d_p)$ and $\mathbf{c} \equiv (c_1, \cdots, c_p)$ are two such points,

$$D_1 \equiv \left(\sum_{i=1}^p |d_i - c_i|^2\right)^{1/2} \leq 2^{-1}\,\text{diam}(I_0)$$

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 $|\mathbf{x}-\mathbf{y}| \leq 2^{-k}\,\text{diam}(I_0)$, 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 $2^{-k}\,\text{diam}(I_0)$ and so $\lim_{k\to\infty} \mathbf{a}_{n_k} = \mathbf{c} \in I_0$. This proves the theorem.

Theorem 18.2.19 *Every Cauchy sequence in $\mathbb{R}^p$ converges.*

Proof: Let $\{\mathbf{a}_k\}$ be a Cauchy sequence. By Theorem 18.2.15 there is some interval, $\prod_{i=1}^p [a_i, b_i]$ containing all the terms of $\{\mathbf{a}_k\}$. Therefore, by Theorem 18.2.18 a subsequence converges to a point of this interval. By Theorem 18.2.16 the original sequence converges. This proves the theorem.

18.2.4 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 18.2.20 *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.

Part VIII

Differentiability

Outcomes

Differentiability and the Chain Rule

A. Define differentiability for a function of several variables.

B. Evaluate partial derivatives from the definition. Describe the relationship between the derivative of a multivariable function and its partial derivatives.

C. Describe the relationship between the existence of partial derivatives and the existence of a derivative for a function of several variables.

D. Apply the chain rule to evaluate derivatives.

E. Solve related rates problems using the chain rule.

Reading: Multivariable Calculus 2.4
Outcome Mapping:

A. H1

B. H1,1

C. G2

D. 3,6

E. 4,7,8

Directional Derivatives

A. Give a graphical interpretation of the gradient.

B. Evaluate the directional derivative of a function.

C. Give a graphical interpretation of directional derivative.

D. Prove that a differential function f increases most rapidly in the direction of the gradient (the rate of change is then $\|f(\vec{x})\|$) and it decreases most rapidly in the opposite direction (the rate of change is then $-\|f(\vec{x})\|$).

E. Find the path of a heat seeking or a heat repelling particle.

Reading: Multivariable Calculus 2.5
Outcome Mapping:

A. 1,11,H2

B. 4,6

C. 2,3,H3

D. H4

E. 5,7,8

Normal Vectors and Tangent Planes

A. Interpret the gradient of a function as a normal to a level curve or a level surface.

B. Find the normal line and tangent plane to a smooth surface at a given point.

C. Find the angles between curves and surfaces.

Reading: Multivariable Calculus 2.6
Outcome Mapping:

A. 1,3,4

B. 9,11

C. 14,15,16,17

Extrema of Functions of Several Variables

A. Identify local extreme values graphically.

B. Determine the local extreme values and saddle points of a function of two variables. When possible, apply the second partial derivatives test.

C. Identify the extreme values of a function defined on a closed and bounded region.

D. Solve word problems involving maximum and minimum values.

Reading: Multivariable Calculus 2.7
Outcome Mapping:

A. 1,2

B. 3

C. 4

D. 6,9,12

Constrained Extrema

A. Graphically interpret the method of Lagrange.

B. Determine the extreme values of a function subject to side constraints by applying the method of Lagrange.

C. Apply the method of Lagrange to solve word problems.

Reading: Multivariable Calculus 2.9
Outcome Mapping:

A. 1,2

B. 3

C. 4,8,14

Differentiability 24-26 Oct.

19.1 The Definition Of Differentiability

Quiz

1. Let $f(x,y) = x^2y + \sin(xy)$. Find $\nabla f(x,y)$.
2. Let $f(x,y) = x^2y + \sin(xy)$. Find $D_{\mathbf{v}} f(1,1)$ where $\mathbf{v}$ is in the direction of $(1,2)$.
3. Let $f(x,y) = x^2y + \sin(xy)$. Find the largest value of $D_{\mathbf{v}} f(1,2)$ for all $\mathbf{v}$. That is, find the largest directional derivative of this function.

First remember what it means for a function of one variable to be differentiable.

$$f'(x) \equiv \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}$$

Another way to say this is contained in the following observation.

Observation 19.1.1 *Suppose a function, f of one variable has a derivative at x. Then*

$$\lim_{h\to 0} \frac{|f(x+h) - f(x) - f'(x)\,h|}{|h|} = 0.$$

For a function of n variables, there is a similar definition of what it means for a function to be differentiable.

Definition 19.1.2 *Let U be an open set in $\mathbb{R}^n$ and suppose $f : U \to \mathbb{R}$ is a function. Then f is differentiable at $\mathbf{x} \in U$ if for $\mathbf{v} = (v_1, \cdots, v_n)$*

$$\lim_{|\mathbf{v}|\to 0} \frac{\left| f(\mathbf{x}+\mathbf{v}) - f(\mathbf{x}) - \sum_{k=1}^n \frac{\partial f}{\partial x_k}(\mathbf{x})\, v_k \right|}{|\mathbf{v}|} = 0.$$

Definition 19.1.3 *A function of a vector, $\mathbf{v}$ is called $\mathbf{o}(\mathbf{v})$ if*

$$\lim_{|\mathbf{v}|\to 0} \frac{\mathbf{o}(\mathbf{v})}{|\mathbf{v}|} = \mathbf{0}. \tag{19.1}$$

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

$$\mathbf{o}(\mathbf{v}) = \mathbf{o}(\mathbf{v}) + \mathbf{o}(\mathbf{v}), \mathbf{o}(\mathbf{v}) = 45\mathbf{o}(\mathbf{v}), \mathbf{o}(\mathbf{v}) = \mathbf{o}(\mathbf{v}) - \mathbf{o}(\mathbf{v}), \textit{etc.}$$

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 obsession over things which are not relevant and should be ignored.

Definition 19.1.2 is then equivalent to the following very simple statement.

Definition 19.1.4 *Let U be an open set in $\mathbb{R}^n$ and suppose $f : U \to \mathbb{R}$ is a function. Then f is differentiable at $\mathbf{x} \in U$ if*

$$f(\mathbf{x}+\mathbf{v}) - f(\mathbf{x}) - \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\mathbf{x})\, v_k = o(\mathbf{v}).$$

The first definition says nothing more than $f(\mathbf{x}+\mathbf{v}) - f(\mathbf{x}) - \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\mathbf{x})\, v_k = o(\mathbf{v})$ because it says

$$\lim_{|\mathbf{v}|\to 0} \frac{\left| f(\mathbf{x}+\mathbf{v}) - f(\mathbf{x}) - \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\mathbf{x})\, v_k \right|}{|\mathbf{v}|} = 0.$$

The following is fundamental.

Proposition 19.1.5 *If f is differentiable at $\mathbf{x}$, then f is continuous at $\mathbf{x}$.*

Proof: From the definition of differentiability,

$$|f(\mathbf{x}+\mathbf{v}) - f(\mathbf{x})| \le \left| \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\mathbf{x})\, v_k + o(\mathbf{v}) \right|$$

Let $\varepsilon > 0$ be given. Then clearly if $|\mathbf{v}|$ is sufficiently small, the right side of the above is less than ε. Thus the function is continuous at $\mathbf{x}$.

So which functions are differentiable? Are there simple ways to look at a function and say that it is clearly differentiable? Existence of partial derivatives is needed in order to even write the above expression but it turns out this is not enough. Here is a simple example.

Example 19.1.6 *Let*

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$$

Then

$$f_x(0,0) \equiv \lim_{h\to 0} \frac{f(h,0) - f(0,0)}{h} = \lim_{h\to 0} \frac{0}{h} = 0$$

Also

$$f_y(0,0) \equiv \lim_{h\to 0} \frac{f(0,h) - f(0,0)}{h} = \lim_{h\to 0} \frac{0}{h} = 0$$

so both partial derivatives exist. However, the function is not even continuous at $(0,0)$. This is because it equals zero on the entire y axis but along the line, $y = x$ the function equals $1/2$. By Proposition 19.1.5 it cannot be differentiable.

19.2 C^1 Functions And Differentiability

It turns out that if the partial derivatives are continuous then the function is differentiable. I will show this next. First, remember the Cauchy Schwarz inequality, which I will list here for convenience.

$$\left|\sum_{i=1}^{n} a_i b_i\right| \leq \left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \left(\sum_{i=1}^{n} b_i^2\right)^{1/2}.$$

Theorem 19.2.1 *Suppose $f : U \to \mathbb{R}$ where U is an open set. Suppose also that all partial derivatives of f exist on U and are continuous. Then f is differentiable at every point of U.*

Proof: If you fix all the variables but one, you can apply the fundamental theorem of calculus as follows.

$$f(\mathbf{x}+v_k\mathbf{e}_k) - f(\mathbf{x}) = \int_0^1 \frac{\partial f}{\partial x_k}(\mathbf{x}+tv_k\mathbf{e}_k)\, v_k dt. \tag{19.2}$$

Here is why. Let $h(t) = f(\mathbf{x}+tv_k\mathbf{e}_k)$. Then

$$\frac{h(t+\Delta t)-h(t)}{\Delta t} = \frac{f(\mathbf{x}+tv_k\mathbf{e}_k+\Delta tv_k\mathbf{e}_k)-f(\mathbf{x}+tv_k\mathbf{e}_k)}{\Delta tv_k}v_k$$

and so, taking the limit as $\Delta t \to 0$ yields

$$h'(t) = \frac{\partial f}{\partial x_k}(\mathbf{x}+tv_k\mathbf{e}_k)\, v_k$$

Therefore,

$$f(\mathbf{x}+v_k\mathbf{e}_k) - f(\mathbf{x}) = h(1)-h(0) = \int_0^1 h'(t)\, dt = \int_0^1 \frac{\partial f}{\partial x_k}(\mathbf{x}+tv_k\mathbf{e}_k)\, v_k dt.$$

Now I will use this observation to prove the theorem. Let $\mathbf{v} = (v_1, \cdots, v_n)$ with $|\mathbf{v}|$ sufficiently small. Thus $\mathbf{v} = \sum_{k=1}^{n} v_k\mathbf{e}_k$. For the purposes of this argument, define

$$\sum_{k=n+1}^{n} v_k\mathbf{e}_k \equiv \mathbf{0}.$$

Then with this convention, and using 19.2,

$$\begin{aligned}
f(\mathbf{x}+\mathbf{v})-f(\mathbf{x}) &= \sum_{i=1}^{n}\left(f\left(\mathbf{x}+\sum_{k=i}^{n}v_k\mathbf{e}_k\right)-f\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k\right)\right)\\
&= \sum_{i=1}^{n}\int_0^1\frac{\partial f}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)v_i\,dt\\
&= \sum_{i=1}^{n}\int_0^1\left(\frac{\partial f}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)v_i-\frac{\partial f}{\partial x_i}(\mathbf{x})\,v_i\right)dt\\
&\quad+\sum_{i=1}^{n}\int_0^1\frac{\partial f}{\partial x_i}(\mathbf{x})\,v_i\,dt\\
&= \sum_{i=1}^{n}\frac{\partial f}{\partial x_i}(\mathbf{x})\,v_i+\int_0^1\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)-\frac{\partial f}{\partial x_i}(\mathbf{x})\right)v_i\,dt\\
&= \sum_{i=1}^{n}\frac{\partial f}{\partial x_i}(\mathbf{x})\,v_i+o(\mathbf{v})
\end{aligned}$$

and this shows f is differentiable at $\mathbf{x}$ because it satisfies the conditions of Definition 19.1.4. Some explanation of the step to the last line is in order. The messy thing at the end is $o(\mathbf{v})$ because of the continuity of the partial derivatives. In fact, from the Cauchy Schwarz inequality,

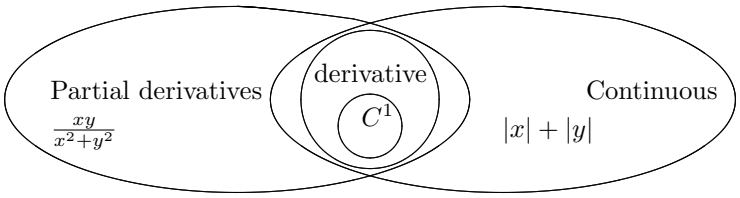
$$\begin{aligned}
&\int_0^1\sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)-\frac{\partial f}{\partial x_i}(\mathbf{x})\right)v_i\,dt\\
\leq\ &\int_0^1\left(\sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)-\frac{\partial f}{\partial x_i}(\mathbf{x})\right|^2\right)^{1/2}dt\left(\sum_{i=1}^{n}v_i^2\right)^{1/2}\\
=\ &\int_0^1\left(\sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)-\frac{\partial f}{\partial x_i}(\mathbf{x})\right|^2\right)^{1/2}dt\,|\mathbf{v}|
\end{aligned}$$

Thus, dividing by $|\mathbf{v}|$ and taking a limit as $|\mathbf{v}|\to 0$, the quotient is nothing but

$$\int_0^1\left(\sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)-\frac{\partial f}{\partial x_i}(\mathbf{x})\right|^2\right)^{1/2}dt$$

which converges to 0 due to continuity of the partial derivatives of f. This proves the theorem.

To help you keep the various terms straight, here is a pretty diagram.



You might ask whether there are examples of functions which are differentiable but not C^1. Of course there are. There are easy examples of this even for functions of one variable. Here is one.

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x} \text{ if } x \neq 0 \\ 0 \text{ if } x = 0 \end{cases}.$$

You should show that $f'(0) = 0$ but the derivative is $2x \sin \frac{1}{x} - \cos \frac{1}{x}$ for $x \neq 0$ and this function fails to even have a limit as $x \to 0$. This is a great test question. You ask for $f'(0)$ and it is really easy if you use the definition. However, people usually find $f'(x)$ and then try to plug in $x = 0$. This is doomed to failure and makes the question very easy to grade.

19.3 The Directional Derivative

Here I will prove the formula for the directional derivative presented earlier. Recall that for $\mathbf{v}$ a unit vector, ($|\mathbf{v}| = 1$)

$$D_{\mathbf{v}}(f)(\mathbf{x}) \equiv \lim_{t \to 0} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t}.$$

Theorem 19.3.1 *Suppose f is differentiable at $\mathbf{x}$. Then $D_{\mathbf{v}}(f)(\mathbf{x})$ exists and is given by*

$$D_{\mathbf{v}}(f)(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v}$$

Proof: By differentiability of f at $\mathbf{x}$,

$$\begin{aligned} \frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t} &= \frac{1}{t}\left(\sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\mathbf{x})\, t v_k + o(t\mathbf{v})\right) \\ &= \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\mathbf{x})\, v_k + \frac{o(t\mathbf{v})}{|t\mathbf{v}|} \end{aligned}$$

Taking the limit as $t \to 0$,

$$D_{\mathbf{v}} f(\mathbf{x}) = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(\mathbf{x})\, v_k \equiv \nabla f(\mathbf{x}) \cdot \mathbf{v}.$$

This proves the theorem.

What is the direction in which the largest directional derivative results? You want to maximize $\nabla f(\mathbf{x}) \cdot \mathbf{v} = |\nabla f(\mathbf{x})|\,|\mathbf{v}| \cos\theta$ where θ is the included angle between $\mathbf{v}$ and $\nabla f(\mathbf{x})$.

Clearly this occurs when $\theta = 0$. Therefore, the largest value of the directional derivative is when $\mathbf{v} = \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$. The value of the directional derivative in this direction, is

$$\nabla f(\mathbf{x}) \cdot \nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})| = |\nabla f(\mathbf{x})|.$$

Similarly, the smallest value for the directional derivative occurs when $\mathbf{v} = -\nabla f(\mathbf{x}) / |\nabla f(\mathbf{x})|$ because this corresponds to $\theta = \pi$ and $\cos\theta = -1$. The smallest value of the directional derivative is then $-|\nabla f(\mathbf{x})|$.

19.3.1 Separable Differential Equations*

If you do not know how to solve simple differential equations, read this section. Otherwise skip it.

Differential equations are just equations which involve an unknown function and some of its derivatives. For example, a differential equation is

$$y' = 1 + y^2. \tag{19.3}$$

You might check and see that a solution to this differential equation is $y = \tan x$.

Here is another easier differential equation.

$$y' = x^2.$$

A solution to this one is of the form

$$y = \frac{x^3}{3} + C$$

where C is any constant. In general, you are familiar with differential equations of the form

$$y' = f(x).$$

The problem is just to find an antiderivative of the given function. However, equations like the one in 19.3 are not so obvious. It turns out there are many recipes for finding solutions to differential equations of various sorts. One of the easiest kinds of differential equations to solve are those which are **separable.**

Separable differential equations are those which can be written in the form

$$\frac{dy}{dx} = \frac{f(x)}{g(y)}.$$

The equation in 19.3 is an example of a separable differential equation. Just let $f(x) = 1$ and $g(y) = \frac{1}{1+y^2}$.

The reason these are called separable is that if you formally cross multiply,

$$g(y)\,dy = f(x)\,dx$$

and the variables are "separated". Here is how you solve these. Find $G'(y) = g(y)$ and $F'(x) = f(x)$. That is, pick $G \in \int g(y)\,dy$ and $F \in \int f(x)\,dx$. Suppose $F(x) - G(y) = c$ specifies y as a differentiable function of x, then $x \to y(x)$ solves the separable differential equation because by the chain rule,

$$F'(x) - G'(y)\,y' = f(x) - g(y)\,y'$$

and so

$$f(x) = g(y)\,y'$$

so

$$y' = \frac{f(x)}{g(y)}.$$

This is why the solutions are given in the form

$$F(x) - G(y) = c$$

where c is a constant, or equivalently

$$G(y) - F(x) = c$$

where c is a constant.

Example 19.3.2 *Find the solutions to the differential equation,*

$$y' = \frac{x^2}{y}$$

which satisfies the initial condition, $y(0) = 1$. Since there is a differential equation along with an initial condition, this is called an initial value problem.

To solve this you separate the variables and write

$$ydy = x^2 dx$$

and then from the above discussion,

$$\frac{y^2}{2} - \frac{x^3}{3} = C. \tag{19.4}$$

You want $y = 1$ when $x = 0$ and so you must have

$$\frac{1}{2} = C.$$

The solution is

$$y = \sqrt{2\left(\frac{x^3}{3} + \frac{1}{2}\right)}$$

where it was necessary to pick the positive square root because otherwise, you would not have $y(0) = 1$.

Sometimes you can't solve for y in terms of x.

Example 19.3.3 *Find the solutions to the differential equation,*

$$y' = \frac{x^2}{y \sin y}.$$

In this case,

$$(y \sin y)\, dy = x^2$$

and so $\int y \sin y = \sin y - y \cos y$

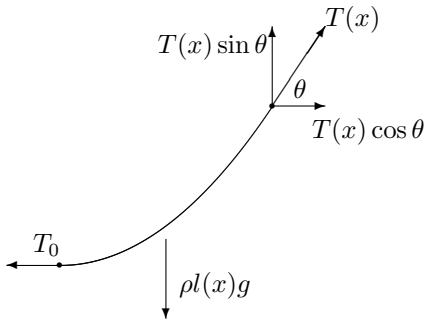
$$\sin y - y \cos y - \frac{x^3}{3} = C$$

gives the solutions. I would not like to try and solve this for y in terms of x. Therefore, in this case, it is customary to leave the solution in this form and refer to it as an implicitly defined solution. The point is, the above equation does define y as a function of x near typical points on the level curve but it might not be possible to algebraically find y as a function of x. You notice in the above argument for finding solutions, it was never assumed that you could algebraically find y as a function of x in $F(x) - G(y) = C$.

Here is an interesting example which is non trivial.

Example 19.3.4 *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

$$y'(x) = cl(x).$$

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

$$\frac{dz}{dx} = c\sqrt{1+z^2},$$

a separable differential equation. Thus

$$\frac{dz}{\sqrt{1+z^2}} = cdx.$$

Now $\int \frac{dz}{\sqrt{1+z^2}} = \operatorname{arcsinh}(z) + C$ and so the solutions are of the form

$$\operatorname{arcsinh}(z) - cx = d$$

where d is some constant. Thus

$$y' = z = \sinh(cx + d)$$

and so

$$y(x) \in \int \sinh(cx + C)\, dx = \frac{\cosh(cx + d)}{c} + k$$

where k is some constant. 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.

19.3.2 Exercises With Answers*

1. Find the solution to the initial value problem,

$$y' = \frac{x}{y^2},\ y(0) = 1.$$

Separating the variables, you get $y^2 dy = x dx$ and so $\frac{y^3}{3} - \frac{x^2}{2} = c$. From the initial condition, $\frac{1}{3} = c$ and so the solution is

$$\frac{y^3}{3} - \frac{x^2}{2} = \frac{1}{3}$$

2. Find the solution to the initial value problem,

$$\tan(y)\, y' = \sin x,\ y\left(\frac{\pi}{4}\right) = \frac{\pi}{4}.$$

Separating the variables, $\tan(y)\, dy = \sin(x)\, dx$ and so $\ln|\sec(y)| + \cos(x) = c$. Now from the initial condition,

$$\ln\left(\sqrt{2}\right) + \frac{\sqrt{2}}{2} = c$$

and so $\ln|\sec(y)| + \cos(x) = \ln\left(\sqrt{2}\right) + \frac{\sqrt{2}}{2}$

3. Find the solution to the initial value problem,

$$y' = \frac{y}{x},\ y(1) = 1.$$

Separating the variables, gives $\frac{dy}{y} = \frac{dx}{x}$ and so $y - x = c$. But from the initial condition, $c = 0$. Hence $y = x$.

19.3.3 A Heat Seaking Particle

Suppose the temperature is given as $T(x,y,z)$ and a particle tries to go in the direction of most rapid rate of change of temperature. In other words this particle likes it hot. This means it moves in the direction of the gradient of T. In other words,

$$(x',y',z')^T = k(x,y,z)\nabla T(x,y,z).$$

Of course you don't know what $k(x,y,z)$ is but if you did and if you also knew T, then you would have a system of differential equations for the position of the particle as a function of time. If you were given an initial position, you could then ask for the solution to the resulting intial value problem. Of course you won't be able to solve the equations in general. These sorts of things require numerical methods. Also, in interesting examples, everything would also depend on t. The following pseudo application has to do with a situation which I will cook up so that I will be able to solve everything.

Example 19.3.5 *A heat seaking particle starts at* $(1,2,1)$. *The temperature is* $T(x,y,z) = x^2+y+z^3$ *and assume that* $k=1$. *Find the motion of the heat seaking particle.*

As explained above, you need $(x',y',z')^T = \nabla T(x,y,z)$ and so

$$\frac{dx}{dt} = 2x, \frac{dy}{dt} = 1, \frac{dz}{dt} = 3z^2$$

because $\nabla T = \left(2x,1,3z^2\right)^T$. It is very fortunate that the equations are not coupled. Consider the first one. Separating the variables,

$$\frac{dx}{x} = 2dt$$

and so $\ln(x) - 2t = c$. From the initial condition which states that at $t=0, x=1$, it follows $c=0$. Therefore, $x = e^{2t}$. Next consider the second of the differential equations. This one says $y = t+c$ and from the initial condition, $c=2$ so the second gives $y = t+2$. Finally the last equation separates to give

$$\frac{dz}{z^2} = 3dt$$

and so

$$\frac{-1}{z} = 3t + c.$$

In this case the initial data gives $c=-1$. Therefore, $z = -\frac{1}{3t-1}$. It follows the path of the particle is of the form

$$\left(e^{2t}, t+2, -\frac{1}{3t-1}\right).$$

Note that this only makes sense for $t \in [0, \frac{1}{3})$. This type of thing is typical of nonlinear differential equations.

I think you can see how to do similar problems in which the particle is heat avoiding. You just put in a minus sign by ∇T.

19.4 The Chain Rule

Remember what this was all about for a function of one variable. You had $z = f(y)$ and $y = g(x)$ and you wanted to find $\frac{dz}{dx}$. Remember the answer was

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}.$$

The chain rule was one of the most important rules for differentiation. Its importance is no less for functions of many variables.

The problem is this: $z = f(y_1, y_2, \cdots, y_n)$ and $y_k = g_k(x_1, \cdots, x_m)$. You want to find $\frac{\partial z}{\partial x_k}$ for each $k = 1, 2, \cdots, m$. It turns out to be exactly the same sort of formula which works. In this case the formula is

$$\frac{\partial z}{\partial x_k} = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x_k}.$$

People who use the repeated index summation convention write this as

$$\frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x_k}$$

which is formally just like it was for a function of one variable. I think this is one reason for the attractiveness of this repeated summation convention. Here is an example.

Example 19.4.1 *Suppose* $z = y_1 + y_2 y_3^2$ *and* $y_1 = \sin(x_1) + x_2, y_2 = \cos(x_3)$, *and* $y_3 = x_1^2 + \sin x_2 + x_4$. *Find* $\frac{\partial z}{\partial x_2}$ *and* $\frac{\partial z}{\partial x_4}$.

From the above formula,

$$\begin{aligned} \frac{\partial z}{\partial x_2} &= \frac{\partial z}{\partial y_1} \frac{\partial y_1}{\partial x_2} + \frac{\partial z}{\partial y_2} \frac{\partial y_2}{\partial x_2} + \frac{\partial z}{\partial y_3} \frac{\partial y_3}{\partial x_2} \\ &= 1 \times 1 + y_3^2 \times 0 + 2 y_2 y_3 \cos x_2 \\ &= 1 + 2 y_2 y_3 \cos x_2. \end{aligned}$$

If you want to put this in terms of the x variables, it is

$$\begin{aligned} \frac{\partial z}{\partial x_2} &= 1 + 2 y_2 y_3 \cos x_2 \\ &= 1 + 2\cos(x_3)\left(x_1^2 + \sin x_2 + x_4\right) \cos x_2 \end{aligned}$$

Now consider the other partial derivative.

$$\begin{aligned} \frac{\partial z}{\partial x_4} &= \frac{\partial z}{\partial y_1} \frac{\partial y_1}{\partial x_4} + \frac{\partial z}{\partial y_2} \frac{\partial y_2}{\partial x_4} + \frac{\partial z}{\partial y_3} \frac{\partial y_3}{\partial x_4} \\ &= \frac{\partial z}{\partial y_1} \times 0 + \frac{\partial z}{\partial y_2} \times 0 + 2 y_2 y_3 \times 1 \\ &= 2\cos(x_3)\left(x_1^2 + \sin x_2 + x_4\right). \end{aligned}$$

Be sure you can find and place the partial derivatives in terms of the independent variables, $\mathbf{x}$. It is just as correct to leave the answer in terms of $\mathbf{y}$ and $\mathbf{x}$ but sometimes people may insist you place the answer in terms of $\mathbf{x}$.

The next task is to explain why the above formula works. The argument I will give applies to one dimension also. Therefore, you can consider it a review of what you should have seen in beginning calculus.

Lemma 19.4.2 *Suppose* U *is an open set in* $\mathbb{R}^n$ *and* $f : U \to \mathbb{R}$. *Suppose* $\mathbf{x} \in U$ *and for all* $\mathbf{v}$ *small enough,*

$$f(\mathbf{x} + \mathbf{v}) - f(\mathbf{x}) = \sum_{i=1}^{n} a_i v_i + o(\mathbf{v}).$$

Then $a_i = \frac{\partial f}{\partial x_i}(\mathbf{x})$ *and* f *is differentiable.*

Proof: Let t be a small nonzero number. Then since the i^{th} component of $\mathbf{e}_k$ equals zero unless $i = k$ when it is 1,

$$\begin{aligned} f(\mathbf{x} + t\mathbf{e}_k) - f(\mathbf{x}) &= a_k t + o(t\mathbf{e}_k) \\ &= a_k t + o(t) \end{aligned}$$

Now divide by t and take a limit.

$$\frac{\partial f}{\partial x_k}(\mathbf{x}) = \lim_{t\to 0} \frac{f(\mathbf{x} + t\mathbf{e}_k) - f(\mathbf{x})}{t} = \lim_{t\to 0} \left(a_k + \frac{o(t)}{t} \right) = a_k.$$

This proves the lemma.

Lemma 19.4.3 *Let U be an open set and suppose g is differentiable at $\mathbf{x} \in U$. Then*

$$\mathbf{o}(g(\mathbf{x} + \mathbf{v}) - g(\mathbf{x})) = \mathbf{o}(\mathbf{v}).$$

Proof: I need to show

$$\lim_{|\mathbf{v}|\to 0} \frac{\mathbf{o}(g(\mathbf{x} + \mathbf{v}) - g(\mathbf{x}))}{|\mathbf{v}|} = \mathbf{0}.$$

Let $\varepsilon > 0$ be given. Since g is continuous at $\mathbf{x}$, there exists $\delta_1 > 0$ such that if $|\mathbf{v}| < \delta_1$, then

$$\frac{|\mathbf{o}(g(\mathbf{x} + \mathbf{v}) - g(\mathbf{x}))|}{|g(\mathbf{x} + \mathbf{v}) - g(\mathbf{x})|} < \varepsilon$$

Hence, for such $\mathbf{v}$,

$$|\mathbf{o}(g(\mathbf{x} + \mathbf{v}) - g(\mathbf{x}))| < \varepsilon |g(\mathbf{x} + \mathbf{v}) - g(\mathbf{x})| \tag{19.5}$$

Since g is differentiable at $\mathbf{x}$, there exists $\delta_2 > 0$ such that if $|\mathbf{v}| < \delta_2$,

$$\frac{\left| g(\mathbf{x} + \mathbf{v}) - g(\mathbf{x}) - \sum_{k=1}^n \frac{\partial g}{\partial x_k}(\mathbf{x}) v_k \right|}{|\mathbf{v}|} < \varepsilon$$

Hence for $|\mathbf{v}| < \delta_2$,

$$|g(\mathbf{x} + \mathbf{v}) - g(\mathbf{x})| < \left| \sum_{k=1}^n \frac{\partial g}{\partial x_k}(\mathbf{x}) v_k \right| + \varepsilon |\mathbf{v}| \tag{19.6}$$

Let $\delta \leq \min(\delta_1, \delta_2)$. Then if $|\mathbf{v}| < \delta$, both 19.5 and 19.6 hold and so by the Cauchy Schwarz inequality,

$$\begin{aligned} |\mathbf{o}(g(\mathbf{x} + \mathbf{v}) - g(\mathbf{x}))| &< \varepsilon |g(\mathbf{x} + \mathbf{v}) - g(\mathbf{x})| \\ &< \varepsilon \left(\left| \sum_{k=1}^n \frac{\partial g}{\partial x_k}(\mathbf{x}) v_k \right| + \varepsilon |\mathbf{v}| \right) \\ &\leq \varepsilon \left(\sum_{k=1}^n \left| \frac{\partial g}{\partial x_k}(\mathbf{x}) \right|^2 \right)^{1/2} |\mathbf{v}| + \varepsilon |\mathbf{v}|. \end{aligned}$$

Dividing both sides by $|\mathbf{v}|$,

$$\frac{\mathbf{o}(g(\mathbf{x} + \mathbf{v}) - g(\mathbf{x}))}{|\mathbf{v}|} \leq \varepsilon \left(\left(\sum_{k=1}^n \left| \frac{\partial g}{\partial x_k}(\mathbf{x}) \right|^2 \right)^{1/2} + 1 \right)$$

and since $\varepsilon > 0$ is arbitrary, this establishes the lemma.

With these lemmas, it is easy to prove the chain rule.

Theorem 19.4.4 *Let V be an open set in $\mathbb{R}^n$ and let U be an open set in $\mathbb{R}^m$. Also let $\mathbf{g}: U \to V$ be a vector valued function having the property that for*

$$\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \cdots, g_m(\mathbf{x})),$$

each g_k is differentiable at $\mathbf{x} \in U$. Also suppose $f: V \to \mathbb{R}$ is differentiable at $\mathbf{g}(\mathbf{x})$. Then for $z \equiv f \circ \mathbf{g}$, $y_i = g_i(\mathbf{x})$,

$$\frac{\partial z}{\partial x_k}(\mathbf{x}) = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i}(\mathbf{g}(\mathbf{x})) \frac{\partial y_i}{\partial x_k}(\mathbf{x}).$$

Proof: Using Lemma 19.4.3 as needed,

$$\begin{aligned}
f \circ \mathbf{g}(\mathbf{x}+\mathbf{v}) - f \circ \mathbf{g}(\mathbf{x}) &= \sum_{i=1}^{n} \frac{\partial z}{\partial y_i}(\mathbf{g}(\mathbf{x}))(g_i(\mathbf{x}+\mathbf{v}) - g_i(\mathbf{x})) + o(g_i(\mathbf{x}+\mathbf{v}) - g_i(\mathbf{x})) \\
&= \sum_{i=1}^{n} \frac{\partial z}{\partial y_i}(\mathbf{g}(\mathbf{x}))(g_i(\mathbf{x}+\mathbf{v}) - g_i(\mathbf{x})) + o(\mathbf{v}) \\
&= \sum_{i=1}^{n} \frac{\partial z}{\partial y_i}(\mathbf{g}(\mathbf{x}))\left(\sum_{k=1}^{m} \frac{\partial y_i}{\partial x_k}(\mathbf{x}) v_k + o(\mathbf{v})\right) + o(\mathbf{v}) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{m} \frac{\partial z}{\partial y_i}(\mathbf{g}(\mathbf{x})) \frac{\partial y_i}{\partial x_k}(\mathbf{x}) v_k + o(\mathbf{v}) \\
&= \sum_{k=1}^{m}\left(\sum_{i=1}^{n} \frac{\partial z}{\partial y_i}(\mathbf{g}(\mathbf{x})) \frac{\partial y_i}{\partial x_k}(\mathbf{x})\right) v_k + o(\mathbf{v})
\end{aligned}$$

Now by Lemma 19.4.2,

$$\frac{\partial f \circ \mathbf{g}}{\partial x_k}(\mathbf{x}) \equiv \frac{\partial z}{\partial x_k} = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i}(\mathbf{g}(\mathbf{x})) \frac{\partial y_i}{\partial x_k}(\mathbf{x}).$$

This proves the theorem.

19.4.1 Related Rates Problems

Sometimes several variables are related and given information about how one variable is changing, you want to find how the others are changing. The following law is discussed later in the book, on Page 519.

Example 19.4.5 *Bernoulli's law states that in an incompressible fluid,*

$$\frac{v^2}{2g} + z + \frac{P}{\gamma} = C$$

where C is a constant. Here v is the speed, P is the pressure, and z is the height above some reference point. The constants, g and γ are the acceleration of gravity and the weight density of the fluid. Suppose measurements indicate that $\frac{dv}{dt} = -3$, and $\frac{dz}{dt} = 2$. Find $\frac{dP}{dt}$ when $v = 7$ and $z = 8$ in terms of g and γ.

This is just an exercise in using the chain rule. Differentiate the two sides with respect to t.

$$\frac{1}{g} v \frac{dv}{dt} + \frac{dz}{dt} + \frac{1}{\gamma} \frac{dP}{dt} = 0.$$

Then when $v = 7$ and $z = 8$, finding $\frac{dP}{dt}$ involves nothing more than solving the following for $\frac{dP}{dt}$.

$$\frac{7}{g}(-3) + 2 + \frac{1}{\gamma}\frac{dP}{dt} = 0$$

Thus

$$\frac{dP}{dt} = \gamma\left(\frac{21}{g} - 2\right)$$

at this instant in time.

Example 19.4.6 *In Bernoulli's law above, each of v, z, and P are functions of (x, y, z), the position of a point in the fluid. Find a formula for $\frac{\partial P}{\partial x}$ in terms of the partial derivatives of the other variables.*

This is an example of the chain rule. Differentiate both sides with respect to x.

$$\frac{v}{g}v_x + z_x + \frac{1}{\gamma}P_x = 0$$

and so

$$P_x = -\left(\frac{vv_x + z_x g}{g}\right)\gamma$$

Example 19.4.7 *Suppose a level curve is of the form $f(x, y) = C$ and that near a point on this level curve, y is a differentiable function of x. Find $\frac{dy}{dx}$.*

This is an example of the chain rule. Differentiate both sides with respect to x. This gives

$$f_x + f_y\frac{dy}{dx} = 0.$$

Solving for $\frac{dy}{dx}$ gives

$$\frac{dy}{dx} = \frac{-f_x(x, y)}{f_y(x, y)}.$$

Example 19.4.8 *Suppose a level surface is of the form $f(x, y, z) = C$. and that near a point, (x, y, z) on this level surface, z is a C^1 function of x and y. Find a formula for z_x.*

This is an exaple of the use of the chain rule. Differentiate both sides of the equation with respect to x. Since $y_x = 0$, this yields

$$f_x + f_z z_x = 0.$$

Then solving for z_x gives

$$z_x = \frac{-f_x(x, y, z)}{f_z(x, y, z)}$$

Example 19.4.9 *Polar coordinates are*

$$x = r\cos\theta,\ y = r\sin\theta.$$

Thus if f is a C^1 scalar valued function you could ask to express f_x in terms of the variables, r and θ. Do so.

This is an example of the chain rule. $f = f(r,\theta)$ and so

$$f_x = f_r r_x + f_\theta \theta_x.$$

This will be done if you can find r_x and θ_x. However you must find these in terms of r and θ, not in terms of x and y. Using the chain rule on the two equations for the transformation,

$$\begin{aligned} 1 &= r_x \cos\theta - (r\sin\theta)\,\theta_x \\ 0 &= r_x \sin\theta + (r\cos\theta)\,\theta_x \end{aligned}$$

Solving these using Cramer's rule yields

$$r_x = \cos(\theta),\ \theta_x = \frac{-\sin(\theta)}{r}$$

Hence f_x in polar coordinates is

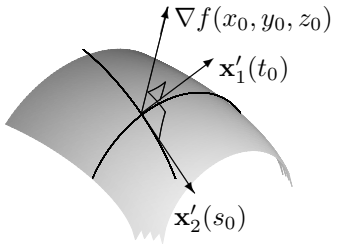
$$f_x = f_r(r,\theta)\cos(\theta) - f_\theta(r,\theta)\left(\frac{\sin(\theta)}{r}\right)$$

19.5 Normal Vectors And Tangent Planes 26 Oct.

Quiz

1. Let $z = xy^2$ and let $x = ts + ps$ while $y = 2s^2 + t$. Find $\frac{\partial z}{\partial s}$ when $(s,t,p) = (1,1,1)$.
2. A level surface is given by $x^3y + z^2 = 2$. Find z_x at the point $(1,1,1)$ on the level surface.
3. Suppose $x = t^3 + s$ and $y = s^3 + t$. Find $\frac{\partial z}{\partial x}$ completely in terms of partial derivativs and functions of the new variables, s,t.

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

$$f(x_1(t), y_1(t), z_1(t)) = c$$

[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 consequence of the implicit function theorem, one of the greatest theorems in all mathematics and a topic for an advanced calculus class. See the the section on the implicit function theorem for the most elementary treatment of this theorem that I know.

for all t in some interval. Therefore, taking the derivative of both sides and using the chain rule on the left,

$$\frac{\partial f}{\partial x}\left(x_1\left(t\right), y_1\left(t\right), z_1\left(t\right)\right) x_1'\left(t\right) +$$
$$\frac{\partial f}{\partial y}\left(x_1\left(t\right), y_1\left(t\right), z_1\left(t\right)\right) y_1'\left(t\right) + \frac{\partial f}{\partial z}\left(x_1\left(t\right), y_1\left(t\right), z_1\left(t\right)\right) z_1'\left(t\right) = 0.$$

In terms of the gradient, this merely states

$$\nabla f\left(x_1\left(t\right), y_1\left(t\right), z_1\left(t\right)\right) \cdot \mathbf{x}_1'\left(t\right) = 0.$$

Similarly,

$$\nabla f\left(x_2\left(s\right), y_2\left(s\right), z_2\left(s\right)\right) \cdot \mathbf{x}_2'\left(s\right) = 0.$$

Letting $s = s_0$ and $t = t_0$, it follows

$$\nabla f\left(x_0, y_0, z_0\right) \cdot \mathbf{x}_1'\left(t_0\right) = 0,\ \nabla f\left(x_0, y_0, z_0\right) \cdot \mathbf{x}_2'\left(s_0\right) = 0.$$

It follows $\nabla f\left(x_0, y_0, z_0\right)$ 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 $\left(x_0, y_0, z_0\right)$ and that $\nabla f\left(x_0, y_0, z_0\right)$ is perpendicular to this tangent plane at the point, $\left(x_0, y_0, z_0\right)$.

Example 19.5.1 *Find the equation of the tangent plane to the level surface, $f\left(x, y, z\right) = 6$ of the function, $f\left(x, y, z\right) = x^2 + 2y^2 + 3z^2$ at the point $\left(1, 1, 1\right)$.*

First note that $\left(1, 1, 1\right)$ is a point on this level surface. To find the desired plane it suffices to find the normal vector to the proposed plane. But $\nabla f\left(x, y, z\right) = \left(2x, 4y, 6z\right)$ and so $\nabla f\left(1, 1, 1\right) = \left(2, 4, 6\right)$. Therefore, from this problem, the equation of the plane is

$$\left(2, 4, 6\right) \cdot \left(x - 1, y - 1, z - 1\right) = 0$$

or in other words,

$$2x - 12 + 4y + 6z = 0.$$

Example 19.5.2 *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

$$\frac{\left(2\sqrt{3}\right)^2 + 4 - 1}{17} = \frac{15}{17}.$$

Example 19.5.3 *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 that is needed 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 desired line is

$$\left(1, \sqrt{3}, 4\right) + t\left(2, 2\sqrt{3}, -1\right).$$

Extrema Of Functions Of Several Variables 30 Oct.

Quiz

1. Let $z = x^2 \sin(y)$ and let $x = t^2 s + r$ while $y = t^2 - s$. Find z_t when $(s,t,r) = (1,1,1)$.
2. The temperature in space is given by $T(x,y,z) = x + 2y^2 + z^2$. Find the path of a heat seeking particle which starts at the point $(0,1,1)$. This is an ill defined problem. Make the usual assumptions.
3. The ideal gas law is $PV = kT$ where k is a constant. Suppose at some time $\frac{dT}{dt} = 1, \frac{dP}{dt} = -1$. Find $\frac{dV}{dt}$ at this instant if $P = 2, V = 6, T = 100$.
4. There are two surfaces, $x^2 + y^2 = 1$ and $x^2 + y^2 + z^2 = 5$ which intersect in a curve. Find an equation of the tangent line to this curve at the point, $\left(\frac{\sqrt{3}}{2}, \frac{1}{2}, 2\right)$.
5. A level surface is $x^2 + 2y^2 + 3z^2 = 6$. Find the tangent plane at the point, $(1,1,1)$.

Quiz

1. Let $z = x \sin\left(x^2 + y^2\right)$. Find $\frac{\partial z}{\partial x}$.
2. Let $z^3 \sin(x) + y^4 z = 7$. Find $\frac{\partial z}{\partial x}$.
3. Suppose $f(x,y)$ is given by

$$f(x,y) = \begin{cases} \frac{2yx + x^3 + xy^2}{x^2 + y^2} \text{ if } (x,y) \neq (0,0) \\ 0 \text{ if } (x,y) = (0,0) \end{cases}$$

   Find $f_x(0,0)$ if possible.
4. Find parametric equations of the line which is perpendicular to the surface $3x^2 + 2y^2 + z^2 = 6$ at the point $(1,1,1)$.
5. Let $u(x,y) = f(x+y) + g(x-y)$. Compute $u_{xx} - u_{yy}$. D'Lambert did this problem back in the mid 1700's and it turned out to be very important.
6. A function of two variables, $f(x,y)$ is called homogeneous of degree α if $f(tx,ty) = t^\alpha f(x,y)$. Establish Euler's identity which states that for such homogeneous functions,

$$x f_x(x,y) + y f_y(x,y) = \alpha f(x,y).$$

This identity dates from early in the 1700's also. **Hint:** Use the chain rule and differentiate both sides of $f(tx,ty)=t^{\alpha}f(x,y)$ with respect to t using the chain rule and then plug in $t=1$.

Suppose $f:D(f)\to\mathbb{R}$ where $D(f)\subseteq\mathbb{R}^n$.

20.1 Local Extrema

Definition 20.1.1 *A point $\mathbf{x}\in D(f)\subseteq\mathbb{R}^n$ 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. The plural for minimum is **minima** and the plural for maximum is **maxima.***

Procedure 20.1.2 *To find candidates for local extrema which are interior points of $D(f)$ where f is a differentiable 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*

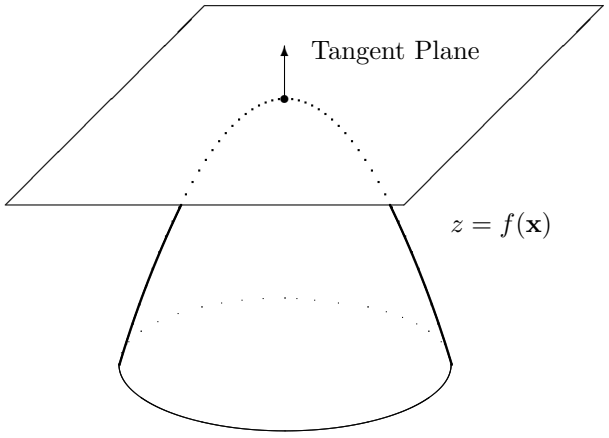
$$F(\mathbf{x},z)\equiv z-f(\mathbf{x})=0$$

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 $\nabla f(\mathbf{x})=\mathbf{0}$.

This is illustrated in the following picture.



A more rigorous explanation is as follows. Let $\mathbf{v}$ be any vector in $\mathbb{R}^n$ and suppose $\mathbf{x}$ is a local maximum (minimum) for $\mathbf{f}$. Then consider the real valued function of one variable, $h(t)\equiv f(\mathbf{x}+t\mathbf{v})$ for small $|t|$. Since $\mathbf{f}$ has a local maximum (minimum), it follows that h is a differentiable function of the single variable t for small t which has a local maximum (minimum) when $t=0$. Therefore, $h'(0)=0$. But $h'(t)=\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}(\mathbf{x}+t\mathbf{v})\,v_i$ by the chain rule. Therefore,

$$h'(0)=\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}(\mathbf{x})\,v_i=0$$

and since $\mathbf{v}$ is arbitrary, it follows $\frac{\partial f}{\partial x_i}(\mathbf{x}) = 0$ for each i. Thus

$$\left(\begin{array}{ccc} f_{x_1}(\mathbf{x}) & \cdots & f_{x_n}(\mathbf{x}) \end{array} \right)^T = \mathbf{0}$$

and so $\nabla f(\mathbf{x}) = \mathbf{0}$. This proves the following theorem.

Theorem 20.1.3 *Suppose U is an open set contained in $D(f)$ such that f is C^1 on U and suppose $\mathbf{x} \in U$ is a local minimum or local maximum for f. Then $\nabla f(\mathbf{x}) = \mathbf{0}$.*

Definition 20.1.4 *A **singular point** for f is a point $\mathbf{x}$ where $\nabla f(\mathbf{x}) = \mathbf{0}$. This is also called a **critical point**. By analogy with the one variable case, a point where the gradient does not exist will also be called a critical point.*

Example 20.1.5 *Find the critical points 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 critical point, $(1,1)$.

Example 20.1.6 *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, letting (x, y) take the place of (x_0, y_0) for simplicity, 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. How do you know this is not the largest tetrahedron?

Example 20.1.7 *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 $xyz = 32$. How do you know this gives the least surface area? Why doesn't this give the largest surface area?

20.2 The Second Derivative Test

20.2.1 Functions Of Two Variables

In the special case of a function of two variables, $f(x, y)$ which is the only case considered in most calculus books, the second derivative test is given in the following theorem. It is a black box formulation of the second derivative test.

Theorem 20.2.1 *(Second Derivative Test) Let f be a function of two variables defined on an open set, U whose second order partial derivatives exist and are continuous. That is, $f \in C^2(U)$. Suppose $(a, b) \in U$ is a point where both partial derivatives of f vanishes. That is $f_x(a, b) = f_y(a, b) = 0$. Let*

$$D \equiv f_{xx}(a, b) f_{yy}(a, b) - (f_{xy}(a, b))^2.$$

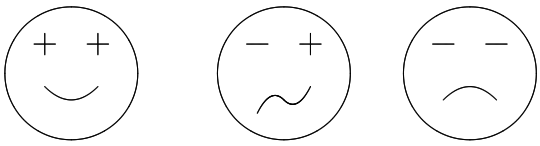
Then

1. *If $D > 0$ and $f_{xx}(a, b) < 0$, then f has a local maximum at (a, b).*
2. *If $D > 0$ and $f_{xx}(a, b) > 0$, then f has a local mimimum at (a, b).*
3. *If $D < 0$, then f has a saddle point at (a, b).*
4. *If $D = 0$, the test fails.*

The above is really a statement about the eigenvalues of the **Hessian matrix**,

$$H \equiv \begin{pmatrix} f_{xx}(a, b) & f_{x,y}(a, b) \\ f_{x,y}(a, b) & f_{yy}(a, b) \end{pmatrix}$$

at a point (a, b) where the partial derivatives of f vanish. It reduces to the following much simpler statement. If both eigenvalues of H are positive, then f has a local minimum at (a, b). If both eigenvalues are negative, then f has a local maximum at (a, b). If one eigenvalue is positive and one is negative, then you have a saddle point at (a, b). If at least one eigenvalue equals zero, then the test fails. Here is a picture which may help you remember this second version of this test.



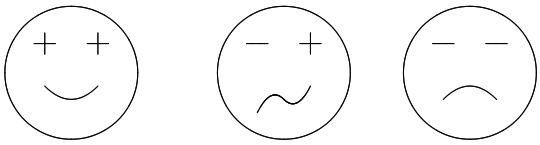
Use whichever version of this theorem you find easiest to remember. However, in the case of a function of many variables, the description I just gave has an obvious generalization. This is presented next. If you are not interested in it, I think you can skip it because it isn't included in the book for the course.

20.2.2 Functions Of Many Variables*

There is a version of the second derivative test for a function of many variables 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 21.4.

Definition 20.2.2 *The matrix, $H(\mathbf{x})$ whose ij^{th} entry at the point $\mathbf{x}$ is $\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})$ is called the* ***Hessian matrix****.*

The following theorem says that if all the eigenvalues of the Hessian matrix at a critical point are positive, then the critical point is a local minimum. If all the eigenvalues of the Hessian matrix at a critical point are negative, then the critical point is a local maximum. Finally, if some of the eigenvalues of the Hessian matrix at the critical point are positive and some are negative then the critical point is a saddle point. The following picture illustrates the situation.



Theorem 20.2.3 *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, $H(\mathbf{x})$. 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 20.2.4 *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 minima, local maxima, 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

$$\begin{pmatrix} 24x^2+28+24y-24x & 24x-12 \\ 24x-12 & 4 \end{pmatrix}.$$

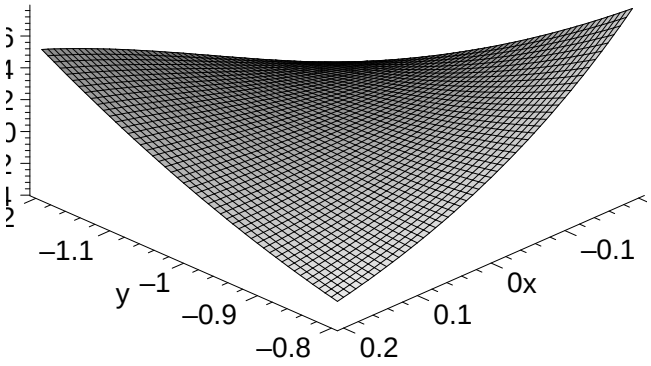
and the thing to determine is the sign of its eigenvalues evaluated at the critical points.

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.

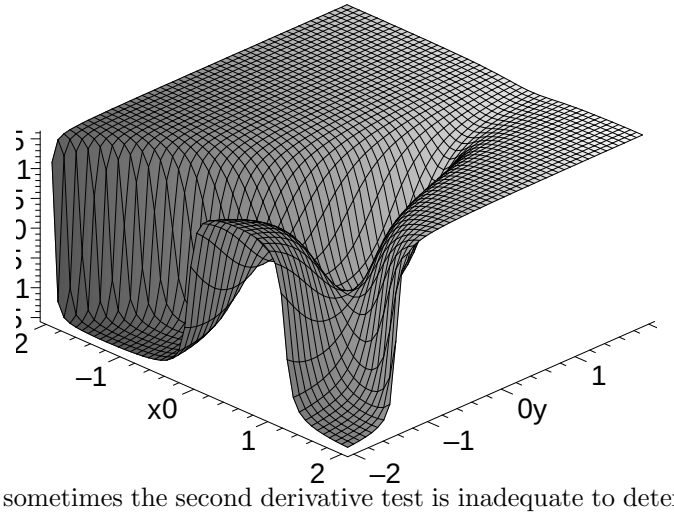
The geometric significance of a saddle point was explained above. In one direction it looks like a local minimum while in another it looks like a local maximum. In fact, they do look like a saddle. Here is a picture of the graph of the above function near the saddle point, $(0,-1)$.



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

$$f(x,y) = \arctan\left(2x^4-4x^3+14x^2+12yx^2-12yx-12x+2y^2+4y+2\right).$$

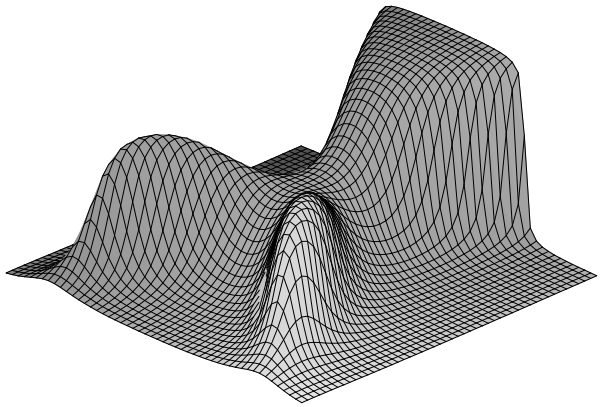
Since arctan is a strictly increasing function, it preserves all the information about whether the given function is increasing or decreasing in certain directions. Below is a graph of this function which illustrates the behavior near the point $(1,-1)$.



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 20.2.5 *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

$$g\left(x,y\right) = 6xy^2 - 2x^3 - 3y^4$$

$g_x\left(0,0\right) = g_y\left(0,0\right) = 0$.

$$g_x\left(x,y\right) = 6y^2 - 6x^2, \quad g_y\left(x,y\right) = 12xy - 12y^3$$

and clearly $(0,0)$ is a critical point. So are $(1,1)$ and $(1,-1)$. Now $g_{xx}(0,0)=0$ and so does $g_{xy}(0,0)$ and $g_{yy}(0,0)$. 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 $h(t)=f(t,t)=\arctan(4t^3-3t^4)$, 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 $p(t)\equiv f(0,t)=-3t^4$. Therefore, along the line, $(0,t)$, f has a local maximum at $(0,0)$.

The following example is for a function of three variables.

Example 20.2.6 *Find the critical points of the following function of three variables and classify them as local minimums, local maximums or saddle points.*

$$f(x,y,z)=\frac{5}{6}x^2+4x+16-\frac{7}{3}xy-4y-\frac{4}{3}xz+12z+\frac{5}{6}y^2-\frac{4}{3}zy+\frac{1}{3}z^2$$

First you need to locate the critical points. This involves taking the gradient.

$$\begin{aligned}&\nabla\left(\frac{5}{6}x^2+4x+16-\frac{7}{3}xy-4y-\frac{4}{3}xz+12z+\frac{5}{6}y^2-\frac{4}{3}zy+\frac{1}{3}z^2\right)\\ =\;&\left(\frac{5}{3}x+4-\frac{7}{3}y-\frac{4}{3}z,-\frac{7}{3}x-4+\frac{5}{3}y-\frac{4}{3}z,-\frac{4}{3}x+12-\frac{4}{3}y+\frac{2}{3}z\right)\end{aligned}$$

Next you need to set the gradient equal to zero and solve the equations. This yields $y=5, x=3, z=-2$. Now to use the second derivative test, you assemble the Hessian matrix which is

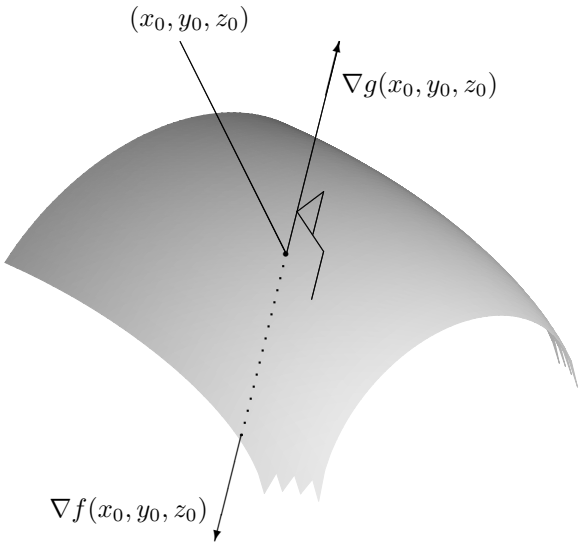
$$\begin{pmatrix}\frac{5}{3}&-\frac{7}{3}&-\frac{4}{3}\\ -\frac{7}{3}&\frac{5}{3}&-\frac{4}{3}\\ -\frac{4}{3}&-\frac{4}{3}&\frac{2}{3}\end{pmatrix}.$$

Note that in this simple example, the Hessian matrix is constant and so all that is left is to consider the eigenvalues. Writing the characteristic equation and solving yields the eigenvalues are $2,-2,4$. Thus the given point is a saddle point.

20.3 Lagrange Multipliers, Constrained Extrema 31 Oct.

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. I think he did it in the context of the calculus of variations in the presence of constraints.

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 20.3.1 *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.

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 20.3.2 *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 $\left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right)$.

Example 20.3.3 *Find candidates for 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.

$$\begin{gathered} y - 2x - \lambda(2x+2y) = 0 \\ x - 2\lambda(x+y) = 0 \\ x^2+2xy+y^2 = 4 \end{gathered}$$

[1]There exists such a point because the sphere is closed and bounded.

From the first two equations,

$$\begin{gathered}(2+2\lambda)\,x-(1-2\lambda)\,y=0\\(1-2\lambda)\,x-2\lambda y=0\end{gathered}$$

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,

$$y=3x \tag{20.1}$$

From the constraint equation,

$$x^2+2x\,(3x)+(3x)^2=4$$

and so

$$x=\frac{1}{2}\text{ or }-\frac{1}{2}$$

Now from 20.1, the points of interest on the boundary of this set are

$$\left(\frac{1}{2},\frac{3}{2}\right),\text{ and }\left(-\frac{1}{2},-\frac{3}{2}\right). \tag{20.2}$$

$$\begin{aligned}f\left(\frac{1}{2},\frac{3}{2}\right) &= \left(\frac{1}{2}\right)\left(\frac{3}{2}\right)-\left(\frac{1}{2}\right)^2\\ &= \frac{1}{2}\end{aligned}$$

$$\begin{aligned}f\left(-\frac{1}{2},-\frac{3}{2}\right) &= \left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right)-\left(-\frac{1}{2}\right)^2\\ &= \frac{1}{2}\end{aligned}$$

It follows the candidates for maximum and minimum are $\left(\frac{1}{2},\frac{3}{2}\right),(0,0)$, and $\left(-\frac{1}{2},-\frac{3}{2}\right)$. Therefore it appears that $(0,0)$ yields a minimum and either $\left(\frac{1}{2},\frac{3}{2}\right)$ or $\left(-\frac{1}{2},-\frac{3}{2}\right)$ yields a maximum. However, this is a little misleading. How do you even know a maximum or a minimum exists? The set, $x^2+2xy+y^2\leq 4$ is an unbounded set which lies between the two lines $x+y=2$ and $x+y=-2$. In fact there is no minimum. For example, take $x=100, y=-98$. Then $xy-x^2=x\,(y-x)=100\,(-98-100)$ which is a large negative number much less than 0, the answer for the point $(0,0)$.

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.

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 or inverse 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. I have also given a complete proof later starting on Page 389. First here is a simple definition explaining one of the terms in the statement of this theorem.

Definition 20.3.4 *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 20.3.5 *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*

$$g_i(\mathbf{x}) = 0, \ i = 1, \cdots, m \tag{20.3}$$

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

$$\begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \tag{20.4}$$

holds.

To help remember how to use 20.4 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 20.4 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 20.3.6 *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.

$$\begin{array}{c} 5y^2 + z^2 = 4 \\ y^2 - \lambda z = 0 \\ yz - \lambda y + \mu = 0 \\ yz - 4\lambda y - \mu = 0 \end{array}$$

From the last equation, $\mu = (yz - 4\lambda y)$. Substitute this into the third and get

$$\begin{array}{c} 5y^2 + z^2 = 4 \\ y^2 - \lambda z = 0 \\ yz - \lambda y + yz - 4\lambda y = 0 \end{array}$$

$y = 0$ 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

$$\begin{array}{c} 5y^2 + z^2 = 4 \\ y^2 - (2/5)\, z^2 = 0 \end{array} ,$$

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.

20.3.1 Exercises With Answers

1. Maximize $x + 3y - 6z$ subject to the constraint, $x^2 + 2y^2 + z^2 = 9$.

   The Lagrangian is $L = x + 3y - 6z - \lambda\left(x^2 + 2y^2 + z^2 - 9\right)$. Now take the derivative with respect to x. This gives the equation $1 - 2\lambda x = 0$. Next take the derivative with respect to y. This gives the equation $3 - 4\lambda y = 0$. The derivative with respect to z gives $-6 - 2\lambda z = 0$. Clearly $\lambda \neq 0$ since this would contradict the first of these equations. Similarly, none of the variables, x, y, z can equal zero. Solving each of these equations for λ gives $\frac{1}{2x} = \frac{3}{4y} = \frac{-3}{z}$. Thus $y = \frac{3x}{2}$ and $z = -6x$. Now you use the constraint equation plugging in these values for y and z. $x^2 + 2\left(\frac{3x}{2}\right)^2 + (-6x)^2 = 9$. This gives the values for x as $x = \frac{3}{83}\sqrt{166}, x = -\frac{3}{83}\sqrt{166}$. From the

three equations above, this also determines the values of z and y. $y = \frac{9}{166}\sqrt{166}$ or $-\frac{9}{166}\sqrt{166}$ and $z = -\frac{18}{83}\sqrt{166}$ or $\frac{18}{83}\sqrt{166}$. Thus there are two points to look at. One will give the minimum value and the other will give the maximum value. You know the minimum and maximum exist because of the extreme value theorem. The two points are $\left(\frac{3}{83}\sqrt{166}, \frac{9}{166}\sqrt{166}, -\frac{18}{83}\sqrt{166}\right)$ and $\left(-\frac{3}{83}\sqrt{166}, -\frac{9}{166}\sqrt{166}, \frac{18}{83}\sqrt{166}\right)$. Now you just need to find which is the minimum and which is the maximum. Plug these in to the function you are trying to maximize. $\left(\frac{3}{83}\sqrt{166}\right) + 3\left(\frac{9}{166}\sqrt{166}\right) - 6\left(-\frac{18}{83}\sqrt{166}\right)$ will clearly be the maximum value occuring at $\left(\frac{3}{83}\sqrt{166}, \frac{9}{166}\sqrt{166}, -\frac{18}{83}\sqrt{166}\right)$. The other point will obviously yield the minimum because this one is positive and the other one is negative. If you use a calculator to compute this you get $\left(\frac{3}{83}\sqrt{166}\right) + 3\left(\frac{9}{166}\sqrt{166}\right) - 6\left(-\frac{18}{83}\sqrt{166}\right) = 19.326$.

2. Find the dimensions of the largest rectangle which can be inscribed in a the ellipse $x^2 + 4y^2 = 4$.

   This is one which you could do without Lagrange multipliers. However, it is easier with Lagrange multipliers. Let a corner of the rectangle be at (x, y). Then the area of the rectangle will be $4xy$ and since (x, y) is on the ellipse, you have the constraint $x^2+4y^2 = 4$. Thus the problem is to maximize $4xy$ subject to $x^2+4y^2 = 4$. The Lagrangian is then $L = 4xy-\lambda\left(x^2 + 4y^2 - 4\right)$ and so you get the equations $4y-2\lambda x = 0$ and $4x-8\lambda y = 0$. You can't have both x and y equal to zero and satisfy the constraint. Therefore, the determinant of the matrix of coefficients must equal zero. Thus $\begin{vmatrix} -2\lambda & 4 \\ 4 & -8\lambda \end{vmatrix} = 16\lambda^2 - 16 = 0$. This is because the system of equations is of the form

$$\begin{pmatrix} -2\lambda & 4 \\ 4 & -8\lambda \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

   If the matrix has an inverse, then the only solution would be $x = y = 0$ which as noted above can't happen. Therefore, $\lambda = \pm 1$. First suppose $\lambda = 1$. Then the first equation says $2y = x$. Pluggin this in to the constraint equation, $x^2 + x^2 = 4$ and so $x = \pm\sqrt{2}$. Therefore, $y = \pm\frac{\sqrt{2}}{2}$. This yields the dimensions of the largest rectangle to be $2\sqrt{2} \times \sqrt{2}$. You can check all the other cases and see you get the same thing in the other cases as well.

3. Maximize $2x + y$ subject to the condition that $\frac{x^2}{4} + y^2 \leq 1$.

   The maximum of this function clearly exists because of the extreme value theorem since the condition defines a closed and bounded set in $\mathbb{R}^2$. However, this function does not achieve its maximum on the interior of the given ellipse defined by $\frac{x^2}{4}+y^2 \leq 1$ because the gradient of the function which is to be maximized is never equal to zero. Therefore, this function must achieve its maximum on the set $\frac{x^2}{4} + y^2 = 1$. Thus you want to maximuze $2x + y$ subject to $\frac{x^2}{4} + y^2 = 1$. This is just like Problem 1. You can finish this.

4. Find the points on $y^2x = 16$ which are closest to $(0, 0)$.

   You want to minimize $x^2 + y^2$ subject to $y^2x - 16$. Of course you really want to minimize $\sqrt{x^2 + y^2}$ but the ordered pair which minimized $x^2 + y^2$ is the same as the ordered pair which minimize $\sqrt{x^2 + y^2}$ so it is pointless to drag around the square root. The Lagrangian is $x^2 + y^2 - \lambda\left(y^2x - 16\right)$. Differentiating with respect to x and y gives the equations $2x - \lambda y^2 = 0$ and $2y - 2\lambda yx = 0$. Neither x nor y can equal zero and solve the constraint. Therefore, the second equation implies $\lambda x = 1$. Hence $\lambda = \frac{1}{x} = \frac{2x}{y^2}$. Therefore, $2x^2 = y^2$ and so $2x^3 = 16$ and so $x = 2$. Therefore, $y = \pm 2\sqrt{2}$.

The points are $\left(2, 2\sqrt{2}\right)$ and $\left(2, -2\sqrt{2}\right)$. They both give the same answer. Note how ad hoc these procedures are. I can't give you a simple strategy for solving these systems of nonlinear equations by algebra because there is none. Sometimes nothing you do will work.

5. Find points on $xy = 1$ farthest from $(0,0)$ if any exist. If none exist, tell why. What does this say about the method of Lagrange multipliers?

   If you graph $xy = 1$ you see there is no farthest point. However, there is a closest point and the method of Lagrange multipliers will find this closest point. This shows that the answer you get has to be carefully considered to determine whether you have a maximum or a minimum or perhaps neither.

6. A curve is formed from the intersection of the plane, $2x + y + z = 3$ and the cylinder $x^2 + y^2 = 4$. Find the point on this curve which is closest to $(0,0,0)$.

   You want to maximize $x^2 + y^2 + z^2$ subject to the two constraints $2x + y + z = 3$ and $x^2 + y^2 = 4$. This means the Lagrangian will have two multipliers.

$$L = x^2 + y^2 + z^2 - \lambda\left(2x + y + z - 3\right) - \mu\left(x^2 + y^2 - 4\right)$$

   Then this yields the equations $2x - 2\lambda - 2\mu x = 0$, $2y - \lambda - 2\mu y$, and $2z - \lambda = 0$. The last equation says $\lambda = 2z$ and so I will replace λ with $2z$ where ever it occurs. This yields

$$x - 2z - \mu x = 0, 2y - 2z - 2\mu y = 0.$$

   This shows $x\left(1-\mu\right) = 2y\left(1-\mu\right)$. First suppose $\mu = 1$. Then from the above equations, $z = 0$ and so the two constraints reduce to $2y + x = 3$ and $x^2 + y^2 = 4$ and $2y + x = 3$. The solutions are $\left(\frac{3}{5} - \frac{2}{5}\sqrt{11}, \frac{6}{5} + \frac{1}{5}\sqrt{11}, 0\right), \left(\frac{3}{5} + \frac{2}{5}\sqrt{11}, \frac{6}{5} - \frac{1}{5}\sqrt{11}, 0\right)$. The other case is that $\mu \neq 1$ in which case $x = 2y$ and the second constraint yields that $y = \pm\frac{2}{\sqrt{5}}$ and $x = \pm\frac{4}{\sqrt{5}}$. Now from the first constraint, $z = -2\sqrt{5} + 3$ in the case where $y = \frac{2}{\sqrt{5}}$ and $z = 2\sqrt{5} + 3$ in the other case. This yields the points $\left(\frac{4}{\sqrt{5}}, \frac{2}{\sqrt{5}}, -2\sqrt{5} + 3\right)$ and $\left(-\frac{4}{\sqrt{5}}, -\frac{2}{\sqrt{5}}, 2\sqrt{5} + 3\right)$. This appears to have exhausted all the possibilities and so it is now just a matter of seeing which of these points gives the best answer. An answer exists because of the extreme value theorem. After all, this constraint set is closed and bounded. The first candidate listed above yields for the answer $\left(\frac{3}{5} - \frac{2}{5}\sqrt{11}\right)^2 + \left(\frac{6}{5} + \frac{1}{5}\sqrt{11}\right)^2 = 4$. The second candidate listed above yields $\left(\frac{3}{5} + \frac{2}{5}\sqrt{11}\right)^2 + \left(\frac{6}{5} - \frac{1}{5}\sqrt{11}\right)^2 = 4$ also. Thus these two give equally good results. Now consider the last two candidates. $\left(\frac{4}{\sqrt{5}}\right)^2 + \left(\frac{2}{\sqrt{5}}\right)^2 + \left(-2\sqrt{5} + 3\right)^2 = 4 + \left(-2\sqrt{5} + 3\right)^2$ which is larger than 4. Finally the last candidate yields $\left(-\frac{4}{\sqrt{5}}\right)^2 + \left(-\frac{2}{\sqrt{5}}\right)^2 + \left(2\sqrt{5} + 3\right)^2 = 4 + \left(2\sqrt{5} + 3\right)^2$ also larger than 4. Therefore, there are two points on the curve of intersection which are closest to the origin, $\left(\frac{3}{5} - \frac{2}{5}\sqrt{11}, \frac{6}{5} + \frac{1}{5}\sqrt{11}, 0\right)$ and $\left(\frac{3}{5} + \frac{2}{5}\sqrt{11}, \frac{6}{5} - \frac{1}{5}\sqrt{11}, 0\right)$. Both are a distance of 4 from the origin.

7. Here are two lines. $\mathbf{x} = \left(1 + 2t, 2 + t, 3 + t\right)^T$ and $\mathbf{x} = \left(2 + s, 1 + 2s, 1 + 3s\right)^T$. 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.

   Hint: Do you need to use Lagrange multipliers for this?

8. Find the point on $x^2 + y^2 + z^2 = 1$ closest to the plane $x + y + z = 10$.

   You want to minimize $(x-a)^2+(y-b)^2+(z-c)^2$ subject to the constraints $a+b+c = 10$ and $x^2 + y^2 + z^2 = 1$. There seem to be a lot of variables in this problem, 6 in all. Start taking derivatives and hope for a miracle. This yields $2(x-a) - 2\mu x = 0, 2(y-b) - 2\mu y = 0, 2(z-c) - 2\mu z = 0$. Also, taking derivatives with respect to a, b, and c you obtain $2(x-a) + \lambda = 0, 2(y-b) + \lambda = 0, 2(z-c) + \lambda = 0$. Comparing the first equations in each list, you see $\lambda = 2\mu x$ and then comparing the second two equations in each list, $\lambda = 2\mu y$ and similarly, $\lambda = 2\mu z$. Therefore, if $\mu \neq 0$, it must follow that $x = y = z$. Now you can see by sketching a rough graph that the answer you want has each of x, y, and z nonnegative. Therefore, using the constraint for these variables, the point desired is $\left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$ which you could probably see was the answer from the sketch. However, this could be made more difficult rather easily such that the sketch won't help but Lagrange multipliers will.

The Derivative Of Vector Valued Functions, What Is The Derivative?*

If you are going to do this stuff, you might as well do it right and include the case of vector valued functions. You know everything about matrices at this point for it to all make perfect sense. Therefore, I think it is well worth your time to read this although you are unlikely to see it on a test. It is not any harder than what has been presented. It also tells you what the derivative is. This is essential information if you are going to understand Newton's method for nonlinear systems. It is also essential if you want to read a really good book on continuum mechanics and is needed in many other physical and engineering applications. Also included is a proof of the second derivative test.

Recall the following definition.

Definition 21.0.7 *A function, T which maps $\mathbb{R}^n$ to $\mathbb{R}^p$ is called a linear transformation if for every pair of scalars, a, b and vectors, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, it follows that $T(a\mathbf{x} + b\mathbf{y}) = aT(\mathbf{x}) + bT(\mathbf{y})$.*

Recall that from the properties of matrix multiplication, it follows that if A is an $n \times p$ matrix, and if $\mathbf{x}, \mathbf{y}$ are vectors in $\mathbb{R}^n$, then $A(a\mathbf{x} + b\mathbf{y}) = aA(\mathbf{x}) + bA(\mathbf{y})$. Thus you can define a linear transformation by multiplying by a matrix. Of course the simplest example is that of a 1×1 matrix or number. You can think of the number 3 as a linear transformation, T mapping $\mathbb{R}$ to $\mathbb{R}$ according to the rule $Tx = 3x$. It satisfies the properties needed for a linear transformation because $3(ax + by) = a3x + b3y = aTx + bTy$. The case of the derivative of a scalar valued function of one variable is of this sort. You get a number for the derivative. However, you can think of this number as a linear transformation. Of course it is not worth the fuss to do so for a function of one variable but this is the way you must think of it for a function of n variables.

Definition 21.0.8 *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 $\mathbf{h} = (h_1 \cdots, h_n)^T$,*

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) = \mathbf{f}(\mathbf{x}) + \sum_{i=1}^{n} \mathbf{v}_i h_i + \mathbf{o}(\mathbf{h}). \tag{21.1}$$

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 21.0.9 *Suppose* $\mathbf{f}$ *is as given above in 21.1. 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 21.1 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

$$\mathbf{f}(\mathbf{x}+\mathbf{v}) = \mathbf{f}(\mathbf{x}) + J(\mathbf{x})\mathbf{v} + \mathbf{o}(\mathbf{v}) \tag{21.2}$$

and to reiterate, the linear transformation which results by multiplication by this $q \times p$ matrix is known as the derivative.

Sometimes x, y, z is written 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 21.0.10 *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, $f(x) = ax$.

21.1 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 differentiable from the definition. For example, you may have seen the following important example in one variable calculus.

Example 21.1.1 *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 21.1.2 *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

$$\mathbf{f}(1+h_1, 2+h_2) - \mathbf{f}(1,2)$$

$$
\begin{aligned}
&= \begin{pmatrix} (1+h_1)^2(2+h_2)+(2+h_2)^2 \\ (2+h_2)^3(1+h_1) \end{pmatrix} - \begin{pmatrix} 6 \\ 8 \end{pmatrix} \\
&= \begin{pmatrix} 5h_2+4h_1+2h_1h_2+2h_1^2+h_1^2h_2+h_2^2 \\ 8h_1+12h_2+12h_1h_2+6h_2^2+6h_2^2h_1+h_2^3+h_2^3h_1 \end{pmatrix} \\
&= \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \begin{pmatrix} 2h_1h_2+2h_1^2+h_1^2h_2+h_2^2 \\ 12h_1h_2+6h_2^2+6h_2^2h_1+h_2^3+h_2^3h_1 \end{pmatrix} \\
&= \begin{pmatrix} 4 & 5 \\ 8 & 12 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \mathbf{o}(\mathbf{h}).
\end{aligned}
$$

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 21.1.3 *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 21.0.9 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 theorem.

Theorem 21.1.4 *Suppose $\mathbf{f} : U \to \mathbb{R}^p$ where U is an open set in $\mathbb{R}^n$. Suppose also that all partial derivatives of $\mathbf{f}$ exist on U and are continuous. Then $\mathbf{f}$ is differentiable at every point of U.*

Proof: If you fix all the variables but one, you can apply the fundamental theorem of calculus as follows.

$$\mathbf{f}(\mathbf{x}+v_k\mathbf{e}_k) - \mathbf{f}(\mathbf{x}) = \int_0^1 \frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x}+tv_k\mathbf{e}_k)\, v_k dt. \tag{21.3}$$

Here is why. Let $\mathbf{h}(t) = \mathbf{f}(\mathbf{x}+tv_k\mathbf{e}_k)$. Then

$$\frac{\mathbf{h}(t+h)-\mathbf{h}(t)}{h} = \frac{\mathbf{f}(\mathbf{x}+tv_k\mathbf{e}_k+hv_k\mathbf{e}_k)-\mathbf{f}(\mathbf{x}+tv_k\mathbf{e}_k)}{hv_k}v_k$$

and so, taking the limit as $h \to 0$ yields

$$\mathbf{h}'(t) = \frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x}+tv_k\mathbf{e}_k)\, v_k$$

Therefore,

$$\mathbf{f}(\mathbf{x}+v_k\mathbf{e}_k) - \mathbf{f}(\mathbf{x}) = \mathbf{h}(1) - \mathbf{h}(0) = \int_0^1 \mathbf{h}'(t)\, dt = \int_0^1 \frac{\partial \mathbf{f}}{\partial x_k}(\mathbf{x}+tv_k\mathbf{e}_k)\, v_k dt.$$

Now I will use this observation to prove the theorem. Let $\mathbf{v} = (v_1, \cdots, v_n)$ with $|\mathbf{v}|$ sufficiently small. Thus $\mathbf{v} = \sum_{k=1}^n v_k\mathbf{e}_k$. For the purposes of this argument, define

$$\sum_{k=n+1}^{n} v_k\mathbf{e}_k \equiv \mathbf{0}.$$

Then with this convention,

$$\begin{aligned}
\mathbf{f}(\mathbf{x}+\mathbf{v})-\mathbf{f}(\mathbf{x}) &= \sum_{i=1}^{n}\left(\mathbf{f}\left(\mathbf{x}+\sum_{k=i}^{n}v_k\mathbf{e}_k\right)-\mathbf{f}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k\right)\right)\\
&= \sum_{i=1}^{n}\int_0^1\frac{\partial\mathbf{f}}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)v_i\,dt\\
&= \sum_{i=1}^{n}\int_0^1\left(\frac{\partial\mathbf{f}}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)v_i-\frac{\partial\mathbf{f}}{\partial x_i}(\mathbf{x})\,v_i\right)dt\\
&\quad+\sum_{i=1}^{n}\int_0^1\frac{\partial\mathbf{f}}{\partial x_i}(\mathbf{x})\,v_i\,dt\\
&= \sum_{i=1}^{n}\frac{\partial\mathbf{f}}{\partial x_i}(\mathbf{x})\,v_i\\
&\quad+\int_0^1\sum_{i=1}^{n}\left(\frac{\partial\mathbf{f}}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)-\frac{\partial\mathbf{f}}{\partial x_i}(\mathbf{x})\right)v_i\,dt\\
&= \sum_{i=1}^{n}\frac{\partial\mathbf{f}}{\partial x_i}(\mathbf{x})\,v_i+o(\mathbf{v})
\end{aligned}$$

and this shows $\mathbf{f}$ is differentiable at $\mathbf{x}$.

Some explanation of the step to the last line is in order. The messy thing at the end is $o(\mathbf{v})$ because of the continuity of the partial derivatives. In fact, from the Cauchy Schwarz inequality,

$$\begin{aligned}
&\int_0^1\sum_{i=1}^{n}\left(\frac{\partial\mathbf{f}}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)-\frac{\partial\mathbf{f}}{\partial x_i}(\mathbf{x})\right)v_i\,dt\\
\leq\ &\int_0^1\left(\sum_{i=1}^{n}\left|\frac{\partial\mathbf{f}}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)-\frac{\partial\mathbf{f}}{\partial x_i}(\mathbf{x})\right|^2\right)^{1/2}dt\left(\sum_{i=1}^{n}v_i^2\right)^{1/2}\\
=\ &\int_0^1\left(\sum_{i=1}^{n}\left|\frac{\partial\mathbf{f}}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)-\frac{\partial\mathbf{f}}{\partial x_i}(\mathbf{x})\right|^2\right)^{1/2}dt\,|\mathbf{v}|
\end{aligned}$$

Thus, dividing by $|\mathbf{v}|$ and taking a limit as $|\mathbf{v}|\to 0$, the quotient is nothing but

$$\int_0^1\left(\sum_{i=1}^{n}\left|\frac{\partial\mathbf{f}}{\partial x_i}\left(\mathbf{x}+\sum_{k=i+1}^{n}v_k\mathbf{e}_k+tv_i\mathbf{e}_i\right)-\frac{\partial\mathbf{f}}{\partial x_i}(\mathbf{x})\right|^2\right)^{1/2}dt$$

which converges to 0 due to continuity of the partial derivatives of $\mathbf{f}$. This proves the theorem.

Here is an example to illustrate.

Example 21.1.5 *Let* $\mathbf{f}(x,y)=\begin{pmatrix}x^2y+y^2\\ y^3x\end{pmatrix}$. *Find* $D\mathbf{f}(x,y)$.

From Theorem 21.1.4 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 21.1.6 *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

$$\begin{pmatrix} 2x_1x_2 & x_1^2+2x_2 & 0 \\ x_2 & x_1 & 1 \\ x_2x_3\cos(x_1x_2x_3) & x_1x_3\cos(x_1x_2x_3) & x_1x_2\cos(x_1x_2x_3) \end{pmatrix}$$

Example 21.1.7 *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 $(1,2,3)$,

$$\begin{aligned} f(x,y,z) &\approx f(1,2,3) + 1(x-1) + 2(y-2) + 6(z-3) \\ &= 11 + 1(x-1) + 2(y-2) + 6(z-3) = -12 + x + 2y + 6z \end{aligned}$$

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 21.1.8 *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,*

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)| \leq (Cp+1)|\mathbf{x}-\mathbf{x}_0| \tag{21.4}$$

Proof: Suppose $\mathbf{f}$ is differentiable. Since $\mathbf{o}(\mathbf{v})$ satisfies 19.1, 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,

$$\left|\sum_{i=1}^p \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x}_0)(x_i - x_{0i})\right| \leq C\sum_{i=1}^p |x_i - x_{0i}| \leq Cp|\mathbf{x}-\mathbf{x}_0|$$

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 21.4. 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 $\mathbf{x}_0$.

21.2 The Chain Rule*

21.2.1 The Chain Rule For Functions Of One Variable*

First recall the chain rule for a function of one variable. Consider the following picture.

$$I \xrightarrow{g} J \xrightarrow{f} \mathbb{R}$$

Here I and J are open intervals and it is assumed that $g(I) \subseteq J$. The chain rule says that if $f'(g(x))$ exists and $g'(x)$ exists for $x \in I$, then the composition, $f \circ g$ also has a derivative at x and

$$(f \circ g)'(x) = f'(g(x)) g'(x).$$

Recall that $f \circ g$ is the name of the function defined by $f \circ g(x) \equiv f(g(x))$. In the notation of this chapter, the chain rule is written as

$$Df(g(x)) Dg(x) = D(f \circ g)(x). \tag{21.5}$$

21.2.2 The Chain Rule For Functions Of Many Variables*

Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^p$ be open sets and let $\mathbf{f}$ be a function defined on V having values in $\mathbb{R}^q$ while $\mathbf{g}$ is a function defined on U such that $\mathbf{g}(U) \subseteq V$ as in the following picture.

$$U \xrightarrow{\mathbf{g}} V \xrightarrow{\mathbf{f}} \mathbb{R}^q$$

The chain rule says that if the linear transformations (matrices) on the left in 21.5 both exist then the same formula holds in this more general case. Thus

$$D\mathbf{f}(\mathbf{g}(\mathbf{x})) D\mathbf{g}(\mathbf{x}) = D(\mathbf{f} \circ \mathbf{g})(\mathbf{x})$$

Note this all makes sense because $D\mathbf{f}(\mathbf{g}(\mathbf{x}))$ is a $q \times p$ matrix and $D\mathbf{g}(\mathbf{x})$ is a $p \times n$ matrix. Remember it is all right to do $(q \times p)(p \times n)$. The middle numbers match. More precisely,

Theorem 21.2.1 *(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,*

$$D(\mathbf{f} \circ \mathbf{g})(\mathbf{x}) = D\mathbf{f}(\mathbf{g}(\mathbf{x})) D\mathbf{g}(\mathbf{x}). \tag{21.6}$$

In particular,

$$\frac{\partial (\mathbf{f} \circ \mathbf{g})(\mathbf{x})}{\partial x_j} = \sum_{i=1}^{p} \frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial y_i} \frac{\partial g_i(\mathbf{x})}{\partial x_j}. \tag{21.7}$$

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

$$\frac{\partial \mathbf{z}}{\partial y_i} \frac{\partial y_i}{\partial x_k} = \frac{\partial \mathbf{z}}{\partial x_k}. \tag{21.8}$$

Remember there is a sum on the repeated index. In particular, for each index, r,

$$\frac{\partial z_r}{\partial y_i} \frac{\partial y_i}{\partial x_k} = \frac{\partial z_r}{\partial x_k}.$$

The proof of this major theorem will be given at the end of this section. It will include the chain rule for functions of one variable as a special case. First here are some examples.

Example 21.2.2 *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 21.8,

$$\frac{\partial z}{\partial t} = \frac{\partial z}{\partial u}\frac{\partial u}{\partial t} + \frac{\partial z}{\partial v}\frac{\partial v}{\partial t} = v\cos(uv)\sin(x) + us\cos(uv).$$

Here $y_1 = u, y_2 = v, t = x_k$. 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 21.2.3 *Let* $w = f(u_1,u_2) = u_2\sin(u_1)$ *and* $u_1 = x^2y + z, u_2 = \sin(xy)$. *Find* $\frac{\partial w}{\partial x}$, $\frac{\partial w}{\partial y}$, *and* $\frac{\partial w}{\partial z}$.

The derivative of f is of the form (w_x, w_y, w_z) and so it suffices to find the derivative of f using the chain rule. You need to find $Df(u_1,u_2)\,D\mathbf{g}(x,y,z)$ where $\mathbf{g}(x,y) = \begin{pmatrix} x^2y + z \\ \sin(xy) \end{pmatrix}$. Then $D\mathbf{g}(x,y,z) = \begin{pmatrix} 2xy & x^2 & 1 \\ y\cos(xy) & x\cos(xy) & 0 \end{pmatrix}$. Also $Df(u_1,u_2) = (u_2\cos(u_1), \sin(u_1))$. Therefore, the derivative is

$$Df(u_1,u_2)\,D\mathbf{g}(x,y,z) = (u_2\cos(u_1), \sin(u_1)) \begin{pmatrix} 2xy & x^2 & 1 \\ y\cos(xy) & x\cos(xy) & 0 \end{pmatrix}$$

$$= \left(2u_2(\cos u_1)\,xy + (\sin u_1)\,y\cos xy,\, u_2(\cos u_1)\,x^2 + (\sin u_1)\,x\cos xy,\, u_2\cos u_1\right) = (w_x, w_y, w_z)$$

Thus $\frac{\partial w}{\partial x} = 2u_2(\cos u_1)\,xy + (\sin u_1)\,y\cos xy = 2(\sin(xy))\left(\cos\left(x^2y+z\right)\right)xy + \left(\sin\left(x^2y+z\right)\right)y\cos xy$. Similarly, you can find the other partial derivatives of w in terms of substituting in for u_1 and u_2 in the above. Note

$$\frac{\partial w}{\partial x} = \frac{\partial w}{\partial u_1}\frac{\partial u_1}{\partial x} + \frac{\partial w}{\partial u_2}\frac{\partial u_2}{\partial x}.$$

In fact, in general if you have $w = f(u_1,u_2)$ and $\mathbf{g}(x,y,z) = \begin{pmatrix} u_1(x,y,z) \\ u_2(x,y,z) \end{pmatrix}$, then $D(f\circ\mathbf{g})(x,y,z)$ is of the form

$$\begin{pmatrix} w_{u_1} & w_{u_2} \end{pmatrix} \begin{pmatrix} u_{1x} & u_{1y} & u_{1z} \\ u_{2x} & u_{2y} & u_{2z} \end{pmatrix}$$
$$= \begin{pmatrix} w_{u_1}u_x + w_{u_2}u_{2x} & w_{u_1}u_y + w_{u_2}u_{2y} & w_{u_1}u_z + w_{u_2}u_{2z} \end{pmatrix}.$$

Example 21.2.4 *Let* $w = f(u_1,u_2,u_3) = u_1^2 + u_3 + u_2$ *and* $\mathbf{g}(x,y,z) = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} x + 2yz \\ x^2 + y \\ z^2 + x \end{pmatrix}$. *Find* $\frac{\partial w}{\partial x}$ *and* $\frac{\partial w}{\partial z}$.

By the chain rule,

$$(w_x, w_y, w_z) = \begin{pmatrix} w_{u_1} & w_{u_2} & w_{u_3} \end{pmatrix} \begin{pmatrix} u_{1x} & u_{1y} & u_{1z} \\ u_{2x} & u_{2y} & u_{2z} \\ u_{3x} & u_{3y} & u_{3z} \end{pmatrix}$$

$$= \begin{pmatrix} w_{u_1}u_{1x} + w_{u_2}u_{2x} + w_{u_3}u_{3x} & w_{u_1}u_{1y} + w_{u_2}u_{2y} + w_{u_3}u_{3y} & w_{u_1}u_{1z} + w_{u_2}u_{2z} + w_{u_3}u_{3z} \end{pmatrix}$$

Note the pattern.

$$\begin{aligned} w_x &= w_{u_1}u_{1x} + w_{u_2}u_{2x} + w_{u_3}u_{3x}, \\ w_y &= w_{u_1}u_{1y} + w_{u_2}u_{2y} + w_{u_3}u_{3y}, \\ w_z &= w_{u_1}u_{1z} + w_{u_2}u_{2z} + w_{u_3}u_{3z}. \end{aligned}$$

Therefore,

$$w_x = 2u_1(1) + 1(2x) + 1(1) = 2(x + 2yz) + 2x + 1 = 4x + 4yz + 1$$

and

$$w_z = 2u_1(2y) + 1(0) + 1(2z) = 4(x + 2yz)y + 2z = 4yx + 8y^2z + 2z.$$

Of course to find all the partial derivatives at once, you just use the chain rule. Thus you would get

$$\begin{aligned} \begin{pmatrix} w_x & w_y & w_z \end{pmatrix} &= \begin{pmatrix} 2u_1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2z & 2y \\ 2x & 1 & 0 \\ 1 & 0 & 2z \end{pmatrix} \\ &= \begin{pmatrix} 2u_1 + 2x + 1 & 4u_1z + 1 & 4u_1y + 2z \end{pmatrix} \\ &= \begin{pmatrix} 4x + 4yz + 1 & 4zx + 8yz^2 + 1 & 4yx + 8y^2z + 2z \end{pmatrix} \end{aligned}$$

Example 21.2.5 *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.

In different notation, let $\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \mathbf{f}(u_1,u_2) = \begin{pmatrix} u_1^2 + u_2 \\ \sin(u_2) + u_1 \end{pmatrix}$. Then

$$\frac{\partial z_1}{\partial x_1} = \frac{\partial z_1}{\partial u_1}\frac{\partial u_1}{\partial x_1} + \frac{\partial z_1}{\partial u_2}\frac{\partial u_2}{\partial x_1} = 2u_1x_2 + 1 = 2(x_1x_2 + x_3)x_2 + 1.$$

Example 21.2.6 *Let* $\mathbf{f}(u_1,u_2,u_3) = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} u_1^2 + u_2u_3 \\ u_1^2 + u_2^3 \\ \ln\left(1+u_3^2\right) \end{pmatrix}$ *and let* $\mathbf{g}(x_1,x_2,x_3,x_4) = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} x_1 + x_2^2 + \sin(x_3) + \cos(x_4) \\ x_4^2 - x_1 \\ x_3^2 + x_4 \end{pmatrix}$. *Find* $(\mathbf{f}\circ\mathbf{g})'(\mathbf{x})$.

$$D\mathbf{f}(\mathbf{u}) = \begin{pmatrix} 2u_1 & u_3 & u_2 \\ 2u_1 & 3u_2^2 & 0 \\ 0 & 0 & \frac{2u_3}{\left(1+u_3^2\right)} \end{pmatrix}$$

Similarly,

$$D\mathbf{g}(\mathbf{x}) = \begin{pmatrix} 1 & 2x_2 & \cos(x_3) & -\sin(x_4) \\ -1 & 0 & 0 & 2x_4 \\ 0 & 0 & 2x_3 & 1 \end{pmatrix}.$$

Then by the chain rule, $D(\mathbf{f}\circ\mathbf{g})(\mathbf{x}) = D\mathbf{f}(\mathbf{u})D\mathbf{g}(\mathbf{x})$ where $\mathbf{u} = \mathbf{g}(\mathbf{x})$ as described above. Thus $D(\mathbf{f}\circ\mathbf{g})(\mathbf{x}) =$

$$\begin{pmatrix} 2u_1 & u_3 & u_2 \\ 2u_1 & 3u_2^2 & 0 \\ 0 & 0 & \frac{2u_3}{\left(1+u_3^2\right)} \end{pmatrix} \begin{pmatrix} 1 & 2x_2 & \cos(x_3) & -\sin(x_4) \\ -1 & 0 & 0 & 2x_4 \\ 0 & 0 & 2x_3 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 2u_1 - u_3 & 4u_1x_2 & 2u_1\cos x_3 + 2u_2x_3 & -2u_1\sin x_4 + 2u_3x_4 + u_2 \\ 2u_1 - 3u_2^2 & 4u_1x_2 & 2u_1\cos x_3 & -2u_1\sin x_4 + 6u_2^2x_4 \\ 0 & 0 & 4\frac{u_3}{1+u_3^2}x_3 & 2\frac{u_3}{1+u_3^2} \end{pmatrix} \tag{21.9}$$

where each u_i is given by the above formulas. Thus $\frac{\partial z_1}{\partial x_1}$ equals

$$\begin{aligned} 2u_1 - u_3 &= 2\left(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)\right) - \left(x_3^2 + x_4\right) \\ &= 2x_1 + 2x_2^2 + 2\sin x_3 + 2\cos x_4 - x_3^2 - x_4. \end{aligned}$$

while $\frac{\partial z_2}{\partial x_4}$ equals

$$-2u_1\sin x_4 + 6u_2^2x_4 = -2\left(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)\right)\sin(x_4) + 6\left(x_4^2 - x_1\right)^2 x_4.$$

If you wanted $\frac{\partial \mathbf{z}}{\partial x_2}$ it would be the second column of the above matrix in 21.9. Thus $\frac{\partial \mathbf{z}}{\partial x_2}$ equals

$$\begin{pmatrix} \frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_2}{\partial x_2} \\ \frac{\partial z_3}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 4u_1x_2 \\ 4u_1x_2 \\ 0 \end{pmatrix} = \begin{pmatrix} 4\left(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)\right)x_2 \\ 4\left(x_1 + x_2^2 + \sin(x_3) + \cos(x_4)\right)x_2 \\ 0 \end{pmatrix}.$$

I hope that by now it is clear that all the information you could desire about various partial derivatives is available and it all reduces to matrix multiplication and the consideration of entries of the matrix obtained by multiplying the two derivatives.

21.2.3 The Derivative Of The Inverse Function*

Example 21.2.7 *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 21.0.10 on Page 373 $DI = I$. 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

$$\delta_{ij} = \frac{\partial x_i}{\partial y_k}(\mathbf{f}(\mathbf{x}))\frac{\partial y_k}{\partial x_j}(\mathbf{x}).$$

21.2.4 Acceleration In Spherical Coordinates*

This is an interesting example which can be done with more elegance in a more general setting. However, the more general approach also depends on the chain rule and this is what it is all about, giving examples of the use of the chain rule. Read it if it interests you.

Example 21.2.8 *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 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

$$\mathbf{r}'(t) = \rho'(t)\,\mathbf{e}_\rho(t) + \rho(t)\,(\mathbf{e}_\rho(t))'. \tag{21.10}$$

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 21.10,

$$\mathbf{r}' = \rho'\mathbf{e}_\rho + \frac{d\phi}{dt}\mathbf{e}_\phi + \frac{d\theta}{dt}\mathbf{e}_\theta. \tag{21.11}$$

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 21.11 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 21.11,

$$\mathbf{r}'' = \rho''\mathbf{e}_\rho + \phi''\mathbf{e}_\phi + \theta''\mathbf{e}_\theta + \rho'(\mathbf{e}_\rho)' + \phi'(\mathbf{e}_\phi)' + \theta'(\mathbf{e}_\theta)'$$

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

$$\begin{aligned}\mathbf{r}'' &= \left(\rho'' - \rho\left(\phi'\right)^2 - \rho\left(\theta'\right)^2\sin^2(\phi)\right)\mathbf{e}_\rho + \left(\phi'' + \frac{2\rho'\phi'}{\rho} - \left(\theta'\right)^2\cos\phi\sin\phi\right)\mathbf{e}_\phi + \\ &+ \left(\theta'' + \frac{2\theta'\rho'}{\rho} + 2\phi'\theta'\cot(\phi)\right)\mathbf{e}_\theta.\end{aligned}$$

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.

21.3 Proof Of 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 21.3.1 *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

$$\lim_{\mathbf{v}\to\mathbf{0}} \frac{|\mathbf{o}(\mathbf{g}(\mathbf{x}+\mathbf{v}) - \mathbf{g}(\mathbf{x}))|}{|\mathbf{v}|} = 0. \tag{21.12}$$

From Lemma 21.1.8, there exists $\delta > 0$ such that if $|\mathbf{v}| < \delta$, then

$$|\mathbf{g}(\mathbf{x}+\mathbf{v}) - \mathbf{g}(\mathbf{x})| \leq (Cn+1)|\mathbf{v}|. \tag{21.13}$$

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

$$|\mathbf{o}(\mathbf{g}(\mathbf{x}+\mathbf{v}) - \mathbf{g}(\mathbf{x}))| < \left(\frac{\varepsilon}{Cn+1}\right)|\mathbf{g}(\mathbf{x}+\mathbf{v}) - \mathbf{g}(\mathbf{x})| \tag{21.14}$$

[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.

Let $|\mathbf{v}| < \min\left(\delta, \frac{\eta}{Cn+1}\right)$. For such $\mathbf{v}$, $|\mathbf{g}(\mathbf{x}+\mathbf{v}) - \mathbf{g}(\mathbf{x})| \leq \eta$, 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 21.12. 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 21.3.2 *(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,*

$$D(\mathbf{f} \circ \mathbf{g})(\mathbf{x}) = D\mathbf{f}(\mathbf{g}(\mathbf{x}))\, D\mathbf{g}(\mathbf{x}). \tag{21.15}$$

In particular,

$$\frac{\partial(\mathbf{f} \circ \mathbf{g})(\mathbf{x})}{\partial x_j} = \sum_{i=1}^{p} \frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial y_i} \frac{\partial g_i(\mathbf{x})}{\partial x_j}. \tag{21.16}$$

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 21.3.1 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 21.16 because of Theorem 21.0.9 on Page 372. 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 21.15 follows from the definition of matrix multiplication.

21.4 Proof Of The Second Derivative Test*

Definition 21.4.1 *The matrix,* $\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})\right)$ *is called the Hessian matrix, denoted by* $H(\mathbf{x})$.

Now recall the Taylor formula with the Lagrange form of the remainder. Since most people don't pay any attention to this important topic when they take calculus, here is a statement and proof of this theorem.

Theorem 21.4.2 *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

$$f(x) - \left(f(c) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k + K(x-c)^{n+1}\right) = 0 \tag{21.17}$$

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

$$F(t) \equiv f(x) - \left(f(t) + \sum_{k=1}^{n} \frac{f^{(k)}(c)}{k!}(x-t)^k + K(x-t)^{n+1}\right).$$

The c in 21.17 got replaced by t.

Therefore, $F(c) = 0$ by the way K was chosen 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.

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 21.4.2 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

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x})\cdot\mathbf{v} + \frac{1}{2}\mathbf{v}^T H(\mathbf{x})\,\mathbf{v} +$$

$$\frac{1}{2}\left(\mathbf{v}^T\left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right) \tag{21.18}$$

where the last term satisfies

$$\lim_{|\mathbf{v}|\to 0} \frac{1}{2}\frac{\left(\mathbf{v}^T\left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right)}{|\mathbf{v}|^2} = 0 \tag{21.19}$$

because of the continuity of the entries of $H(\mathbf{x})$.

Recall the following important theorem from linear algebra.

Theorem 21.4.3 *If A is a real symmetric matrix, then A is Hermitian and there exists a real unitary matrix, U such that $U^T AU = D$ where D is a diagonal matrix. In particular, it has all real eigenvalues and an orthonormal basis of eigenvectors.*

Theorem 21.4.4 *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 21.18 implies

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \frac{1}{2}\mathbf{v}^T H(\mathbf{x})\,\mathbf{v} + \frac{1}{2}\left(\mathbf{v}^T\left(H(\mathbf{x}+t\mathbf{v}) - H(\mathbf{x})\right)\mathbf{v}\right) \tag{21.20}$$

and by continuity of the second derivatives, these mixed second derivatives are equal and so $H(\mathbf{x})$ is a symmetric matrix . Thus, by Theorem 21.4.3 $H(\mathbf{x})$ 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

$$\begin{aligned}\mathbf{u}^T H(\mathbf{x})\,\mathbf{u} &= \left(\sum_{k=1}^n u_k \mathbf{v}_k^T\right) H(\mathbf{x}) \left(\sum_{j=1}^n u_j \mathbf{v}_j\right) \\ &= \sum_{k,j} u_k \mathbf{v}_k^T H(\mathbf{x})\,\mathbf{v}_j u_j \\ &= \sum_{j=1}^n u_j^2 \lambda_j \geq \delta^2 \sum_{j=1}^n u_j^2 = \delta^2 |\mathbf{u}|^2 .\end{aligned}$$

From 21.20 and 21.19, 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 21.20, 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.

Implicit Function Theorem*

The implicit function theorem is one of the greatest theorems in mathematics. There are many versions of this theorem which are of far greater generality than the one given here. The proof given here is like one found in one of Caratheodory's books on the calculus of variations. It is not as elegant as some of the others which are based on a contraction mapping principle but it may be more accessible. However, it is an advanced topic. Don't waste your time with it unless you have first read and understood the material on rank and determinants found in the chapter on the mathematical theory of determinants. You will also need to use the extreme value theorem for a function of n variables and the chain rule as well as everything about matrix multiplication.

Definition 22.0.5 *Suppose U is an open set in $\mathbb{R}^n \times \mathbb{R}^m$ and $(\mathbf{x},\mathbf{y})$ will denote a typical point of $\mathbb{R}^n \times \mathbb{R}^m$ with $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Let $\mathbf{f} : U \to \mathbb{R}^p$ be in $C^1(U)$. Then define*

$$D_1\mathbf{f}(\mathbf{x},\mathbf{y}) \equiv \begin{pmatrix} f_{1,x_1}(\mathbf{x},\mathbf{y}) & \cdots & f_{1,x_n}(\mathbf{x},\mathbf{y}) \\ \vdots & & \vdots \\ f_{p,x_1}(\mathbf{x},\mathbf{y}) & \cdots & f_{p,x_n}(\mathbf{x},\mathbf{y}) \end{pmatrix},$$

$$D_2\mathbf{f}(\mathbf{x},\mathbf{y}) \equiv \begin{pmatrix} f_{1,y_1}(\mathbf{x},\mathbf{y}) & \cdots & f_{1,y_m}(\mathbf{x},\mathbf{y}) \\ \vdots & & \vdots \\ f_{p,y_1}(\mathbf{x},\mathbf{y}) & \cdots & f_{p,y_m}(\mathbf{x},\mathbf{y}) \end{pmatrix}.$$

Thus $D\mathbf{f}(\mathbf{x},\mathbf{y})$ is a $p \times (n+m)$ matrix of the form

$$D\mathbf{f}(\mathbf{x},\mathbf{y}) = \left(\begin{array}{c|c} D_1\mathbf{f}(\mathbf{x},\mathbf{y}) & D_2\mathbf{f}(\mathbf{x},\mathbf{y}) \end{array} \right).$$

Note that $D_1\mathbf{f}(\mathbf{x},\mathbf{y})$ is an $p \times n$ matrix and $D_2\mathbf{f}(\mathbf{x},\mathbf{y})$ is a $p \times m$ matrix.

Theorem 22.0.6 *(implicit function theorem) Suppose U is an open set in $\mathbb{R}^n \times \mathbb{R}^m$. Let $\mathbf{f} : U \to \mathbb{R}^n$ be in $C^1(U)$ and suppose*

$$\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0) = \mathbf{0},\ D_1\mathbf{f}(\mathbf{x}_0, \mathbf{y}_0)^{-1} \text{ exists.} \tag{22.1}$$

Then there exist positive constants, δ, η, such that for every $\mathbf{y} \in B(\mathbf{y}_0, \eta)$ there exists a unique $\mathbf{x}(\mathbf{y}) \in B(\mathbf{x}_0, \delta)$ such that

$$\mathbf{f}(\mathbf{x}(\mathbf{y}), \mathbf{y}) = \mathbf{0}. \tag{22.2}$$

Furthermore, the mapping, $\mathbf{y} \to \mathbf{x}(\mathbf{y})$ is in $C^1(B(\mathbf{y}_0, \eta))$.

Proof: Let

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} f_1(\mathbf{x}, \mathbf{y}) \\ f_2(\mathbf{x}, \mathbf{y}) \\ \vdots \\ f_n(\mathbf{x}, \mathbf{y}) \end{pmatrix}.$$

Define for $(\mathbf{x}^1, \cdots, \mathbf{x}^n) \in \overline{B(\mathbf{x}_0, \delta)}^n$ and $\mathbf{y} \in B(\mathbf{y}_0, \eta)$ the following matrix.

$$J(\mathbf{x}^1, \cdots, \mathbf{x}^n, \mathbf{y}) \equiv \begin{pmatrix} f_{1,x_1}(\mathbf{x}^1, \mathbf{y}) & \cdots & f_{1,x_n}(\mathbf{x}^1, \mathbf{y}) \\ \vdots & & \vdots \\ f_{n,x_1}(\mathbf{x}^n, \mathbf{y}) & \cdots & f_{n,x_n}(\mathbf{x}^n, \mathbf{y}) \end{pmatrix}.$$

Then by the assumption of continuity of all the partial derivatives and the extreme value theorem, there exists $r > 0$ and $\delta_0, \eta_0 > 0$ such that if $\delta \leq \delta_0$ and $\eta \leq \eta_0$, it follows that for all $(\mathbf{x}^1, \cdots, \mathbf{x}^n) \in \overline{B(\mathbf{x}_0, \delta)}^n$ and $\mathbf{y} \in \overline{B(\mathbf{y}_0, \eta)}$,

$$\det\left(J(\mathbf{x}^1, \cdots, \mathbf{x}^n, \mathbf{y})\right) > r > 0. \tag{22.3}$$

and $\overline{B(\mathbf{x}_0, \delta_0)} \times \overline{B(\mathbf{y}_0, \eta_0)} \subseteq U$. By continuity of all the partial derivatives and the extreme value theorem, it can also be assumed there exists a constant, K such that for all $(\mathbf{x}, \mathbf{y}) \in \overline{B(\mathbf{x}_0, \delta_0)} \times \overline{B(\mathbf{y}_0, \eta_0)}$ and $i = 1, 2, \cdots, n$, the i^{th} row of $D_2\mathbf{f}(\mathbf{x}, \mathbf{y})$, given by $D_2 f_i(\mathbf{x}, \mathbf{y})$ satisfies

$$|D_2 f_i(\mathbf{x}, \mathbf{y})| < K, \tag{22.4}$$

and for all $(\mathbf{x}^1, \cdots, \mathbf{x}^n) \in \overline{B(\mathbf{x}_0, \delta_0)}^n$ and $\mathbf{y} \in \overline{B(\mathbf{y}_0, \eta_0)}$ the i^{th} row of the matrix, $J(\mathbf{x}^1, \cdots, \mathbf{x}^n, \mathbf{y})^{-1}$ which equals $\mathbf{e}_i^T \left(J(\mathbf{x}^1, \cdots, \mathbf{x}^n, \mathbf{y})^{-1}\right)$ satisfies

$$\left|\mathbf{e}_i^T \left(J(\mathbf{x}^1, \cdots, \mathbf{x}^n, \mathbf{y})^{-1}\right)\right| < K. \tag{22.5}$$

(Recall that $\mathbf{e}_i$ is the column vector consisting of all zeros except for a 1 in the i^{th} position.)

To begin with it is shown that for a given $\mathbf{y} \in B(\mathbf{y}_0, \eta)$ there is at most one $\mathbf{x} \in B(\mathbf{x}_0, \delta)$ such that $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$.

Pick $\mathbf{y} \in B(\mathbf{y}_0, \eta)$ and suppose there exist $\mathbf{x}, \mathbf{z} \in \overline{B(\mathbf{x}_0, \delta)}$ such that $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{f}(\mathbf{z}, \mathbf{y}) = \mathbf{0}$. Consider f_i and let

$$h(t) \equiv f_i(\mathbf{x} + t(\mathbf{z} - \mathbf{x}), \mathbf{y}).$$

Then $h(1) = h(0)$ and so by the mean value theorem, $h'(t_i) = 0$ for some $t_i \in (0, 1)$. Therefore, from the chain rule and for this value of t_i,

$$h'(t_i) = Df_i(\mathbf{x} + t_i(\mathbf{z} - \mathbf{x}), \mathbf{y})(\mathbf{z} - \mathbf{x}) = 0. \tag{22.6}$$

Then denote by $\mathbf{x}^i$ the vector, $\mathbf{x}+t_i(\mathbf{z}-\mathbf{x})$. It follows from 22.6 that

$$J\left(\mathbf{x}^1,\cdots,\mathbf{x}^n,\mathbf{y}\right)(\mathbf{z}-\mathbf{x})=\mathbf{0}$$

and so from 22.3 $\mathbf{z}-\mathbf{x}=\mathbf{0}$. (The matrix, in the above is invertible since its determinant is nonzero.) Now it will be shown that if η is chosen sufficiently small, then for all $\mathbf{y}\in B(\mathbf{y}_0,\eta)$, there exists a unique $\mathbf{x}(\mathbf{y})\in B(\mathbf{x}_0,\delta)$ such that $\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})=\mathbf{0}$.

Claim: If η is small enough, then the function, $h_{\mathbf{y}}(\mathbf{x})\equiv|\mathbf{f}(\mathbf{x},\mathbf{y})|^2$ achieves its minimum value on $\overline{B(\mathbf{x}_0,\delta)}$ at a point of $B(\mathbf{x}_0,\delta)$. (The existence of a point in $\overline{B(\mathbf{x}_0,\delta)}$ at which $h_{\mathbf{y}}$ achieves its minimum follows from the extreme value theorem.)

Proof of claim: Suppose this is not the case. Then there exists a sequence $\eta_k\to 0$ and for some $\mathbf{y}_k$ having $|\mathbf{y}_k-\mathbf{y}_0|<\eta_k$, the minimum of $h_{\mathbf{y}_k}$ on $\overline{B(\mathbf{x}_0,\delta)}$ occurs on a point of $\overline{B(\mathbf{x}_0,\delta)}$, $\mathbf{x}_k$ such that $|\mathbf{x}_0-\mathbf{x}_k|=\delta$. Now taking a subsequence, still denoted by k, it can be assumed that $\mathbf{x}_k\to\mathbf{x}$ with $|\mathbf{x}-\mathbf{x}_0|=\delta$ and $\mathbf{y}_k\to\mathbf{y}_0$. This follows from the fact that $\left\{\mathbf{x}\in\overline{B(\mathbf{x}_0,\delta)}:|\mathbf{x}-\mathbf{x}_0|=\delta\right\}$ is a closed and bounded set and is therefore sequentially compact. Let $\varepsilon>0$. Then for k large enough, the continuity of $\mathbf{y}\to h_{\mathbf{y}}(\mathbf{x}_0)$ implies $h_{\mathbf{y}_k}(\mathbf{x}_0)<\varepsilon$ because $h_{\mathbf{y}_0}(\mathbf{x}_0)=0$ since $\mathbf{f}(\mathbf{x}_0,\mathbf{y}_0)=\mathbf{0}$. Therefore, from the definition of $\mathbf{x}_k$, it is also the case that $h_{\mathbf{y}_k}(\mathbf{x}_k)<\varepsilon$. Passing to the limit yields $h_{\mathbf{y}_0}(\mathbf{x})\le\varepsilon$. Since $\varepsilon>0$ is arbitrary, it follows that $h_{\mathbf{y}_0}(\mathbf{x})=0$ which contradicts the first part of the argument in which it was shown that for $\mathbf{y}\in B(\mathbf{y}_0,\eta)$ there is at most one point, $\mathbf{x}$ of $\overline{B(\mathbf{x}_0,\delta)}$ where $\mathbf{f}(\mathbf{x},\mathbf{y})=\mathbf{0}$. Here two have been obtained, $\mathbf{x}_0$ and $\mathbf{x}$. This proves the claim.

Choose $\eta<\eta_0$ and also small enough that the above claim holds and let $\mathbf{x}(\mathbf{y})$ denote a point of $B(\mathbf{x}_0,\delta)$ at which the minimum of $h_{\mathbf{y}}$ on $\overline{B(\mathbf{x}_0,\delta)}$ is achieved. Since $\mathbf{x}(\mathbf{y})$ is an interior point, you can consider $h_{\mathbf{y}}(\mathbf{x}(\mathbf{y})+t\mathbf{v})$ for $|t|$ small and conclude this function of t has a zero derivative at $t=0$. Now

$$h_{\mathbf{y}}(\mathbf{x}(\mathbf{y})+t\mathbf{v})=\sum_{i=1}^{n}f_i^2(\mathbf{x}(\mathbf{y})+t\mathbf{v},\mathbf{y})$$

and so from the chain rule,

$$\frac{d}{dt}h_{\mathbf{y}}(\mathbf{x}(\mathbf{y})+t\mathbf{v})=\sum_{i=1}^{n}2f_i(\mathbf{x}(\mathbf{y})+t\mathbf{v},\mathbf{y})\frac{\partial f_i(\mathbf{x}(\mathbf{y})+t\mathbf{v},\mathbf{y})}{\partial x_j}v_j.$$

Therefore, letting $t=0$, it is required that for every $\mathbf{v}$,

$$\sum_{i=1}^{n}2f_i(\mathbf{x}(\mathbf{y}),\mathbf{y})\frac{\partial f_i(\mathbf{x}(\mathbf{y}),\mathbf{y})}{\partial x_j}v_j=0.$$

In terms of matrices this reduces to

$$0=2\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})^T D_1\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})\mathbf{v}$$

for every vector $\mathbf{v}$. Therefore,

$$\mathbf{0}=\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})^T D_1\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})$$

From 22.3, it follows $\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})=\mathbf{0}$. This proves the existence of the function $\mathbf{y}\to\mathbf{x}(\mathbf{y})$ such that $\mathbf{f}(\mathbf{x}(\mathbf{y}),\mathbf{y})=\mathbf{0}$ for all $\mathbf{y}\in B(\mathbf{y}_0,\eta)$.

It remains to verify this function is a C^1 function. To do this, let $\mathbf{y}_1$ and $\mathbf{y}_2$ be points of $B(\mathbf{y}_0,\eta)$. Then as before, consider the i^{th} component of $\mathbf{f}$ and consider the same argument using the mean value theorem to write

$$\begin{aligned}0&=f_i(\mathbf{x}(\mathbf{y}_1),\mathbf{y}_1)-f_i(\mathbf{x}(\mathbf{y}_2),\mathbf{y}_2)\\&=f_i(\mathbf{x}(\mathbf{y}_1),\mathbf{y}_1)-f_i(\mathbf{x}(\mathbf{y}_2),\mathbf{y}_1)+f_i(\mathbf{x}(\mathbf{y}_2),\mathbf{y}_1)-f_i(\mathbf{x}(\mathbf{y}_2),\mathbf{y}_2)\\&=D_1f_i\left(\mathbf{x}^i,\mathbf{y}_1\right)(\mathbf{x}(\mathbf{y}_1)-\mathbf{x}(\mathbf{y}_2))+D_2f_i\left(\mathbf{x}(\mathbf{y}_2),\mathbf{y}^i\right)(\mathbf{y}_1-\mathbf{y}_2).\end{aligned}\tag{22.7}$$

where $\mathbf{y}^i$ is a point on the line segment joining $\mathbf{y}_1$ and $\mathbf{y}_2$. Thus from 22.4 and the Cauchy Schwarz inequality,

$$\left|D_2 f_i\left(\mathbf{x}\left(\mathbf{y}_2\right), \mathbf{y}^i\right)\left(\mathbf{y}_1-\mathbf{y}_2\right)\right| \leq K\left|\mathbf{y}_1-\mathbf{y}_2\right|.$$

Therefore, letting $M\left(\mathbf{y}^1, \cdots, \mathbf{y}^n\right) \equiv M$ denote the matrix having the i^{th} row equal to $D_2 f_i\left(\mathbf{x}\left(\mathbf{y}_2\right), \mathbf{y}^i\right)$, it follows

$$\left|M\left(\mathbf{y}_1-\mathbf{y}_2\right)\right| \leq \left(\sum_i K^2\left|\mathbf{y}_1-\mathbf{y}_2\right|^2\right)^{1/2} = \sqrt{m} K\left|\mathbf{y}_1-\mathbf{y}_2\right|. \tag{22.8}$$

Also, from 22.7,

$$J\left(\mathbf{x}^1, \cdots, \mathbf{x}^n, \mathbf{y}_1\right)\left(\mathbf{x}\left(\mathbf{y}_1\right)-\mathbf{x}\left(\mathbf{y}_2\right)\right) = -M\left(\mathbf{y}_1-\mathbf{y}_2\right) \tag{22.9}$$

and so from 22.8 and 22.10,

$$\begin{aligned}
\left|\mathbf{x}\left(\mathbf{y}_1\right)-\mathbf{x}\left(\mathbf{y}_2\right)\right| &= \left|J\left(\mathbf{x}^1, \cdots, \mathbf{x}^n, \mathbf{y}_1\right)^{-1} M\left(\mathbf{y}_1-\mathbf{y}_2\right)\right| && (22.10)\\
&= \left(\sum_{i=1}^n\left|\mathbf{e}_i^T J\left(\mathbf{x}^1, \cdots, \mathbf{x}^n, \mathbf{y}_1\right)^{-1} M\left(\mathbf{y}_1-\mathbf{y}_2\right)\right|^2\right)^{1/2} \\
&\leq \left(\sum_{i=1}^n K^2\left|M\left(\mathbf{y}_1-\mathbf{y}_2\right)\right|^2\right)^{1/2} \leq \left(\sum_{i=1}^n K^2\left(\sqrt{m} K\left|\mathbf{y}_1-\mathbf{y}_2\right|\right)^2\right)^{1/2} \\
&= K^2 \sqrt{m n}\left|\mathbf{y}_1-\mathbf{y}_2\right| && (22.11)
\end{aligned}$$

It follows as in the proof of the chain rule that

$$\mathbf{o}\left(\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right)-\mathbf{x}\left(\mathbf{y}\right)\right) = \mathbf{o}\left(\mathbf{v}\right). \tag{22.12}$$

Now let $\mathbf{y} \in B\left(\mathbf{y}_0, \eta\right)$ and let $|\mathbf{v}|$ be sufficiently small that $\mathbf{y}+\mathbf{v} \in B\left(\mathbf{y}_0, \eta\right)$. Then

$$\begin{aligned}
\mathbf{0} &= \mathbf{f}\left(\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right), \mathbf{y}+\mathbf{v}\right)-\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right) \\
&= \mathbf{f}\left(\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right), \mathbf{y}+\mathbf{v}\right)-\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right), \mathbf{y}\right)+\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right), \mathbf{y}\right)-\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right)
\end{aligned}$$

$$= D_2\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right), \mathbf{y}\right)\mathbf{v}+D_1\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right)\left(\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right)-\mathbf{x}\left(\mathbf{y}\right)\right)+\mathbf{o}\left(\left|\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right)-\mathbf{x}\left(\mathbf{y}\right)\right|\right)$$

$$\begin{aligned}
&= D_2\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right)\mathbf{v}+D_1\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right)\left(\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right)-\mathbf{x}\left(\mathbf{y}\right)\right)+ \\
&\quad \mathbf{o}\left(\left|\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right)-\mathbf{x}\left(\mathbf{y}\right)\right|\right)+\left(D_2\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right), \mathbf{y}\right)\mathbf{v}-D_2\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right)\mathbf{v}\right) \\
&= D_2\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right)\mathbf{v}+D_1\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right)\left(\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right)-\mathbf{x}\left(\mathbf{y}\right)\right)+\mathbf{o}\left(\mathbf{v}\right).
\end{aligned}$$

Therefore,

$$\mathbf{x}\left(\mathbf{y}+\mathbf{v}\right)-\mathbf{x}\left(\mathbf{y}\right) = -D_1\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right)^{-1} D_2\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right)\mathbf{v}+\mathbf{o}\left(\mathbf{v}\right)$$

which shows that $D\mathbf{x}\left(\mathbf{y}\right) = -D_1\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right)^{-1} D_2\mathbf{f}\left(\mathbf{x}\left(\mathbf{y}\right), \mathbf{y}\right)$ and $\mathbf{y} \rightarrow D\mathbf{x}\left(\mathbf{y}\right)$ is continuous. This proves the theorem.

22.1 The Method Of Lagrange Multipliers

As an application of the implicit function theorem, consider the method of Lagrange multipliers. Recall the problem is to maximize or minimize a function subject to equality constraints. Let $f : U \to \mathbb{R}$ be a C^1 function where $U \subseteq \mathbb{R}^n$ and let

$$g_i(\mathbf{x}) = 0,\ i = 1, \cdots, m \tag{22.13}$$

be a collection of equality constraints with $m < n$. Now consider the system of nonlinear equations

$$\begin{aligned} f(\mathbf{x}) &= a \\ g_i(\mathbf{x}) &= 0,\ i = 1, \cdots, m. \end{aligned}$$

Recall $\mathbf{x}_0$ is a local maximum if $f(\mathbf{x}_0) \geq f(\mathbf{x})$ for all $\mathbf{x}$ near $\mathbf{x}_0$ which also satisfies the constraints 22.13. A local minimum is defined similarly. Let $\mathbf{F} : U \times \mathbb{R} \to \mathbb{R}^{m+1}$ be defined by

$$\mathbf{F}(\mathbf{x},a) \equiv \begin{pmatrix} f(\mathbf{x}) - a \\ g_1(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{pmatrix}. \tag{22.14}$$

Now consider the $m+1 \times n$ matrix,

$$\begin{pmatrix} f_{x_1}(\mathbf{x}_0) & \cdots & f_{x_n}(\mathbf{x}_0) \\ g_{1x_1}(\mathbf{x}_0) & \cdots & g_{1x_n}(\mathbf{x}_0) \\ \vdots & & \vdots \\ g_{mx_1}(\mathbf{x}_0) & \cdots & g_{mx_n}(\mathbf{x}_0) \end{pmatrix}.$$

If this matrix has rank $m+1$ then some $m+1 \times m+1$ submatrix has nonzero determinant. It follows from the implicit function theorem there exists $m+1$ variables, $x_{i_1}, \cdots, x_{i_{m+1}}$ such that the system

$$\mathbf{F}(\mathbf{x},a) = \mathbf{0} \tag{22.15}$$

specifies these $m+1$ variables as a function of the remaining $n-(m+1)$ variables and a in an open set of $\mathbb{R}^{n-m}$. Thus there is a solution $(\mathbf{x},a)$ to 22.15 for some $\mathbf{x}$ close to $\mathbf{x}_0$ whenever a is in some open interval. Therefore, $\mathbf{x}_0$ cannot be either a local minimum or a local maximum. It follows that if $\mathbf{x}_0$ is either a local maximum or a local minimum, then the above matrix must have rank less than $m+1$ which requires the rows to be linearly dependent. Thus, there exist m scalars,

$$\lambda_1, \cdots, \lambda_m,$$

and a scalar μ, not all zero such that

$$\mu \begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix}. \tag{22.16}$$

If the column vectors

$$\begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix}, \cdots \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \tag{22.17}$$

are linearly independent, then, $\mu \neq 0$ and dividing by μ yields an expression of the form

$$\begin{pmatrix} f_{x_1}(\mathbf{x}_0) \\ \vdots \\ f_{x_n}(\mathbf{x}_0) \end{pmatrix} = \lambda_1 \begin{pmatrix} g_{1x_1}(\mathbf{x}_0) \\ \vdots \\ g_{1x_n}(\mathbf{x}_0) \end{pmatrix} + \cdots + \lambda_m \begin{pmatrix} g_{mx_1}(\mathbf{x}_0) \\ \vdots \\ g_{mx_n}(\mathbf{x}_0) \end{pmatrix} \tag{22.18}$$

at every point $\mathbf{x}_0$ which is either a local maximum or a local minimum. This proves the following theorem.

Theorem 22.1.1 *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 22.13, then 22.16 must hold for some scalars $\mu, \lambda_1, \cdots, \lambda_m$ not all equal to zero. If the vectors in 22.17 are linearly independent, it follows that an equation of the form 22.18 holds.*

22.2 The Local Structure Of C^1 Mappings

Definition 22.2.1 *Let U be an open set in $\mathbb{R}^n$ and let $\mathbf{h} : U \to \mathbb{R}^n$. Then $\mathbf{h}$ is called primitive if it is of the form*

$$\mathbf{h}(\mathbf{x}) = \begin{pmatrix} x_1 & \cdots & \alpha(\mathbf{x}) & \cdots & x_n \end{pmatrix}^T.$$

Thus, $\mathbf{h}$ is primitive if it only changes one of the variables. A function, $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ is called a flip if

$$\mathbf{F}(x_1, \cdots, x_k, \cdots, x_l, \cdots, x_n) = (x_1, \cdots, x_l, \cdots, x_k, \cdots, x_n)^T.$$

Thus a function is a flip if it interchanges two coordinates. Also, for $m = 1, 2, \cdots, n$,

$$P_m(\mathbf{x}) \equiv \begin{pmatrix} x_1 & x_2 & \cdots & x_m & 0 & \cdots & 0 \end{pmatrix}^T$$

It turns out that if $\mathbf{h}(\mathbf{0}) = \mathbf{0}$,$D\mathbf{h}(0)^{-1}$ exists, and $\mathbf{h}$ is C^1 on U, then $\mathbf{h}$ can be written as a composition of primitive functions and flips. This is a very interesting application of the inverse function theorem.

Theorem 22.2.2 *Let $\mathbf{h} : U \to \mathbb{R}^n$ be a C^1 function with $\mathbf{h}(\mathbf{0}) = \mathbf{0}$,$D\mathbf{h}(\mathbf{0})^{-1}$ exists. Then there an open set, $V \subseteq U$ containing $\mathbf{0}$, flips, $\mathbf{F}_1, \cdots, \mathbf{F}_{n-1}$, and primitive functions, $\mathbf{G}_n, \mathbf{G}_{n-1}, \cdots, \mathbf{G}_1$ such that for $\mathbf{x} \in V$,*

$$\mathbf{h}(\mathbf{x}) = \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{G}_n \circ \mathbf{G}_{n-1} \circ \cdots \circ \mathbf{G}_1(\mathbf{x}).$$

Proof: Let

$$\mathbf{h}_1(\mathbf{x}) \equiv \mathbf{h}(\mathbf{x}) = \begin{pmatrix} \alpha_1(\mathbf{x}) & \cdots & \alpha_n(\mathbf{x}) \end{pmatrix}^T$$

$$D\mathbf{h}(\mathbf{0})\mathbf{e}_1 = \begin{pmatrix} \alpha_{1,1}(\mathbf{0}) & \cdots & \alpha_{n,1}(\mathbf{0}) \end{pmatrix}^T$$

where $\alpha_{k,1}$ denotes $\frac{\partial \alpha_k}{\partial x_1}$. Since $D\mathbf{h}(\mathbf{0})$ is one to one, the right side of this expression cannot be zero. Hence there exists some k such that $\alpha_{k,1}(\mathbf{0}) \neq 0$. Now define

$$\mathbf{G}_1(\mathbf{x}) \equiv \begin{pmatrix} \alpha_k(\mathbf{x}) & x_2 & \cdots & x_n \end{pmatrix}^T$$

Then the matrix of $D\mathbf{G}(\mathbf{0})$ is of the form

$$\begin{pmatrix} \alpha_{k,1}(\mathbf{0}) & \cdots & \cdots & \alpha_{k,n}(0) \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

and its determinant equals $\alpha_{k,1}(\mathbf{0}) \neq 0$. Therefore, by the inverse function theorem, there exists an open set, U_1 containing $\mathbf{0}$ and an open set, V_2 containing $\mathbf{0}$ such that $\mathbf{G}_1(U_1) = V_2$ and $\mathbf{G}_1$ is one to one and onto such that it and its inverse are both C^1. Let $\mathbf{F}_1$ denote the flip which interchanges x_k with x_1. Now define

$$\mathbf{h}_2(\mathbf{y}) \equiv \mathbf{F}_1 \circ \mathbf{h}_1 \circ \mathbf{G}_1^{-1}(\mathbf{y})$$

Thus

$$\begin{aligned} \mathbf{h}_2(\mathbf{G}_1(\mathbf{x})) &\equiv \mathbf{F}_1 \circ \mathbf{h}_1(\mathbf{x}) \\ &= \begin{pmatrix} \alpha_k(\mathbf{x}) & \cdots & \alpha_1(\mathbf{x}) & \cdots & \alpha_n(\mathbf{x}) \end{pmatrix}^T \end{aligned} \tag{22.19}$$

Therefore,

$$P_1\mathbf{h}_2(\mathbf{G}_1(\mathbf{x})) = \begin{pmatrix} \alpha_k(\mathbf{x}) & 0 & \cdots & 0 \end{pmatrix}^T.$$

Also

$$P_1(\mathbf{G}_1(\mathbf{x})) = \begin{pmatrix} \alpha_k(\mathbf{x}) & x_2 & \cdots & x_n \end{pmatrix}^T$$

so $P_1\mathbf{h}_2(\mathbf{y}) = P_1(\mathbf{y})$ for all $\mathbf{y} \in V_2$. Also, $\mathbf{h}_2(\mathbf{0}) = \mathbf{0}$ and $D\mathbf{h}_2(0)^{-1}$ exists because of the definition of $\mathbf{h}_2$ above and the chain rule. Also, since $\mathbf{F}_1^2 = \mathit{identity}$, it follows from 22.19 that

$$\mathbf{h}(\mathbf{x}) = \mathbf{h}_1(\mathbf{x}) = \mathbf{F}_1 \circ \mathbf{h}_2 \circ \mathbf{G}_1(\mathbf{x}). \tag{22.20}$$

Suppose then that for $m \geq 2$,

$$P_{m-1}\mathbf{h}_m(\mathbf{x}) = P_{m-1}(\mathbf{x}) \tag{22.21}$$

for all $\mathbf{x} \in U_m$, an open subset of U containing $\mathbf{0}$ and $\mathbf{h}_m(\mathbf{0}) = \mathbf{0}$,$D\mathbf{h}_m(\mathbf{0})^{-1}$ exists. From 22.21, $\mathbf{h}_m(\mathbf{x})$ must be of the form

$$\mathbf{h}_m(\mathbf{x}) = \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_1(\mathbf{x}) & \cdots & \alpha_n(\mathbf{x}) \end{pmatrix}^T$$

where these α_k are different than the ones used earlier. Then

$$D\mathbf{h}_m(\mathbf{0})\mathbf{e}_m = \begin{pmatrix} 0 & \cdots & 0 & \alpha_{1,m}(\mathbf{0}) & \cdots & \alpha_{n,m}(\mathbf{0}) \end{pmatrix}^T \neq \mathbf{0}$$

because $D\mathbf{h}_m(\mathbf{0})^{-1}$ exists. Therefore, there exists a k such that $\alpha_{k,m}(\mathbf{0}) \neq 0$, not the same k as before. Define

$$\mathbf{G}_{m+1}(\mathbf{x}) \equiv \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_k(\mathbf{x}) & x_{m+1} & \cdots & x_n \end{pmatrix}^T \tag{22.22}$$

Then $G_{m+1}(\mathbf{0}) = \mathbf{0}$ and $D\mathbf{G}_{m+1}(\mathbf{0})^{-1}$ exists similar to the above. In fact $\det(D\mathbf{G}_{m+1}(\mathbf{0})) = \alpha_{k,m}(\mathbf{0})$. Therefore, by the inverse function theorem, there exists an open set, V_{m+1} containing $\mathbf{0}$ such that $V_{m+1} = \mathbf{G}_{m+1}(U_m)$ with $\mathbf{G}_{m+1}$ and its inverse being one to one continuous and onto. Let $\mathbf{F}_m$ be the flip which flips x_m and x_k. Then define $\mathbf{h}_{m+1}$ on V_{m+1} by

$$\mathbf{h}_{m+1}(\mathbf{y}) = \mathbf{F}_m \circ \mathbf{h}_m \circ \mathbf{G}_{m+1}^{-1}(\mathbf{y}).$$

Thus for $\mathbf{x} \in U_m$,

$$\mathbf{h}_{m+1}(\mathbf{G}_{m+1}(\mathbf{x})) = (\mathbf{F}_m \circ \mathbf{h}_m)(\mathbf{x}). \tag{22.23}$$

and consequently,

$$\mathbf{F}_m \circ \mathbf{h}_{m+1} \circ \mathbf{G}_{m+1}(\mathbf{x}) = \mathbf{h}_m(\mathbf{x}) \tag{22.24}$$

It follows

$$\begin{aligned} P_m \mathbf{h}_{m+1}(\mathbf{G}_{m+1}(\mathbf{x})) &= P_m(\mathbf{F}_m \circ \mathbf{h}_m)(\mathbf{x}) \\ &= \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_k(\mathbf{x}) & 0 & \cdots & 0 \end{pmatrix}^T \end{aligned}$$

and

$$P_m(\mathbf{G}_{m+1}(\mathbf{x})) = \begin{pmatrix} x_1 & \cdots & x_{m-1} & \alpha_k(\mathbf{x}) & 0 & \cdots & 0 \end{pmatrix}^T.$$

Therefore, for $\mathbf{y} \in V_{m+1}$,

$$P_m \mathbf{h}_{m+1}(\mathbf{y}) = P_m(\mathbf{y}).$$

As before, $\mathbf{h}_{m+1}(\mathbf{0}) = \mathbf{0}$ and $D\mathbf{h}_{m+1}(\mathbf{0})^{-1}$ exists. Therefore, we can apply 22.24 repeatedly, obtaining the following:

$$\begin{aligned} \mathbf{h}(\mathbf{x}) &= \mathbf{F}_1 \circ \mathbf{h}_2 \circ \mathbf{G}_1(\mathbf{x}) \\ &= \mathbf{F}_1 \circ \mathbf{F}_2 \circ \mathbf{h}_3 \circ \mathbf{G}_2 \circ \mathbf{G}_1(\mathbf{x}) \\ &\vdots \\ &= \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{h}_n \circ \mathbf{G}_{n-1} \circ \cdots \circ \mathbf{G}_1(\mathbf{x}) \end{aligned}$$

where

$$P_{n-1}\mathbf{h}_n(\mathbf{x}) = P_{n-1}(\mathbf{x}) = \begin{pmatrix} x_1 & \cdots & x_{n-1} & 0 \end{pmatrix}^T$$

and so $\mathbf{h}_n(\mathbf{x})$ is of the form

$$\mathbf{h}_n(\mathbf{x}) = \begin{pmatrix} x_1 & \cdots & x_{n-1} & \alpha(\mathbf{x}) \end{pmatrix}^T.$$

Therefore, define the primitive function, $\mathbf{G}_n(\mathbf{x})$ to equal $\mathbf{h}_n(\mathbf{x})$. This proves the theorem.

Part IX

Multiple Integrals

Outcomes

Double Integrals

A. Compare the definition of the double integral to the method of repeated integration geometrically.

B. Evaluate double integrals over a rectangle by repeated integration.

C. Apply a double integral to calculate the volume or mass of a solid.

Reading: Multivariable Calculus 3.1
Outcome Mapping:

A. J1

B. 1,2

C. 3,4

Double Integrals Over General Regions

A. Evaluate double integrals over general regions.

B. Evaluate double integrals by interpreting them as known volumes.

C. Rewrite a double integral changing the order of integration.

D. Apply double integrals to calculate volumes of solids.

E. Evaluate the physical characteristics of a plate such as mass, centroid, center of mass and moment of inertia.

Reading: Multivariable Calculus 3.2
Outcome Mapping:

A. 1

B. 2

C. 3

D. 4

E. 5,6

Double Integrals in Polar Coordinates

A. Represent a region in both Cartesian and polar coordinates.

B. Evaluate double integrals in polar coordinates.

C. Convert a double integral in Cartesian coordinates to a double integral in polar coordinates and then evaluate.

D. Evaluate areas and volumes using polar coordinates

E. Evaluate the physical characteristics of a plate such as centroid, mass, and center of mass using polar coordinates.

F. Make conversions of algebraic expressions between Cartesian coordinates and cylindrical coordinates.

Reading: Multivariable Calculus 3.3
Outcome Mapping:

A. 1,4

B. 5a

C. 8

D. 5b,6,7

E. 5c

F. 9,10,11,12

Triple Integrals

A. Find the volume of a solid using triple integration in Cartesian coordinates.

B. Evaluate the physical characteristics of a solid such as mass, centroid and center of mass using Cartesian coordinates.

Reading: Multivariable Calculus 3.4
Outcome Mapping:

A. 1,4,7

B. 2,4,7

Triple Integrals in Cylindrical Coordinats

A. Describe regions in both Cartesian coordinates and cylindrical coordinates.

B. Evaluate triple integrals using cylindrical coordinates.

C. Find volumes by applying triple integration in cylindrical coordinates.

D. Evaluate the physical characteristics of a solid such as mass, centroid and center of mass using cylindrical coordinates.

Reading: Multivariable Calculus 3.5
Outcome Mapping:

A. 2,3

B. 5

C. 6

D. 7,8,11

Triple Integrals in Spherical Coordinats

A. Describe regions in both Cartesian coordinates and spherical coordinates.

B. Evaluate triple integrals using spherical coordinates.

C. Find volumes by applying triple integration in spherical coordinates.

D. Evaluate the physical characteristics of a solid such as mass, centroid and center of mass using spherical coordinates.

E. Convert a triple integral in Cartesian coordinates to cylindrical or spherical coordinates and then evaluate.

F. Make conversions of algebraic expressions between Cartesian coordinates and spherical coordinates.

Reading: Multivariable Calculus 3.6
Outcome Mapping:

A. 1,2,3

B. 5

C. 6

D. 7

E. 10

F. 11,12,13,14,15

The Jacobian

A. Find the Jacobian of a transformation.

B. Change variables in a multiple integration to obtain a more simple integral and then evaluate.

Reading: Multivariable Calculus 3.7
Outcome Mapping:

A. 1

B. 3,5,6

The Riemann Integral On $\mathbb{R}^n$

23.1 Methods For Double Integrals 1 Nov.

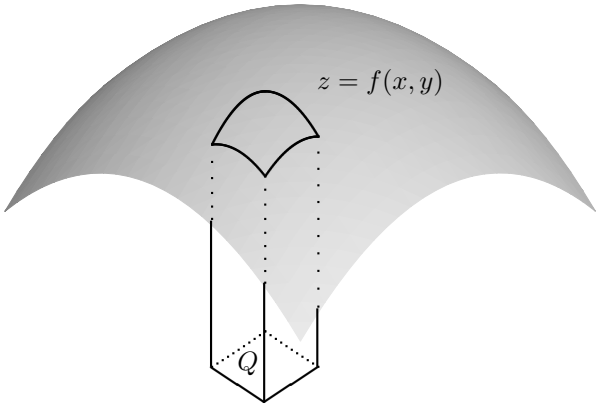
Quiz

1. Maximize $2x + y$ subject to the condition that $\frac{x^2}{4} + \frac{y^2}{9} \leq 1$.

2. 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 $(0, 0, 0)$.

3. Find the points on $y^2 x = 9$ which are closest to $(0, 0)$.

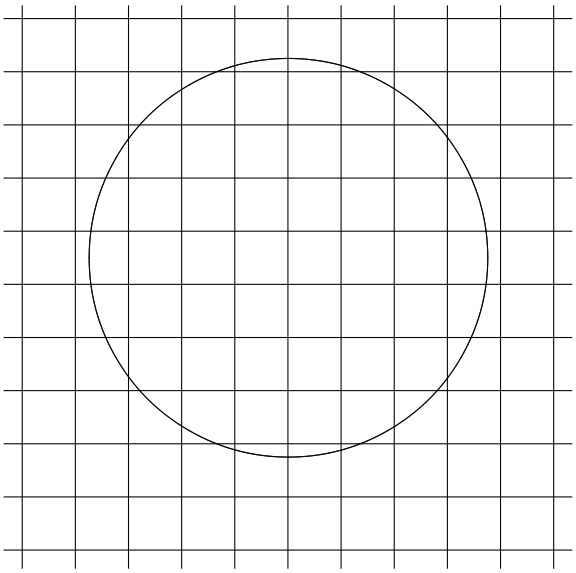
This chapter is on the Riemann integral for a function of n variables. It begins by introducing the basic concepts and applications of the integral. The proofs of the theorems involved are difficult and are left till the end. To begin with consider the problem of finding the volume under a surface of 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

$$M_Q(f) \equiv \sup\{f(\mathbf{x}) : \mathbf{x} \in Q\},\ m_Q(f) \equiv \inf\{f(\mathbf{x}) : \mathbf{x} \in Q\}, \tag{23.1}$$

and $v(Q)$ is defined as the area of Q. Now consider the following picture.

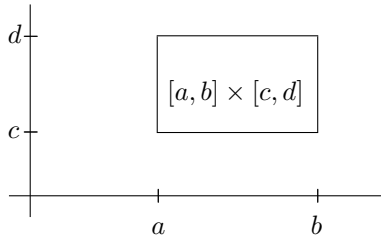


In this picture, it is assumed f equals zero outside the circle and f is a bounded nonnegative function. Then each of those little squares are the base of a prism of the sort in the previous picture and the sum of the volumes of those prisms should be the volume under the surface, $z = f(x,y)$. Therefore, 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 23.1.1 *For $i = 1,2$, let $\left\{\alpha_k^i\right\}_{k=-\infty}^{\infty}$ be points on $\mathbb{R}$ which satisfy*

$$\lim_{k\to\infty} \alpha_k^i = \infty,\ \lim_{k\to-\infty} \alpha_k^i = -\infty,\ \alpha_k^i < \alpha_{k+1}^i. \tag{23.2}$$

*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*

$$Q = \left[\alpha_k^1, \alpha_{k+1}^1\right] \times \left[\alpha_l^2, \alpha_{l+1}^2\right]. \tag{23.3}$$

*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,

$$\sum_{Q\in\mathcal{G}} M_Q(f)\, v(Q)$$

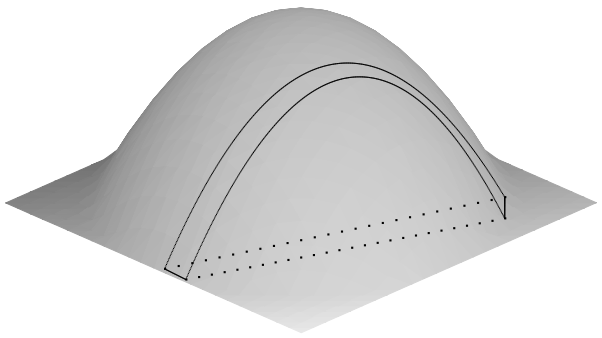
is called the upper sum associated with the grid, $\mathcal{G}$ as described above in the discussion of the volume under a surface. Again, this means to take a rectangle from $\mathcal{G}$ multiply $M_Q(f)$ defined in 23.1 by its area, $v(Q)$ and sum all these products for every $Q \in \mathcal{G}$. 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 23.1.2 *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

$$\int_c^d \left(\int_a^b f(x,y)\, dx \right) dy \tag{23.4}$$

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,

$$\int_a^b \left(\int_c^d f(x,y)\, dy \right) dx \tag{23.5}$$

These expressions in 23.4 and 23.5 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 23.1.3 *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 23.1.4 *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.

$$\int_0^2 \int_0^1 \left(x^2y + yx\right)\,dx\,dy$$

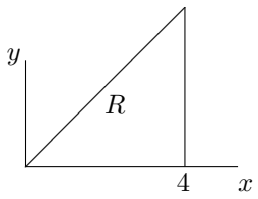
$$\int_0^1 \left(x^2y + yx\right)\,dx = \frac{5}{6}y$$

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 23.1.5 *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.

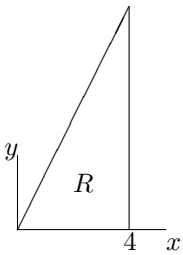
$$\int_y^4 \left(x^2y + yx\right)\,dx = \frac{88}{3}y - \frac{1}{3}y^4 - \frac{1}{2}y^3$$

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 23.1.6 *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

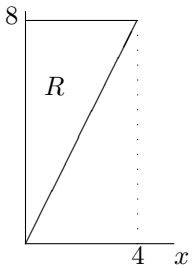
$$\int_R f\, dV = \int_0^8 \left(\frac{64}{3} y - \frac{1}{24} y^4\right)\, dy = \frac{2048}{5}$$

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 23.1.7 *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.

23.1.1 Density Mass And Center Of Mass

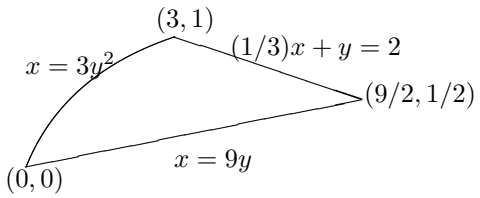
Consider a two dimensional material. Of course there is no such thing but a flat plate might be modeled as one. The density ρ is a function of position and is defined as follows. Consider a small chunk of area, dV located at the point whose Cartesian coordinates are (x, y). Then the mass of this small 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 is

$$\int_U \rho\, dV$$

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 23.1.8 *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.

The concept of center of mass of a plate occupying the bounded open set, U is also easy to express in terms of double integrals. Letting ρ denote the density of the plate, the moment of a small chunk of mass having coordinates (x, y) about the y axis is $x\rho(x, y)\, dV$ and the moment of the same small chunk of mass about the x axis is $y\rho(x, y)\, dV$. Therefore the center of mass, (x_c, y_c) is defined in the usual way.

Definition 23.1.9 *The center of mass of a plate occuying the bounded open set, U is defined as (x_c, y_c) where*

$$x_c \equiv \frac{\int_U x\rho(x, y)\, dV}{\int_U \rho(x, y)\, dV},\ y_c \equiv \frac{\int_U y\rho(x, y)\, dV}{\int_U \rho(x, y)\, dV}.$$

In other words, the total moment about the y axis equals x_c times the total mass. That is, if you placed the total mass at the single point, (x_c, y_c) this point mass would produce the same moments about the x and y axes as the original plate.

Example 23.1.10 *In Example 23.1.8, suppose the density is $\rho(x, y) = y$ as it is in that example. Find the total mass and the center of mass.*

First, the total mass was found above. Then the center of mass is

$$x_c = \frac{\int_0^3 \int_{x/9}^{\sqrt{x/3}} xy\, dy\, dx + \int_3^{\frac{9}{2}} \int_{x/9}^{2-\frac{1}{3}x} xy\, dy\, dx}{\int_3^{\frac{9}{2}} \int_{x/9}^{2-\frac{1}{3}x} y\, dy\, dx + \int_0^3 \int_{x/9}^{\sqrt{x/3}} y\, dy\, dx} = \frac{\frac{39}{16}}{1} = \frac{39}{16}$$

$$y_c = \frac{\int_0^3 \int_{x/9}^{\sqrt{x/3}} y^2\, dy\, dx + \int_3^{\frac{9}{2}} \int_{x/9}^{2-\frac{1}{3}x} y^2\, dy\, dx}{\int_3^{\frac{9}{2}} \int_{x/9}^{2-\frac{1}{3}x} y\, dy\, dx + \int_0^3 \int_{x/9}^{\sqrt{x/3}} y\, dy\, dx} = \frac{47}{80}.$$

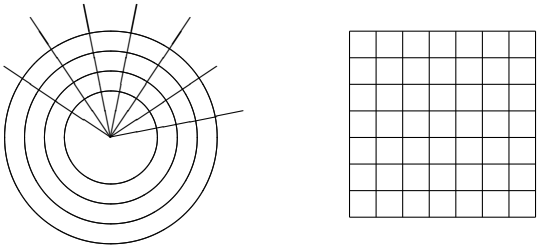
Thus the center of mass is $\left(\frac{39}{16}, \frac{47}{80}\right)$.

23.2 Double Integrals In Polar Coordinates

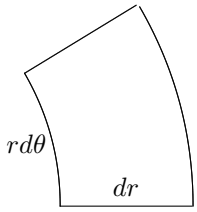
Remember polar coordinates,

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned}$$

where $\theta \in [0, 2\pi]$ and $r > 0$. If you assign a given value to r, the points obtained yield a circle and if you give a value to θ the points yield a ray from the origin. Thus assigning many different values for r and many different values for θ yields a grid of the sort illustrated in the following picture.



By contrast, the grid on the right is obtained by assigning different values for x and y. For the grid on the right, if the vertical lines are dx apart and the horizontal lines are dy apart, the area of one of those little boxes would be $dxdy$. This is the **increment of area** in rectangular coordinates. Now consider the grid on the left which is obtained by setting each of the two polar variables equal to various constants. What is the area of one of those little curvy rectangles if the values for r and θ are very small? Zoom in on one of them as illustrated in the following picture.



The angle between the two straight lines is $d\theta$ and so the length of one of the curved sides is approximately $rd\theta$ while the length of the straight sides is dr. Therefore, the area of the little curvy rectangle is approximately equal to $rdrd\theta$. This is the increment of area in polar coordinates.

Later, I will present a unified way to change variables. For now, consider the following problems which illustrate the use of polar coordinates to compute integrals over areas.

Example 23.2.1 *Find the area of a circle of radius R.*

Denote by D this circle. Then the area of the circle is $\int dA$ and you need to write dA in terms of polar coordinates. As described above, $dA = rdrd\theta$. To compute the integral, note that in terms of the variables, θ and r, this region is actually the rectangle, $[0,R] \times [0,2\pi]$. Therefore, the integral equals

$$\int_0^{2\pi} \int_0^R rdrd\theta = \pi R^2$$

which you have already heard about.

Example 23.2.2 *Find the volume of the ball of radius R.*

It is enough to find the volume of the top half of this ball and then multiply it by 2. Corresponding to the small curvy rectangle as described above having polar coordinates (r,θ) the height of the ball over this point is $\sqrt{R^2 - r^2}$. Therefore, the volume of a small prism having as a base the small curvy rectangle described above is $\sqrt{R^2 - r^2} rdrd\theta$. Summing these using the integral, the desired volume is

$$\int_D \sqrt{R^2 - r^2} dA = \int_0^{2\pi} \int_0^R \sqrt{R^2 - r^2} rdrd\theta = \frac{2}{3}\pi R^3$$

and so the volume of the whole ball is

$$\frac{4}{3}\pi R^3$$

which is another formula you might have seen.

Example 23.2.3 *Find the area inside $r = 1 + \cos\theta$ for $\theta \in [0, 2\pi]$.*

This is the graph of a cardioid. You saw this in beginning calculus. Let its inside be denoted by C for cardioid. Then the desired area is

$$\int_C dA$$

and you need to set up an iterated integral and put in the correct form for dA. For each θ, you have that r goes from 0 to $1+\cos\theta$. Therefore, the desired area is given by the iterated integral,

$$\int_0^{2\pi}\int_0^{1+\cos\theta} rdrd\theta = \frac{3}{2}\pi$$

This was really easy because of polar coordinates. If you try to do this in rectangular coordinates it will not work very well.

Example 23.2.4 *A plate occupies the region inside the curve, $r = 2\cos\theta$ for $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. The density in terms of polar coordinates is $\delta = r$. Find the mass and center of mass of this plate.*

First of all the mass is given by

$$\int_{-\pi/2}^{\pi/2}\int_0^{2\cos(\theta)} r^2 drd\theta = \frac{32}{9}$$

The volume element is $rdrd\theta$ and I summed these up multiplied, by the density which was r and that is the above integral.

Now to compute the center of mass, recall

$$x_c \equiv \frac{\int_U x\rho(x,y)\,dV}{\int_U \rho(x,y)\,dV},\ y_c \equiv \frac{\int_U y\rho(x,y)\,dV}{\int_U \rho(x,y)\,dV}$$

I need to place x and y in terms of the polar coordinates. Thus $x = r\cos\theta, y = r\sin\theta$. A small contribution to the moment about the y axis is $r\cos(\theta)\times r\times rdrd\theta$. Thus

$$x_c = \frac{\int_{-\pi/2}^{\pi/2}\int_0^{2\cos(\theta)} r^2\cos(\theta)\overbrace{rdrd\theta}^{dA}}{\int_{-\pi/2}^{\pi/2}\int_0^{2\cos(\theta)} r\times\overbrace{rdrd\theta}^{dA}} = \frac{6}{5}.$$

Similarly,

$$y_c = \frac{\int_{-\pi/2}^{\pi/2}\int_0^{2\cos(\theta)} r^2\sin(\theta)\,rdrd\theta}{\int_{-\pi/2}^{\pi/2}\int_0^{2\cos(\theta)} r^2 drd\theta} = 0$$

Example 23.2.5 *Let a plate occupy the region, C which is inside the polar graph, $r = 2+\cos\theta$ for $\theta \in [0, 2\pi]$. Suppose the density of this plate is given by $\delta(r,\theta) = r$. Find the mass and center of mass of the plate.*

Here you need to evaluate the following to get the total mass.

$$\int_0^{2\pi}\int_0^{2+\cos\theta} r\times r dr d\theta = \int_0^{2\pi}\int_0^{2+\cos\theta} r^2 dr d\theta = \frac{22}{3}\pi$$

Now recall the center of mass is given by

$$x_c \equiv \frac{\int_U x\rho(x,y)\,dV}{\int_U \rho(x,y)\,dV},\ y_c \equiv \frac{\int_U y\rho(x,y)\,dV}{\int_U \rho(x,y)\,dV}$$

Thus

$$x_c = \frac{\int_0^{2\pi}\int_0^{2+\cos\theta}(r\cos\theta)\,r^2 dr d\theta}{\frac{22}{3}\pi} = \frac{57}{44}$$

and

$$y_c = \frac{\int_0^{2\pi}\int_0^{2+\cos\theta}(r\sin\theta)\,r^2 dr d\theta}{\frac{22}{3}\pi} = 0$$

I think this might be impossible if you tried to do it in rectangular coordinates. However, it is just a little tedious in polar coordinates. Be sure you understand the set up. This is usually the thing which gives people the most trouble in these kinds of problems.

Example 23.2.6 *Let $f(x,y) = \sin\left(x^2+y^2\right)$ for (x,y) in the circle, $D = \left\{(x,y) : x^2+y^2 \le 9\right\}$. Find*

$$\int_D f dA.$$

You don't want to try this in rectangular coordinates even though the function is given in rectangular coordinates. You should change it to polar coordinates for two reasons. The first is that $x^2+y^2 = r^2$ and it is easier to look at $\sin\left(r^2\right)$ than $\sin\left(x^2+y^2\right)$. The main reason is that the integration is taking place on a circle which is a rectangle in polar coordinates and as explained earlier, it is easy to integrate over rectangles. In this case the rectangle is $[0,2\pi]\times[0,3]$. Thus the integral to work is

$$\int_0^{2\pi}\int_0^3 \sin\left(r^2\right)\overbrace{r dr d\theta}^{dA} = -\pi\cos 9 + \pi.$$

Example 23.2.7 *Remember the formula for the area between two polar graphs $r = f(\theta)$ and $r = g(\theta)$, $g(\theta) > f(\theta)$ for $\theta \in [a,b]$ is given by*

$$\frac{1}{2}\int_a^b \left(g(\theta)^2 - f(\theta)^2\right) d\theta.$$

Show this formula from one variable calculus follows from the form of the area increment given here.

Denote by R the region between the two graphs. Then you need to find

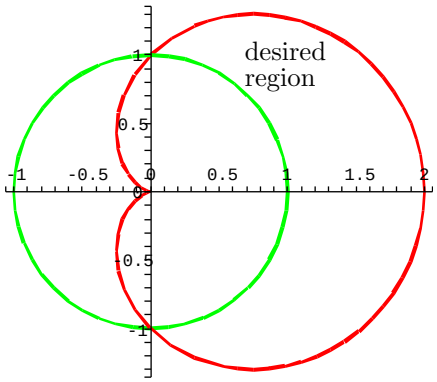
$$\int_R dA = \int_a^b \int_{f(\theta)}^{g(\theta)} r dr d\theta = \frac{1}{2}\int_a^b \left(g(\theta)^2 - f(\theta)^2\right) d\theta \tag{23.6}$$

which is the formula done earlier.

Here is an example.

Example 23.2.8 *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. It is very important to figure out which function is farther from the origin.



Then you need

$$\int_{-\pi/2}^{\pi/2}\int_{1}^{1+\cos\theta} r dr d\theta = 2 + \frac{1}{4}\pi$$

You could also work it using the formula derived in 23.6 which is like what you did in one variable calculus.

Example 23.2.9 *Let* $f(x,y) = e^{x^2+y^2}$ *for* (x,y) *in the pie shaped region* P *defined by* $r \in [0,2]$ *and* $\theta \in [0,\pi/6]$.

Be sure you can see why the integral wanted is

$$\int_0^2 \int_0^{\pi/6} e^{r^2} r d\theta dr = \frac{1}{12}e^4\pi - \frac{1}{12}\pi$$

Example 23.2.10 *Find*

$$\int_{-1}^{1}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \sqrt{1+x^2+y^2}dydx.$$

In this example you are integrating the function, $f(x,y) = \sqrt{1+x^2+y^2}$ over the circle of radius 1 centered at the origin. Therefore, changing to polar coordinates it equals

$$\int_0^{2\pi}\int_0^1 \sqrt{1+r^2}rdrd\theta = \frac{4}{3}\pi\sqrt{2} - \frac{2}{3}\pi$$

In this case, I think you could have done it without changing to polar coordinates but it would involve wading through much affliction and sorrow. Of course if you like adversity, you could try to do it this way.

Example 23.2.11 *Find* $\int_0^\infty e^{-x^2}dx$.

Let $I = \int_0^\infty e^{-x^2} dx$. Then

$$\begin{aligned} I^2 &= \left(\int_0^\infty e^{-x^2} dx\right)\left(\int_0^\infty e^{-y^2} dy\right) = \int_0^\infty \int_0^\infty e^{-\left(x^2+y^2\right)} dxdy \\ &= \int_0^{\pi/2} \int_0^\infty e^{-r^2} r dr d\theta = \frac{1}{4}\pi \end{aligned}$$

It follows $I = \frac{\sqrt{\pi}}{2}$. This is a very important formula. You showed, (hopefully) in one variable calculus that this integral exists. Now with the aid of polar coordinates you can actually find it.

23.3 Methods For Triple Integrals 2-7 Nov.

23.3.1 Definition Of The Integral

The integral of a function of three variables is similar to the integral of a function of two variables.

Definition 23.3.1 *For $i = 1, 2, 3$ let $\left\{\alpha_k^i\right\}_{k=-\infty}^\infty$ be points on $\mathbb{R}$ which satisfy*

$$\lim_{k\to\infty} \alpha_k^i = \infty, \ \lim_{k\to-\infty} \alpha_k^i = -\infty, \ \alpha_k^i < \alpha_{k+1}^i. \tag{23.7}$$

*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*

$$Q = \left[\alpha_k^1, \alpha_{k+1}^1\right] \times \left[\alpha_l^2, \alpha_{l+1}^2\right] \times \left[\alpha_p^3, \alpha_{p+1}^3\right]. \tag{23.8}$$

*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 23.3.2 *Let $f : \mathbb{R}^3 \to \mathbb{R}$ be a bounded function which equals zero outside 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 23.3.3 *For $i = 1, \cdots, n$, let $\{\alpha_k^i\}_{k=-\infty}^{\infty}$ be points on $\mathbb{R}$ which satisfy*

$$\lim_{k\to\infty} \alpha_k^i = \infty,\ \lim_{k\to-\infty} \alpha_k^i = -\infty,\ \alpha_k^i < \alpha_{k+1}^i. \tag{23.9}$$

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

$$Q = \prod_{i=1}^n \left[\alpha_{j_i}^i, \alpha_{j_i+1}^i\right]. \tag{23.10}$$

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 23.3.4 *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*

$$M_Q(f) \equiv \sup\{f(\mathbf{x}) : \mathbf{x} \in Q\},\ m_Q(f) \equiv \inf\{f(\mathbf{x}) : \mathbf{x} \in Q\}. \tag{23.11}$$

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.

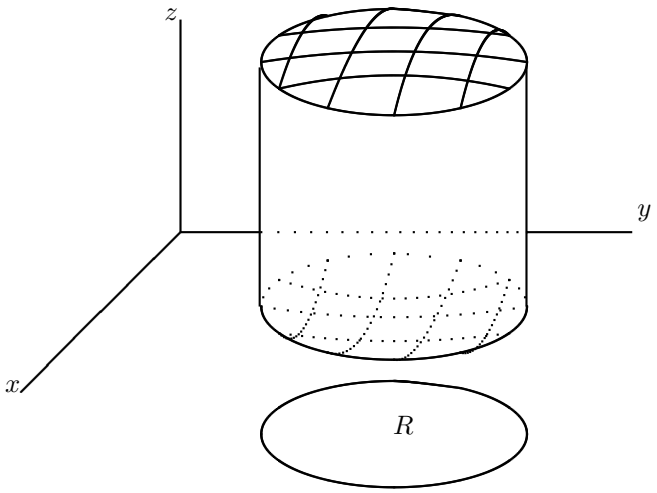
Definition 23.3.5 *For $E \subseteq \mathbb{R}^n$,*

$$\mathcal{X}_E(\mathbf{x}) \equiv \begin{cases} 1 \text{ if } \mathbf{x}\in E \\ 0 \text{ if } \mathbf{x}\notin E \end{cases}.$$

Define $\int_E f\,dV \equiv \int \mathcal{X}_E f\,dV$ when $f\mathcal{X}_E \in \mathcal{R}(\mathbb{R}^n)$.

23.3.2 Iterated Integrals

As before, the integral is often computed by using an iterated integral. In general it is impossible to set up an iterated integral for finding $\int_E f dV$ for arbitrary regions, E but when the region is sufficiently simple, one can make progress. Suppose the region, E over which the integral is to be taken is of the form $E = \{(x,y,z) : a(x,y) \le z \le b(x,y)\}$ for $(x,y) \in R$, a two dimensional region. This is illustrated in the following picture in which the bottom surface is the graph of $z = a(x,y)$ and the top is the graph of $z = b(x,y)$.



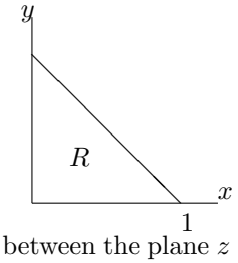
Then

$$\int_E f dV = \int_R \int_{a(x,y)}^{b(x,y)} f(x,y,z)\, dz dA$$

It might be helpful to think of $dV = dzdA$. Now $\int_{a(x,y)}^{b(x,y)} f(x,y,z)\,dz$ is a function of x and y and so you have reduced the triple integral to a double integral over R of this function of x and y. Similar reasoning would apply if the region in $\mathbb{R}^3$ were of the form $\{(x,y,z) : a(y,z) \le x \le b(y,z)\}$ or $\{(x,y,z) : a(x,z) \le y \le b(x,z)\}$.

Example 23.3.6 *Find the volume of the region, E in the first octant between $z = 1-(x+y)$ and $z = 0$.*

In this case, R is the region shown.



Thus the region, E is between the plane $z = 1-(x+y)$ on the top, $z = 0$ on the bottom, and over R shown above. Thus

$$\begin{aligned}\int_E 1dV &= \int_R \int_0^{1-(x+y)} dzdA \\ &= \int_0^1 \int_0^{1-x} \int_0^{1-(x+y)} dzdydx = \frac{1}{6}\end{aligned}$$

Of course iterated integrals have a life of their own although this will not be explored here. You can just write them down and go to work on them. Here are some examples.

Example 23.3.7 *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 23.3.8 *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.

$$\begin{aligned}\int_0^{3y} \int_0^{y+z} \cos(x+y)\,dx\,dz &= \int_0^{3y} \left(2\cos z \sin y \cos y + 2\sin z \cos^2 y - \sin z - \sin y\right) dz \\ &= -1 - 16\cos^5 y + 20\cos^3 y - 5\cos y - 3(\sin y)\,y + 2\cos^2 y.\end{aligned}$$

Finally, this last expression must be integrated from 0 to π. Thus

$$\int_0^\pi \int_0^{3y} \int_0^{y+z} \cos(x+y)\,dx\,dz\,dy$$

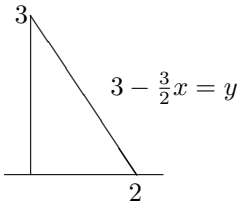
$$\begin{aligned}&= \int_0^\pi \left(-1 - 16\cos^5 y + 20\cos^3 y - 5\cos y - 3(\sin y)\,y + 2\cos^2 y\right) dy \\ &= -3\pi\end{aligned}$$

Example 23.3.9 *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

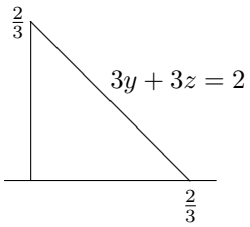
$$\int_0^3 \int_0^{\frac{2}{3}(3-y)} f(x,y)\, dx\,dy.$$

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 23.3.10 *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

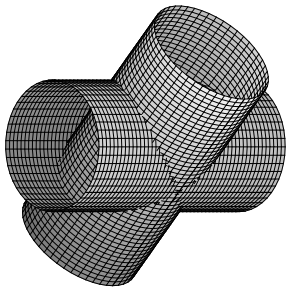
$$\int_0^{\frac{2}{3}} \int_0^{\frac{2}{3}-y} \int_0^{16-y^2} dx\,dz\,dy = \frac{860}{243}$$

Example 23.3.11 *Find the volume of the region determined by the intersection of the two cylinders,* $x^2+y^2 \le 9$ *and* $y^2+z^2 \le 9$.

The first listed cylinder intersects the xy plane in the disk, $x^2+y^2 \le 9$. 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.

23.3.3 Mass And Density

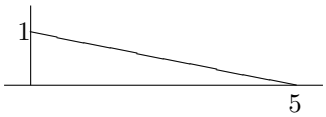
As an example of the use of triple integrals, consider a solid occupying a set of points, $U \subseteq \mathbb{R}^3$ having density ρ. Thus ρ is a function of position and the total mass of the solid equals

$$\int_U \rho \, dV.$$

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 23.3.12 *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 23.3.13 *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 23.3.14 *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 23.3.15 *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 $x + \frac{1}{7}y + \frac{1}{4}z = 1$.*

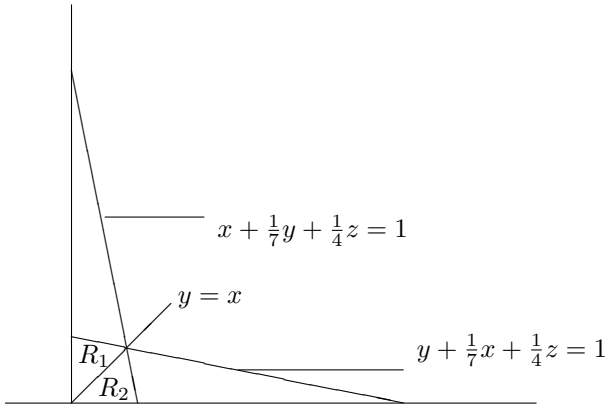
When $z = 0$, the plane $\frac{1}{7}x + y + \frac{1}{4}z = 1$ intersects the xy plane in the line whose equation is

$$\frac{1}{7}x + y = 1$$

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

$$\int_0^{7/8} \int_x^{1-\frac{x}{7}} \int_0^{4\left(1-\frac{1}{7}x-y\right)} dz\, dy\, dx + \int_0^{7/8} \int_y^{1-\frac{y}{7}} \int_0^{4\left(1-\frac{1}{7}y-x\right)} dz\, dx\, dy = \frac{7}{6}.$$

23.3.4 Exercises With Answers

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.

$$\int_0^2 \int_0^{30-5z} \int_{\frac{1}{5}y}^{6-z} \frac{\sin x}{x}\, dx\, dy\, dz = \int_0^2 \int_0^{6-z} \int_0^{5x} \frac{\sin x}{x}\, dy\, dx\, dz$$

$$= -5\sin 2\cos 6 + 5\cos 2\sin 6 + 10 - 5\sin 6$$

The Integral In Other Coordinates 8-10 Nov.

24.1 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 or parallelogram for example. In what follows, this is what it will be.

First recall the definition of the box product given in Definition 3.2.8 on Page 52. The absolute value of the box product of three vectors gave the volume of the parallelepiped determined by the three vectors.

Definition 24.1.1 *Let $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ be vectors in $\mathbb{R}^3$. The parallelepiped determined by these vectors will be denoted by $P(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ and it is defined as*

$$P(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3) \equiv \left\{ \sum_{j=1}^{3} s_j \mathbf{u}_j : s_j \in [0,1] \right\}.$$

Lemma 24.1.2 *The volume of the parallelepiped, $P(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ is given by $\left|\det\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix}\right|$ where $\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix}$ is the matrix having columns $\mathbf{u}_1, \mathbf{u}_2,$ and $\mathbf{u}_3$.*

Proof: Recall from the discussion of the box product or triple product,

$$\text{volume of } P(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3) \equiv |[\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]| = \left|\det\begin{pmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \mathbf{u}_3^T \end{pmatrix}\right|$$

where $\begin{pmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \mathbf{u}_3^T \end{pmatrix}$ is the matrix having rows equal to the vectors, $\mathbf{u}_1, \mathbf{u}_2$ and $\mathbf{u}_3$ arranged horizontally. Since the determinant of a matrix equals the determinant of its transpose,

$$\text{volume of } P(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3) = |[\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]| = \left|\det\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix}\right|.$$

This proves the lemma.

Definition 24.1.3 *In the case of two vectors, $P(\mathbf{u}_1, \mathbf{u}_2)$ will denote the parallelogram determined by $\mathbf{u}_1$ and $\mathbf{u}_2$. Thus*

$$P(\mathbf{u}_1, \mathbf{u}_2) \equiv \left\{ \sum_{j=1}^{2} s_j \mathbf{u}_j : s_j \in [0,1] \right\}.$$

Lemma 24.1.4 *The area of the parallelogram, $P(\mathbf{u}_1, \mathbf{u}_2)$ is given by $\left|\det\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{pmatrix}\right|$ where $\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{pmatrix}$ is the matrix having columns $\mathbf{u}_1$ and $\mathbf{u}_2$.*

Proof: Letting $\mathbf{u}_1 = (a,b)^T$ and $\mathbf{u}_2 = (c,d)^T$, consider the vectors in $\mathbb{R}^3$ defined by $\widehat{\mathbf{u}}_1 \equiv (a,b,0)^T$ and $\widehat{\mathbf{u}}_2 \equiv (c,d,0)^T$. Then the area of the parallelogram determined by the vectors, $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$ is the norm of the cross product of $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$. This follows directly from the geometric definition of the cross product given in Definition 3.2.2 on Page 49. But this is the same as the area of the parallelogram determined by the vectors $\mathbf{u}_1, \mathbf{u}_2$. Taking the cross product of $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$ yields $\mathbf{k}(ad - bc)$. Therefore, the norm of this cross product is

$$|ad - bc|$$

which is the same as

$$\left|\det\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{pmatrix}\right|$$

where $\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{pmatrix}$ denotes the matrix having the two vectors $\mathbf{u}_1, \mathbf{u}_2$ as columns. This proves the lemma.

It always works this way. The n dimensional volume of the n dimensional parallelepiped determined by the vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is always

$$\left|\det\begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}\right|$$

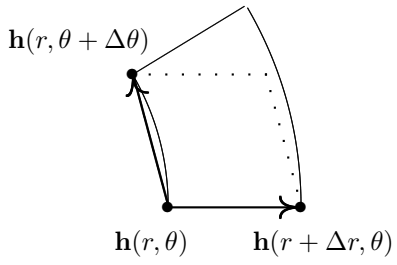
This general fact will not be used in what follows.

24.1.1 Review Of Polar Coordinates

Earlier it was shown based on geometric reasoning that the appropriate increment of area in polar coordinates is $rdrd\theta$. Consider this again in a slightly different way. Recall the transformation equations for polar coordinates,

$$\mathbf{h}(r,\theta) = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix}$$

As before, you need to approximate the area of the curvy shape shown in the following picture.



As $\Delta\theta$ gets smaller, the curvy shape will be better and better approximated by the dotted parallelogram shown in the picture. So what is the area of this parallelogram? The parallelogram is determined by the vectors, $\mathbf{h}(r,\theta+\Delta\theta)-\mathbf{h}(r,\theta)$ and $\mathbf{h}(r+\Delta r,\theta)-\mathbf{h}(r,\theta)$. For very small Δr and $\Delta\theta$, these two vectors are essentially equal to $\mathbf{h}_\theta(r,\theta)\Delta\theta$ and $\mathbf{h}_r(r,\theta)\Delta r$. Therefore, from the above discussion, the increment of area is essentially equal to

$$\left|\det\left(\begin{array}{cc}\mathbf{h}_\theta(r,\theta)\Delta\theta & \mathbf{h}_r(r,\theta)\Delta r\end{array}\right)\right| = \left|\det\left(\begin{array}{cc}\mathbf{h}_\theta(r,\theta) & \mathbf{h}_r(r,\theta)\end{array}\right)\right|\Delta r\Delta\theta$$

Now

$$\mathbf{h}_\theta(r,\theta)=\begin{pmatrix}-r\sin\theta\\ r\cos\theta\end{pmatrix},\ \mathbf{h}_r(r,\theta)=\begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix}$$

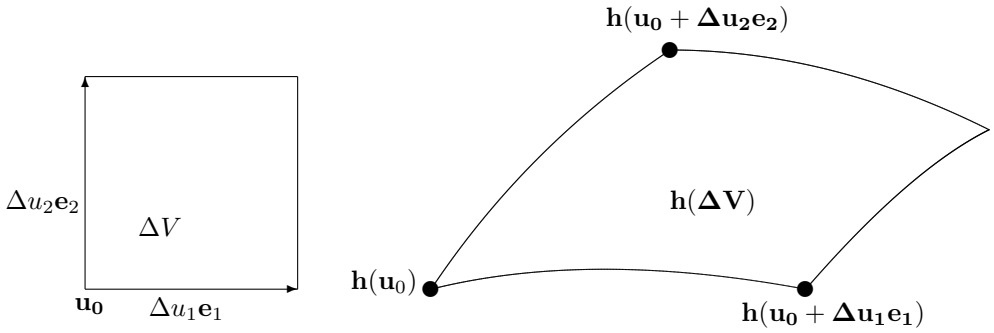
and so the above reduces to

$$\left|\begin{pmatrix}-r\sin\theta & \cos\theta\\ r\cos\theta & \sin\theta\end{pmatrix}\right|\Delta r\Delta\theta = r\Delta r\Delta\theta$$

and this is why the area increment in polar coordinates is $rdrd\theta$. Note the emphasis on algebraic techniques to find the area increment. The same approach works for other coordinates, not just polar coordinates.

24.1.2 General Two Dimensional Coordinates

Suppose U is a set in $\mathbb{R}^2$ and $\mathbf{h}$ is a C^1 function[1] mapping U one to one onto $\mathbf{h}(U)$, a set in $\mathbb{R}^2$. Consider a small square inside U. The following picture is of such a square having a corner at the point, $\mathbf{u}_0$ and sides as indicated. The image of this square is also represented.



For small Δu_i you would expect the sides going from $\mathbf{h}(\mathbf{u}_0)$ to $\mathbf{h}(\mathbf{u}_0+\Delta u_1\mathbf{e}_1)$ and from $\mathbf{h}(\mathbf{u}_0)$ to $\mathbf{h}(\mathbf{u}_0+\Delta u_2\mathbf{e}_2)$ to be almost the same as the vectors, $\mathbf{h}(\mathbf{u}_0+\Delta u_1\mathbf{e}_1)-\mathbf{h}(\mathbf{u}_0)$ and $\mathbf{h}(\mathbf{u}_0+\Delta u_2\mathbf{e}_2)-\mathbf{h}(\mathbf{u}_0)$ which are approximately equal to $\frac{\partial\mathbf{h}}{\partial u_1}(\mathbf{u}_0)\Delta u_1$ and $\frac{\partial\mathbf{h}}{\partial u_2}(\mathbf{u}_0)\Delta u_2$ respectively. Therefore, the area of $\mathbf{h}(\Delta V)$ for small Δu_i is essentially equal to the area of the parallelogram determined by the two vectors, $\frac{\partial\mathbf{h}}{\partial u_1}(\mathbf{u}_0)\Delta u_1$ and $\frac{\partial\mathbf{h}}{\partial u_2}(\mathbf{u}_0)\Delta u_2$. By Lemma 24.1.4 this equals

$$\left|\det\left(\begin{array}{cc}\frac{\partial\mathbf{h}}{\partial u_1}(\mathbf{u}_0)\Delta u_1 & \frac{\partial\mathbf{h}}{\partial u_2}(\mathbf{u}_0)\Delta u_2\end{array}\right)\right| = \left|\det\left(\begin{array}{cc}\frac{\partial\mathbf{h}}{\partial u_1}(\mathbf{u}_0) & \frac{\partial\mathbf{h}}{\partial u_2}(\mathbf{u}_0)\end{array}\right)\right|\Delta u_1\Delta u_2$$

[1] By this is meant $\mathbf{h}$ is the restriction to U of a function defined on an open set containing U which is C^1. If you like, you can assume U is open but this is not necessary. Neither is C^1.

Thus an infinitesimal chunk of area in $\mathbf{h}(U)$ located at $\mathbf{u}_0$ is of the form

$$\left|\det\left(\begin{array}{cc} \frac{\partial \mathbf{h}}{\partial u_1}(\mathbf{u}_0) & \frac{\partial \mathbf{h}}{\partial u_2}(\mathbf{u}_0) \end{array}\right)\right| dV$$

where dV is a corresponding chunk of area located at the point $\mathbf{u}_0$. This shows the following change of variables formula is reasonable.

$$\int_{\mathbf{h}(U)} f(\mathbf{x})\, dV(\mathbf{x}) = \int_U f(\mathbf{h}(\mathbf{u}))\left|\det\left(\begin{array}{cc} \frac{\partial \mathbf{h}}{\partial u_1}(\mathbf{u}) & \frac{\partial \mathbf{h}}{\partial u_2}(\mathbf{u}) \end{array}\right)\right| dV(\mathbf{u})$$

Definition 24.1.5 *Let $\mathbf{h} : U \to \mathbf{h}(U)$ be a one to one and C^1 mapping. The (volume) area element in terms of $\mathbf{u}$ is defined as $\left|\det\left(\begin{array}{cc} \frac{\partial \mathbf{h}}{\partial u_1}(\mathbf{u}) & \frac{\partial \mathbf{h}}{\partial u_2}(\mathbf{u}) \end{array}\right)\right| dV(\mathbf{u})$. The factor, $\left|\det\left(\begin{array}{cc} \frac{\partial \mathbf{h}}{\partial u_1}(\mathbf{u}) & \frac{\partial \mathbf{h}}{\partial u_2}(\mathbf{u}) \end{array}\right)\right|$ is called the Jacobian. It equals*

$$\left|\det\begin{pmatrix} \frac{\partial h_1}{\partial u_1}(u_1,u_2) & \frac{\partial h_1}{\partial u_2}(u_1,u_2) \\ \frac{\partial h_2}{\partial u_1}(u_1,u_2) & \frac{\partial h_2}{\partial u_2}(u_1,u_2) \end{pmatrix}\right|.$$

It is traditional to call two dimensional volumes area. However, it is probably better to simply always refer to it as volume. Thus there is 2 dimensional volume, 3 dimensional volume, etc. Sometimes you can get confused by too many different words to describe things which are really not essentially different.

Example 24.1.6 *Find the area element for polar coordinates.*

Here the $\mathbf{u}$ coordinates are θ and r. The polar coordinate transformations are

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix}$$

Therefore, the volume (area) element is

$$\left|\det\begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}\right| d\theta dr = r\, d\theta dr.$$

Example 24.1.7 *Suppose $x = u^2 - v^2$ and $y = 2uv$ Find the area element in terms of u and v.*

You are given

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} u^2 - v^2 \\ 2uv \end{pmatrix}$$

and so the area element is

$$\left|\det\begin{pmatrix} 2u & -2v \\ 2v & 2u \end{pmatrix}\right| dudv = \left(4u^2 + 4v^2\right) dudv$$

Example 24.1.8 *Suppose new coordinates are given by $u = x + y$ and $v = y/x$ find the area element in terms of the new coordinates, u and v.*

You need to solve for x and y first. This yields $y = \frac{uv}{v+1}, x = \frac{u}{v+1}$. Thus

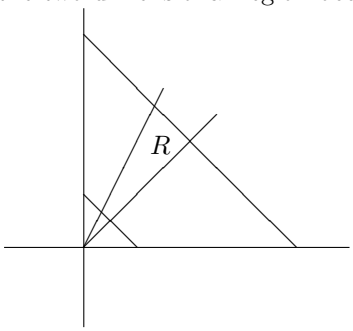
$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{u}{v+1} \\ \frac{uv}{v+1} \end{pmatrix}$$

and so the area increment equals

$$\left|\det\begin{pmatrix} \frac{1}{v+1} & -\frac{u}{(v+1)^2} \\ \frac{v}{v+1} & \frac{u}{v+1} - v\frac{u}{(v+1)^2} \end{pmatrix}\right| dudv = \left|\frac{u}{(v+1)^2}\right| dudv$$

Example 24.1.9 *The area density is given by $\rho(x,y,z) = x$ and a plate occupies the region between $x+y=1, x+y=4$ and the lines $y=x$ and $y=2x$. Find the x coordinate of the center of mass of this plate.*

Here is a picture of the two dimensional region occupied by this plate.



It is labeled as R. Thus the total mass and the x coordinate of the center of mass are given respectively as

$$\int_R x dA,\ x_c = \frac{\int_R x^2 dA}{\int_R x dA}$$

So now you just set up the iterated integrals and go to work. Good luck if you try this. It is much better to change the variables. Let $u = x+y$ and $v = y/x$ as in Example 24.1.8. In these new coordinates the horrible quadrilateral becomes the rectangle

$$(u,v) \in [1,4] \times [1,2].$$

That is u goes between 1 and 4 while v goes between 1 and 2. Remember it is easy to integrate over rectangles. Thus using the result of Example 24.1.8

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{u}{v+1} \\ \frac{uv}{v+1} \end{pmatrix}$$

and the area element is

$$\frac{u}{(v+1)^2} dudv$$

Therefore, the total mass is

$$\int_1^4 \int_1^2 \overbrace{\left(\frac{u}{v+1}\right)}^{x} \overbrace{\frac{u}{(v+1)^2} dudv}^{dA} = \frac{49}{200}$$

and the x coordinate of the center of mass is

$$\frac{\int_1^4 \int_1^2 \left(\frac{u}{v+1}\right)^2 \frac{u}{(v+1)^2} dudv}{\frac{49}{200}} = \frac{117}{196}$$

It is a tedious problem but not all that hard if you change the variables.

24.1.3 Three Dimensions

Quiz

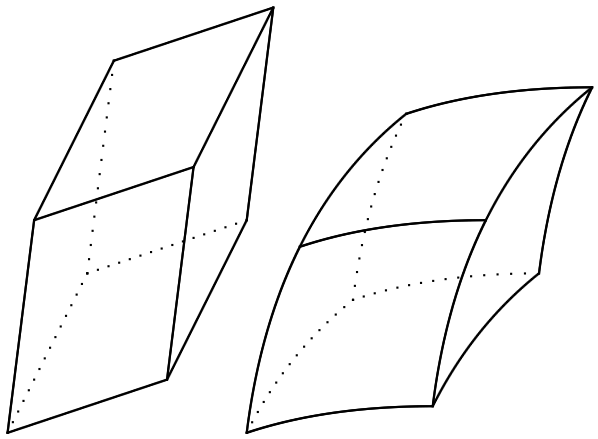
1. Find $\int_0^1 \int_{\sqrt{y}}^1 \sin\left(x^3\right) dxdy$.

2. Find the x coordinate of the center of mass of the solid whose density is $\delta(x,y,z) = x$ which occupies the three dimensional region below the plane $z = y$ and above the triangular region in the xy plane determined by $x \in [0,1]$ and $0 \leq y \leq x$.

3. Maximize xy subject to the constraint $x^2 + y^2 = 1$.

4. Find the tangent plane to the surface, $z^2 + x^2 + 3y^2 = 5$ at the point $(1,1,1)$.

The situation is no different for coordinate systems in any number of dimensions although I will concentrate here on three dimensions. A rectangular chunk of volume in the $\mathbf{u}$ space corresponds to the curvy parallelepiped shown below which is approximated by the parallelepiped shown on the left determined by the 3 vectors $\left\{\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} du_i\right\}_{i=1}^3$ for $\mathbf{x} = \mathbf{f}(\mathbf{u})$ where $\mathbf{u} \in U$, a subset of $\mathbb{R}^3$ and $\mathbf{x}$ is a point in $V = \mathbf{f}(U)$, a subset of 3 dimensional space.



Thus, letting the Cartesian coordinates of $\mathbf{x}$ be given by $\mathbf{x} = (x_1, x_2, x_3)^T$, 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 3 vectors $\left\{\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_i} du_i\right\}_{i=1}^3$. From Lemma 24.1.2, the volume of this infinitesimal parallelepiped located at $\mathbf{f}(\mathbf{u}_0)$ is given by

$$\begin{aligned} \left|\left[\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1} du_1, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_2} du_2, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_3} du_3\right]\right| &= \left|\left[\frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1}, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_2}, \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_3}\right]\right| du_1 du_2 du_3 \\ &= \left|\det\left(\begin{array}{ccc} \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1} & \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_2} & \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_3} \end{array}\right)\right| du_1 du_2 du_3 \end{aligned} \tag{24.1}$$

There is also no change in going to higher dimensions than 3.

Definition 24.1.10 *Let $\mathbf{x} = \mathbf{f}(\mathbf{u})$ be as described above. Then for $n = 2, 3$, the symbol, $\frac{\partial(x_1, \cdots x_n)}{\partial(u_1, \cdots, u_n)}$, called the Jacobian determinant, is defined by*

$$\det\left(\begin{array}{ccc} \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_1} & \cdots & \frac{\partial \mathbf{x}(\mathbf{u}_0)}{\partial u_n} \end{array}\right) \equiv \frac{\partial(x_1, \cdots, x_n)}{\partial(u_1, \cdots, u_n)}.$$

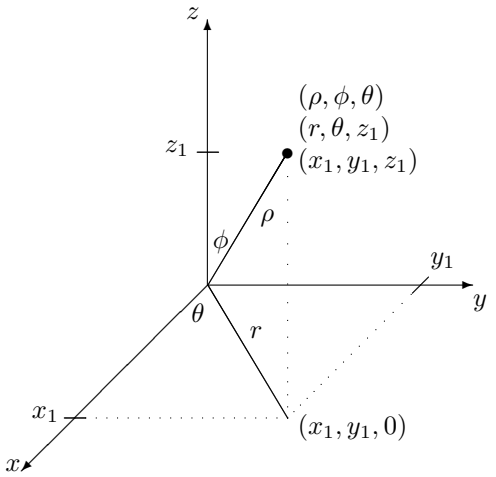
Also, the symbol, $\left|\frac{\partial(x_1, \cdots, x_n)}{\partial(u_1, \cdots, u_n)}\right| du_1 \cdots du_n$ is called the volume element.

This has given motivation for the following fundamental procedure often called the **change of variables formula** which holds under fairly general conditions.

Procedure 24.1.11 *Suppose U is an open subset of $\mathbb{R}^n$ for $n = 2, 3$ and suppose $\mathbf{f} : U \to \mathbf{f}(U)$ is a C^1 function which is one to one, $\mathbf{x} = \mathbf{f}(\mathbf{u})$.[2] Then if $h : \mathbf{f}(U) \to \mathbb{R}$,*

$$\int_U h(\mathbf{f}(\mathbf{u})) \left| \frac{\partial(x_1, \cdots, x_n)}{\partial(u_1, \cdots, u_n)} \right| dV = \int_{\mathbf{f}(U)} h(\mathbf{x})\, dV.$$

Now consider spherical coordinates. Recall the geometrical meaning of these coordinates illustrated in the following picture.



Thus there is a relationship between these coordinates and rectangular coordinates given by

$$x = \rho \sin\phi \cos\theta, y = \rho \sin\phi \sin\theta,\ z = \rho \cos\phi \tag{24.2}$$

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. From the above argument, the volume element is

$$\left| \det \left(\begin{array}{ccc} \frac{\partial \mathbf{x}(\rho_0,\phi_0,\theta_0)}{\partial \rho} & \frac{\partial \mathbf{x}(\rho_0,\phi_0,\theta_0)}{\partial \phi} & \frac{\partial \mathbf{x}(\rho_0,\phi_0,\theta_0)}{\partial \theta} \end{array} \right) \right| d\rho d\theta d\phi.$$

The mapping between spherical and rectangular coordinates is written as

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \rho \sin\phi \cos\theta \\ \rho \sin\phi \sin\theta \\ \rho \cos\phi \end{pmatrix} = \mathbf{f}(\rho, \phi, \theta) \tag{24.3}$$

[2] 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 set U 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. For a typical precise theorem see the appendix on the Riemann integral.

Therefore, $\det\left(\frac{\partial \mathbf{x}(\rho_0,\phi_0,\theta_0)}{\partial\rho}, \frac{\partial \mathbf{x}(\rho_0,\phi_0,\theta_0)}{\partial\phi}, \frac{\partial \mathbf{x}(\rho_0,\phi_0,\theta_0)}{\partial\theta}\right) =$

$$\det\begin{pmatrix} \sin\phi\cos\theta & \rho\cos\phi\cos\theta & -\rho\sin\phi\sin\theta \\ \sin\phi\sin\theta & \rho\cos\phi\sin\theta & \rho\sin\phi\cos\theta \\ \cos\phi & -\rho\sin\phi & 0 \end{pmatrix} = \rho^2\sin\phi$$

which is positive because $\phi \in [0,\pi]$.

Example 24.1.12 *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 24.3 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 24.1.13 *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 $r\,d\theta\,dr\,dz$.

Example 24.1.14 *This example uses spherical coordinates to verify an important conclusion about gravitational force. Let the hollow sphere, H be defined by $a^2 < x^2+y^2+z^2 < b^2$ and suppose this hollow sphere has constant density taken to equal α. 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 $\left(\alpha G\int_H \frac{(z-z_0)}{\left[x^2+y^2+(z-z_0)^2\right]^{3/2}}\,dV\right)\mathbf{k}$ 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*[3].

[3]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.

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

$$\int_H (x\mathbf{i} + y\mathbf{j} + (z - z_0)\,\mathbf{k}) \frac{1}{\left(x^2 + y^2 + (z - z_0)^2\right)^{3/2}} G\alpha\, dV.$$

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

$$\alpha G\mathbf{k} \int_H \frac{(z - z_0)}{\left[x^2 + y^2 + (z - z_0)^2\right]^{3/2}}\, dV$$

as claimed. Now for the interesting part, the integral is evaluated. In spherical coordinates this integral is.

$$\int_0^{2\pi} \int_a^b \int_0^{\pi} \frac{(\rho\cos\phi - z_0)\,\rho^2 \sin\phi}{\left(\rho^2 + z_0^2 - 2\rho z_0\cos\phi\right)^{3/2}}\, d\phi\, d\rho\, d\theta. \tag{24.4}$$

Rewrite the inside integral and use integration by parts to obtain this inside integral equals

$$\frac{1}{2z_0} \int_0^{\pi} \left(\rho^2\cos\phi - \rho z_0\right) \frac{(2z_0\rho\sin\phi)}{\left(\rho^2 + z_0^2 - 2\rho z_0\cos\phi\right)^{3/2}}\, d\phi =$$

$$\frac{1}{2z_0}\left(-2\frac{-\rho^2 - \rho z_0}{\sqrt{(\rho^2 + z_0^2 + 2\rho z_0)}} + 2\frac{\rho^2 - \rho z_0}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0)}} - \int_0^{\pi} 2\rho^2 \frac{\sin\phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0\cos\phi)}}\, d\phi\right). \tag{24.5}$$

There are some cases to consider here.

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

$$\frac{2\rho\,(\rho + z_0)}{\sqrt{(\rho + z_0)^2}} + \frac{2\rho\,(\rho - z_0)}{\sqrt{(\rho - z_0)^2}} = \frac{2\rho\,(\rho + z_0)}{(\rho + z_0)} + \frac{2\rho\,(\rho - z_0)}{\rho - z_0} = 4\rho$$

and so the expression in 24.5 equals

$$\frac{1}{2z_0}\left(4\rho - \int_0^{\pi} 2\rho^2 \frac{\sin\phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0\cos\phi)}}\, d\phi\right)$$

$$= \frac{1}{2z_0}\left(4\rho - \frac{1}{z_0}\int_0^{\pi} \rho \frac{2\rho z_0\sin\phi}{\sqrt{(\rho^2 + z_0^2 - 2\rho z_0\cos\phi)}}\, d\phi\right)$$

$$= \frac{1}{2z_0}\left(4\rho - \frac{2\rho}{z_0}\left(\rho^2 + z_0^2 - 2\rho z_0\cos\phi\right)^{1/2}|_0^\pi\right)$$
$$= \frac{1}{2z_0}\left(4\rho - \frac{2\rho}{z_0}\left[(\rho + z_0) - (\rho - z_0)\right]\right) = 0.$$

Therefore, in this case the inner integral of 24.4 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 24.5 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, 24.5 equals

$$\frac{1}{2z_0}\left(-\int_0^\pi 2\rho^2 \frac{\sin\phi}{\sqrt{\left(\rho^2 + z_0^2 - 2\rho z_0\cos\phi\right)}}\,d\phi\right)$$
$$= \frac{-\rho}{2z_0^2}\left(\int_0^\pi \frac{2\rho z_0\sin\phi}{\sqrt{\left(\rho^2 + z_0^2 - 2\rho z_0\cos\phi\right)}}\,d\phi\right)$$

which equals

$$\frac{-\rho}{z_0^2}\left(\left(\rho^2 + z_0^2 - 2\rho z_0\cos\phi\right)^{1/2}|_0^\pi\right)$$
$$= \frac{-\rho}{z_0^2}\left[(\rho + z_0) - (z_0 - \rho)\right] = -\frac{2\rho^2}{z_0^2}.$$

Thus the inner integral of 24.4 reduces to the above simple expression. Therefore, 24.4 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

$$\alpha G\mathbf{k}\int_H \frac{(z - z_0)}{\left[x^2 + y^2 + (z - z_0)^2\right]^{3/2}}\,dV = \alpha G\mathbf{k}\left(-\frac{4}{3}\pi\frac{b^3 - a^3}{z_0^2}\right) = -\mathbf{k}G\frac{\text{total mass}}{z_0^2}.$$

24.1.4 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 density is $4u + 4v + 3w$

   $$\int_R x \, dV = \int_Q (4u + 4v + 3w)\,(8)\, dV$$

   $$= \int_0^1 \int_0^1 \int_0^1 (4u + 4v + 3w)\,(8)\, du\, dv\, dw = 44.$$

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 $\int_S \frac{y}{x}\,dV$ 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 $\int_S \left(\left(\frac{y}{x}\right)^2 + 1\right) dV$ 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. $\int_S \sin\left(4x^2+4y^2\right) dV$ where $r \le 2$ and $0 \le \theta \le \frac{1}{6}\pi$.

    Answer:

$$\int_0^{\frac{1}{6}\pi}\int_0^2 \sin\left(4r^2\right) r\,dr\,d\theta = -\frac{1}{48}\pi\cos 16 + \frac{1}{48}\pi$$

19. Find $\int_S e^{2x^2+2y^2}\,dV$ 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:

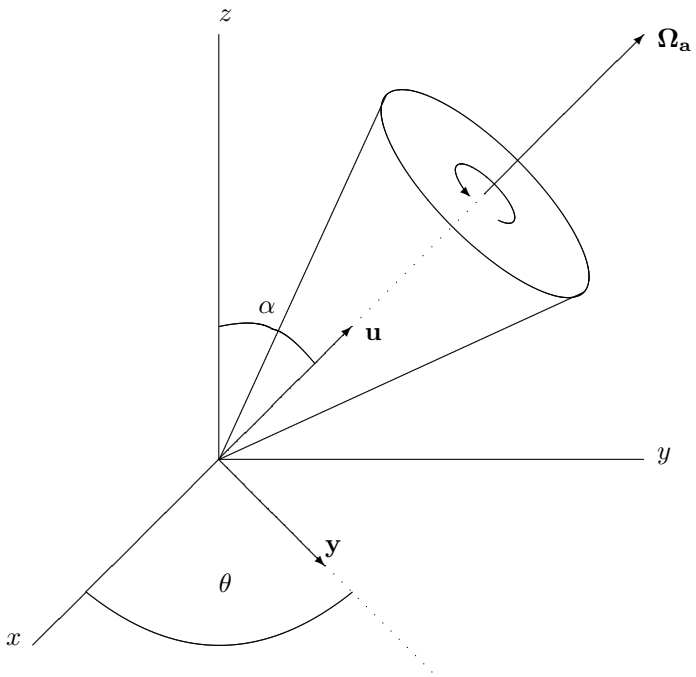
    Using cylindrical coordinates, the integral is $\int_0^{2\pi}\int_0^R \int_{-\sqrt{R^2-r^2}}^{\sqrt{R^2-r^2}} r\,dz\,dr\,d\theta = \frac{4}{3}\pi R^3$.

24.2 The Moment Of Inertia *

In order to appreciate the importance of this concept, it is necessary to discuss its physical significance.

24.2.1 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 [24].

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 16.4.2 on Page 300, 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

$$\mathbf{L} \equiv \sum_{i=1}^{N} \mathbf{r}_i \times m_i\mathbf{v}_i \tag{24.6}$$

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

$$\mathbf{L} = I\omega\mathbf{u}. \tag{24.7}$$

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

$$(\mathbf{u} \cdot \mathbf{i}) = (\mathbf{y} \cdot \mathbf{i}) = \sin\alpha \cos\theta. \tag{24.8}$$

Now also from 24.6,

$$\begin{aligned}\frac{d\mathbf{L}}{dt} &= \sum_{i=1}^{N} m_i \overbrace{\mathbf{r}_i' \times \mathbf{v}_i}^{=0} + \mathbf{r}_i \times m_i \mathbf{v}_i' \\ &= \sum_{i=1}^{N} \mathbf{r}_i \times m_i \mathbf{v}_i' = -\sum_{i=1}^{N} \mathbf{r}_i \times m_i g \mathbf{k}\end{aligned}$$

where g is the acceleration of gravity. From 24.7, 24.8, and the above,

$$\begin{aligned}\frac{d\mathbf{L}}{dt}\cdot\mathbf{i} &= I\omega\left(\frac{d\mathbf{u}}{dt}\cdot\mathbf{i}\right) = I\omega\left(\frac{d\mathbf{y}}{dt}\cdot\mathbf{i}\right)\\ &= \left(-I\omega\sin\alpha\sin\theta\right)\theta' = -\sum_{i=1}^{N}\mathbf{r}_i\times m_i g\mathbf{k}\cdot\mathbf{i}\\ &= -\sum_{i=1}^{N} m_i g\mathbf{r}_i\cdot\mathbf{k}\times\mathbf{i} = -\sum_{i=1}^{N} m_i g\mathbf{r}_i\cdot\mathbf{j}. \end{aligned} \tag{24.9}$$

To simplify this further, recall the following definition of the center of mass.

Definition 24.2.1 *Define the total mass, M by*

$$M = \sum_{i=1}^{N} m_i$$

and the center of mass, $\mathbf{r}_0$ by

$$\mathbf{r}_0 \equiv \frac{\sum_{i=1}^{N}\mathbf{r}_i m_i}{M}. \tag{24.10}$$

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 24.9,

$$\left(-I\omega\sin\alpha\sin\theta\right)\theta' = -Mg\left|\mathbf{r}_0\right|\sin\alpha\sin\theta.$$

which shows

$$\theta' = \frac{Mg\left|\mathbf{r}_0\right|}{I\omega}. \tag{24.11}$$

From 24.11, 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.

24.2.2 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 277.) The total kinetic energy of the collection of masses is then

$$E = \sum_{i=1}^{N} \frac{1}{2} m_i \left|\mathbf{r}_i'\right|^2 . \tag{24.12}$$

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 24.12 the total kinetic energy is

$$\begin{aligned} E &= \sum_{i=1}^{N} \frac{1}{2} m_i \left|\mathbf{r}_i' - \mathbf{r}_0' + \mathbf{r}_0'\right|^2 \\ &= \sum_{i=1}^{N} \frac{1}{2} m_i \left[\left|\mathbf{r}_i' - \mathbf{r}_0'\right|^2 + \left|\mathbf{r}_0'\right|^2 + 2\left(\mathbf{r}_i' - \mathbf{r}_0' \cdot \mathbf{r}_0'\right)\right] . \end{aligned} \tag{24.13}$$

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 24.10

$$\begin{aligned} \sum_{i=1}^{N} m_i \left(\mathbf{r}_i - \mathbf{r}_0\right) &= \sum_{i=1}^{N} m_i \mathbf{r}_i - \sum_{i=1}^{N} m_i \mathbf{r}_0 \\ &= \sum_{i=1}^{N} m_i \mathbf{r}_i - \sum_{i=1}^{N} m_i \left(\frac{\sum_{i=1}^{N} \mathbf{r}_i m_i}{\sum_{i=1}^{N} m_i}\right) = \mathbf{0}. \end{aligned}$$

Let $M \equiv \sum_{i=1}^{N} m_i$ be the total mass. Then 24.13 reduces to

$$\begin{aligned} E &= \sum_{i=1}^{N} \frac{1}{2} m_i \left[\left|\mathbf{r}_i' - \mathbf{r}_0'\right|^2 + \left|\mathbf{r}_0'\right|^2\right] \\ &= \frac{1}{2} M \left|\mathbf{r}_0'\right|^2 + \sum_{i=1}^{N} \frac{1}{2} m_i \left|\mathbf{r}_i' - \mathbf{r}_0'\right|^2 . \end{aligned} \tag{24.14}$$

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 one 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.

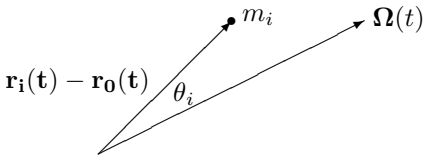
Now let $\mathbf{e}_1(t), \mathbf{e}_2(t)$, and $\mathbf{e}_3(t)$ be an orthonormal set of vectors which is fixed in the body undergoing rigid body motion. This means that $\mathbf{r}_i(t) - \mathbf{r}_0(t)$ has components which are constant in t with respect to the vectors, $\mathbf{e}_i(t)$. By Theorem 16.4.2 on Page 300 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 24.14,

$$\begin{aligned} E &= \frac{1}{2}M|\mathbf{r}_0'|^2 + \sum_{i=1}^{N}\frac{1}{2}m_i|\mathbf{\Omega}(t) \times (\mathbf{r}_i(t) - \mathbf{r}_0(t))|^2 \\ &= \frac{1}{2}M|\mathbf{r}_0'|^2 + \frac{1}{2}\left(\sum_{i=1}^{N} m_i|\mathbf{r}_i(t) - \mathbf{r}_0(t)|^2 \sin^2\theta_i\right)|\mathbf{\Omega}(t)|^2 \\ &= \frac{1}{2}M|\mathbf{r}_0'|^2 + \frac{1}{2}\left(\sum_{i=1}^{N} m_i|\mathbf{r}_i(0) - \mathbf{r}_0(0)|^2 \sin^2\theta_i\right)|\mathbf{\Omega}(t)|^2 \end{aligned}$$

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.

24.3 Finding The Moment Of Inertia And Center Of Mass 13 Nov.

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. This leads to the following definition.

Definition 24.3.1 *Let a solid occupy a region R such that its density is $\delta(\mathbf{x})$ for $\mathbf{x}$ a point in R and let L be a line. For $\mathbf{x} \in R$, let $l(\mathbf{x})$ be the distance from the point, $\mathbf{x}$ to the line L. The* ***moment of inertia*** *of the solid is defined as*

$$\int_R l(\mathbf{x})^2 \delta(\mathbf{x})\, dV.$$

Letting (x_c, y_c, z_c) denote the Cartesian coordinates of the ***center of mass****,*

$$\begin{aligned} x_c &= \frac{\int_R x\delta(\mathbf{x})\,dV}{\int_R \delta(\mathbf{x})\,dV}, y_c = \frac{\int_R y\delta(\mathbf{x})\,dV}{\int_R \delta(\mathbf{x})\,dV}, \\ z_c &= \frac{\int_R z\delta(\mathbf{x})\,dV}{\int_R \delta(\mathbf{x})\,dV} \end{aligned}$$

where x, y, z are the Cartesian coordinates of the point at $\mathbf{x}$.

Example 24.3.2 *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.

$$\int_R \left(x^2+y^2\right)\rho(\mathbf{x})\;dV. \tag{24.15}$$

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 24.3.3 *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, z_c. In polar coordinates, $\rho = r$ and so,

$$z_c = \frac{\int_R z\rho\,dV}{\int_R \rho\,dV} = \frac{\int_0^{2\pi}\int_0^3\int_0^{9-r^2} zr^2\,dz\,dr\,d\theta}{\int_0^{2\pi}\int_0^3\int_0^{9-r^2} r^2\,dz\,dr\,d\theta} = \frac{18}{7}.$$

Thus the center of mass will be $\left(0,0,\frac{18}{7}\right)$.

A short comment about terminology is in order. When the density is constant, the center of mass is called the **centroid.** Thus the centroid is a purely geometrical concept because the densities will cancel from the integrals.

24.4 Exercises With Answers

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.

   The region, R is a dome shaped region above the circle centered at the origin having radius 2. Therefore, using polar or cylindrical coordinates

   $$z_c = \frac{\int_R z\sigma dV}{\int_R \sigma dV} = \frac{\int_0^{2\pi}\int_0^2\int_0^{4-r^2} zrdzdrd\theta}{\int_0^{2\pi}\int_0^2\int_0^{4-r^2} rdzdrd\theta} = \frac{4}{3}$$

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 $\sigma(x,y,z) = z$.

   This problem is just like the one above except here the density is not constant. Thus

   $$z_c = \frac{\int_R z^2 dV}{\int_R z dV} = \frac{\int_0^{2\pi}\int_0^2\int_0^{4-r^2} z^2 rdzdrd\theta}{\int_0^{2\pi}\int_0^2\int_0^{4-r^2} zrdzdrd\theta} = 2$$

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 $\sigma = 1$.

   To find where (x,y) is you let $-y^2 + 8 = 2x^2 + y^2$ and this shows the two surfaces intersect in the circle $x^2 + y^2 = 4$. Using cylindrical coordinates,

   $$\begin{aligned} z_c &= \frac{\int_R z dV}{\int_R 1 dV} = \frac{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} zrdzdrd\theta}{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} rdzdrd\theta} \\ &= \frac{\left(\frac{224}{3}\pi\right)}{16\pi} = \frac{14}{3} \end{aligned}$$

   You can find the the others the same way.

   $$x_c = \frac{\int_R x dV}{\int_R 1 dV} = \frac{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} (r\cos(\theta))\, rdzdrd\theta}{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} rdzdrd\theta} = 0$$

   $$y_c = \frac{\int_R y dV}{\int_R 1 dV} = \frac{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} (r\sin(\theta))\, rdzdrd\theta}{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} rdzdrd\theta} = 0$$

   Thus the center of mass is $\left(0,0,\frac{14}{3}\right)$. The mass is

   $$\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} rdzdrd\theta = 16\pi$$

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 $\sigma(x,y,z)=x^2$.

   This is just like the problem above only now the density is not constant.

$$z_c=\frac{\int_R zx^2dV}{\int_R x^2dV}=\frac{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} z\left(r^2\cos^2(\theta)\right)rdzdrd\theta}{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} r\left(r^2\cos^2(\theta)\right)dzdrd\theta}=\frac{11}{2}$$

$$y_c=\frac{\int_R yx^2dV}{\int_R x^2dV}=\frac{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} \left(r\sin(\theta)\right)\left(r^2\cos^2(\theta)\right)rdzdrd\theta}{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} r\left(r^2\cos^2(\theta)\right)dzdrd\theta}=0$$

$$x_c=\frac{\int_R xx^2dV}{\int_R x^2dV}=\frac{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} \left(r\cos(\theta)\right)\left(r^2\cos^2(\theta)\right)rdzdrd\theta}{\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} r\left(r^2\cos^2(\theta)\right)dzdrd\theta}=0$$

   So in this case the center of mass is $\left(0,0,\frac{11}{2}\right)$. The mass is

$$\int_0^{2\pi}\int_0^2\int_{2r^2\cos^2(\theta)+r^2\sin^2(\theta)}^{8-r^2\sin^2(\theta)} r\left(r^2\cos^2(\theta)\right)dzdrd\theta=\frac{32}{3}\pi$$

5. The two cylinders, $x^2+y^2=4$ and $y^2+z^2=4$ intersect in a region, R. Find the mass and center of mass if the density, σ, is given by $\sigma(x,y,z)=z^2$.

   The first cylinder is parallel to the z axis. Let D denote the circle of radius 2 in the xy plane. Then the region just described has (x,y) in the circle of radius 2 and z between $-\sqrt{4-y^2}$ and $\sqrt{4-y^2}$. It follows the total mass is

$$\int_{-2}^2\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}}\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} z^2dzdxdy=\frac{2048}{45}.$$

   By symmetry, the center of mass will be $(0,0,0)$.

6. The two cylinders, $x^2+y^2=4$ and $y^2+z^2=4$ intersect in a region, R. Find the mass and center of mass if the density, σ, is given by $\sigma(x,y,z)=4+z$.

   The total mass is

$$\int_{-2}^2\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}}\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}}(4+z)\,dzdxdy=\frac{512}{3}$$

$$z_c=\frac{\int_{-2}^2\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}}\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} z(4+z)\,dzdxdy}{\left(\frac{512}{3}\right)}=\frac{4}{15}$$

$$x_c=\frac{\int_{-2}^2\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}}\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} x(4+z)\,dzdxdy}{\left(\frac{512}{3}\right)}=0$$

$$y_c=\frac{\int_{-2}^2\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}}\int_{-\sqrt{4-y^2}}^{\sqrt{4-y^2}} y(4+z)\,dzdxdy}{\left(\frac{512}{3}\right)}=0$$

   and so the center of mass is $\left(0,0,\frac{4}{15}\right)$.

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 $\sigma(x,y,z)=4+y+z$.

   This is the inside of an ellipsoid. Denote this by R. Then the total mass is

$$\int_R \sigma dV = \int_R (4+y+z)\,dV$$

   Lets change the variables. Let $x=2u, y=3v, z=w$. When this is done, (u,v,w) will be in the unit ball. The Jacobian of this transformation is 6. Now changing the variables the above integral equals

$$\int_B (4+3v+w)\,6dV$$

   where here B is the unit ball. When integrating over a ball, you ought to suspect that spherical coordinates would be a good idea. Change the variables again in the above integral to spherical coordinates.

$$w=\rho\cos\phi, v=\rho\sin\phi\sin\theta, u=\rho\sin\phi\cos\theta.$$

   Then the above integral in spherical coordinates is

$$\int_0^{\pi}\int_0^{2\pi}\int_0^1 (4+3\rho\sin(\phi)\sin(\theta)+\rho\cos\phi)\,6\rho^2\sin\phi d\rho d\theta d\phi = 32\pi.$$

   To find the center of mass, it would be

$$z_c=\frac{\int_R z\,(4+y+z)\,dV}{32\pi}=\frac{\int_R z^2 dV}{32\pi}.$$

   Now to get this, I have used symmetry of the region. This equals

$$\frac{\int_R w^2 dV}{32\pi}=\frac{\int_0^{\pi}\int_0^{2\pi}\int_0^1 (\rho\cos(\phi))^2\,6\rho^2\sin\phi d\rho d\theta d\phi}{32\pi}=\frac{1}{20}$$

$$\begin{aligned} y_c &= \frac{\int_R y\,(4+y+z)\,dV}{32\pi}=\frac{\int_R y^2 dV}{32\pi} \\ &= \frac{\int_0^{\pi}\int_0^{2\pi}\int_0^1 (3\,(\rho\sin(\phi)\sin(\theta)))^2\,6\rho^2\sin\phi d\rho d\theta d\phi}{32\pi}=\frac{9}{20} \end{aligned}$$

   I think you get the idea. You can now find x_c in the same way.

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 given the density equals 1.

   Using cylindrical coordinates, this is

$$\int_R r^2 dV=\int_0^{2\pi}\int_0^3\int_0^{9-r^2} r^2 r dz dr d\theta=\frac{243}{2}\pi$$

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 given the density equals 1.

   It is like the above except different.

$$\int_0^{2\pi}\int_0^3\int_0^{9-r^2}\left(r^2\sin^2(\theta)+z^2\right) r dz dr d\theta=\frac{1215}{2}\pi$$

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 $\frac{2}{5}R^2M$ where M is the mass.

    The constant density of the ball is $\frac{3}{4}\frac{M}{\pi R^3}$. For simplicity let the line be the z axis. I will also use spherical coordinates since this is a ball. Then the moment of inertia is

$$\int_0^{\pi}\int_0^{2\pi}\int_0^{R}\frac{3}{4}\frac{M}{\pi R^3}\left(\rho\sin\left(\phi\right)\right)^2\rho^2\sin\left(\phi\right)d\rho d\theta d\phi=\frac{2}{5}R^2M.$$

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.

$$\int_0^{\pi}\int_0^{2\pi}\int_0^{R}\rho\left(\rho\sin\left(\phi\right)\right)^2\rho^2\sin\left(\phi\right)d\rho d\theta d\phi=\frac{4}{9}\pi R^6$$

    Also the total mass is

$$M=\int_0^{\pi}\int_0^{2\pi}\int_0^{R}\rho\rho^2\sin\left(\phi\right)d\rho d\theta d\phi=\pi R^4$$

    Therefore, the moment of inertia is

$$\frac{4}{9}MR^2.$$

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 $\frac{1}{2}R^2M$ where M is the mass.

    The density is $\frac{M}{\pi R^2h}$ where h is the height of the cylinder. Using cylindrical coordinates, the moment of inertia is

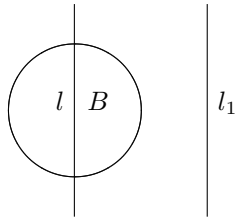
$$\int_0^{2\pi}\int_0^{R}\int_0^{h}\left(\frac{M}{\pi R^2h}\right)r^2rdzdrd\theta=\frac{1}{2}R^2M$$

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?

    The sphere will win. This is because it takes less torque to produce a given angular acceleration in the sphere than in the cylinder because the moment of inertia for the sphere is less than the moment of inertia of the cylinder. Thus a given torque about the axis of roation, which will be identical in both will produce faster rotation in the sphere than in the cylinder. Another way to look at it is that they both have the same total energy when they get to the bottom. This energy comes from two parts, one involving rotation and the other translation of the center of mass. If the center of mass of both were moving at the same speed, this would be a contradiction because the different moments of inertia would then require the kinetic energy of one to be greater than that of the other.

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 $I_{l_1} = I_l + a^2 M$. 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.

    Consider the following picture in which, as suggested, the line, l is the z axis and l_1 goes through $(a, 0)$ in the xy plane and is parallel to l.



    For $\mathbf{x}$ a point in B, let the coordinates of this point be (x, y, z). Then the displacement vector from a point, $(a, 0, z)$ on l_1 to the point, (x, y, z) is $(x - a, -y, 0)$ and so the square of the distance is $x^2 - 2xa + a^2 + y^2$. Therefore, from the definition of moment of inertia, the moment of inertia about l_1 is

$$I_{l_1} \equiv \int_B \delta(x, y, z)\left(x^2 - 2xa + a^2 + y^2\right) dV$$

    Since the line goes through the center of mass, this reduces to

$$\int_B \delta(x, y, z)\left(x^2 + y^2\right) dV + \int_B \delta(x, y, z)\, a^2 dV \equiv I_l + a^2 M$$

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?

    The answer is yes. To see this, consider the parallel axis theorem above.

Part X

Line Integrals

Outcomes

Line Integrals

A. Evaluate the work done by a varying force over a curved path.

B. Evaluate line integrals in general including line integrals with respect to arc length.

C. Evaluate the physical characteristics of a wire such as centroid, mass, and center of mass using line integrals.

Reading: Multivariable Calculus 4.2
Outcome Mapping:

A. 1,9

B. 2,3,4

C. 5,6

Path Independent Line Integrals

A. Recall and apply the Fundamental Theorem for Line Integrals.

B. Determine whether or not a force field is conservative, and if so, find its potential.

C. Evaluate the circulation of a force field or the work done by a force field on a object moving along a given path.

Reading: Multivariable Calculus 4.3
Outcome Mapping:

A. K1,1

B. 3,6

C. 2,4

Recovering a Function from its Gradient

A. Analyze the characteristics of a vector field. Sketch a vector field.

B. Determine whether a vector field is a gradient.

C. Determine whether a differential form is exact.

D. Recover a function from its gradient or differential form, if possible.

Reading: Multivariable Calculus 4.1
Outcome Mapping:

A. 1,7

B. 5

C. 4,6

D. 4,5,7

Line Integrals 14 Nov.

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.

25.0.1 Orientations And Smooth Curves

Recall the notion of a smooth 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 proof that curve length is well defined for a smooth curve contains a result which deserves to be stated as a corollary. It is proves in the Section which starts on Page 295. This is one of those sections you should read only if you are interested.

Corollary 25.0.1 *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 25.0.2 *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 based only on the assumption that the parameterizations of the curve are one to one.

Lemma 25.0.3 *The following hold for* $\sim$.

$$\mathbf{f} \sim \mathbf{f}, \tag{25.1}$$

$$\textit{If } \mathbf{f} \sim \mathbf{g} \textit{ then } \mathbf{g} \sim \mathbf{f}, \tag{25.2}$$

$$\textit{If } \mathbf{f} \sim \mathbf{g} \textit{ and } \mathbf{g} \sim \mathbf{h}, \textit{ then } \mathbf{f} \sim \mathbf{h}. \tag{25.3}$$

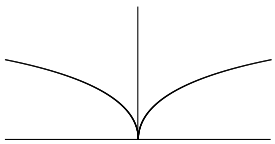
Proof: Formula 25.1 is obvious because $\mathbf{f}^{-1} \circ \mathbf{f}(t) = t$ so it is clearly an increasing function. If $\mathbf{f} \sim \mathbf{g}$ then $\mathbf{f}^{-1} \circ \mathbf{g}$ is increasing. Now $\mathbf{g}^{-1} \circ \mathbf{f}$ must also be increasing because it is the inverse of $\mathbf{f}^{-1} \circ \mathbf{g}$. This verifies 25.2. To see 25.3, $\mathbf{f}^{-1} \circ \mathbf{h} = \left(\mathbf{f}^{-1} \circ \mathbf{g}\right) \circ \left(\mathbf{g}^{-1} \circ \mathbf{h}\right)$ 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 25.0.1, 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.

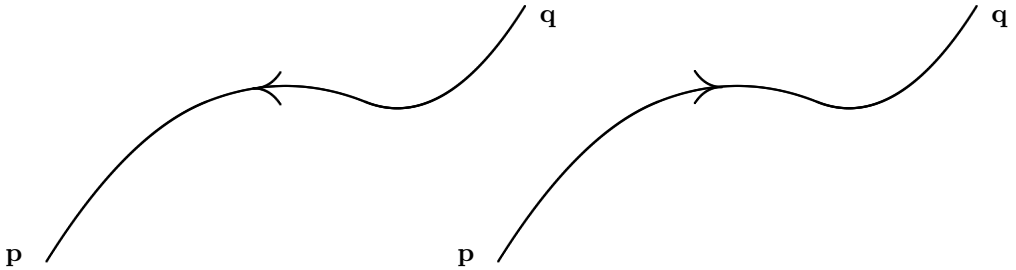
Sometimes people wonder why it is required, in the definition of a smooth curve that $\mathbf{p}'(t) \neq \mathbf{0}$. Imagine t is time and $\mathbf{p}(t)$ gives the location of a point in space. If $\mathbf{p}'(t)$ is allowed to equal zero, the point can stop and change directions abruptly, producing a pointy place in C. Here is an example.

Example 25.0.4 *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! If it were a banister and you were sliding down it, it would be clear at a certain point that the curve is not smooth. I think you may even get the point of this from the picture below.



So what is the thing to remember from all this? First, there are certain conditions which must be satisfied for a curve to be smooth. These are listed in 1 - 5. Next, if you have any curve, there are two directions you can move over this curve, each called an orientation. This is illustrated in the following picture.



Either you move from $\mathbf{p}$ to $\mathbf{q}$ or you move from $\mathbf{q}$ to $\mathbf{p}$.

Definition 25.0.5 *A curve C is piecewise smooth if there exist points on this curve, $\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_n$ such that, denoting $C_{\mathbf{p}_{k-1}\mathbf{p}_k}$ the part of the curve joining $\mathbf{p}_{k-1}$ and $\mathbf{p}_k$, it follows $C_{\mathbf{p}_{k-1}\mathbf{p}_k}$ is a smooth curve and $\cup_{k=1}^{n} C_{\mathbf{p}_{k-1}\mathbf{p}_k} = C$. In other words, it is piecewise smooth if it consists of a finite number of smooth curves linked together.*

Note that Example 25.0.4 is an example of a piecewise smooth curve although it is not smooth.

25.0.2 The Integral Of A Function Defined On A Smooth Curve

Letting $\mathbf{r}(t), t \in [a,b]$ be the position vector of a smooth curve, recall that the total length of this curve is given by

$$l = \int_a^b |\mathbf{r}'(t)|\, dt. \tag{25.4}$$

Remember that if you interpret t as time, $|\mathbf{r}'(t)|$ is the speed and the above integral says that to get the total distance you simply integrate the speed. A small chunk of distance traveled is $dl = |\mathbf{r}'(t)|\, dt$. This says the same thing as

$$\frac{dl}{dt} = |\mathbf{r}'(t)|$$

which was discussed earlier. Of course it follows from 25.4 and the fundamental theorem of calculus. The distance for the parameter between a and t is

$$l(t) = \int_a^t |\mathbf{r}'(s)|\, ds$$

and so by the fundamental theorem of calculus,

$$l'(t) = \frac{dl}{dt} = |\mathbf{r}'(t)|.$$

For this reason, the increment of arc length is $dl = |\mathbf{r}'(t)|\, dt$. Think of it as giving an infinitesimal contribution to the integral. For C a smooth curve with a parameterization, $\mathbf{r} : [a,b] \to C$ and a function, f defined on C, define the symbol,

$$\int_C f dl \equiv \int_a^b f(\mathbf{r}(t)) |\mathbf{r}'(t)|\, dt.$$

Example 25.0.6 *Let C be a smooth curve which has parameterization given by $\mathbf{r}(t) = (\cos 2t, \sin(2t), t)$ for $t \in [0, 2\pi]$. Suppose $f(x,y,z) = x^2 + y$. Find $\int_C f dl$.*

The increment of length is $\sqrt{4\cos^2(2t)+4\sin^2(2t)+1}dt=\sqrt{5}dt$. *Now the desired integral is*

$$\int_0^{2\pi}\left(\cos^2 2t+\sin(2t)\right)\sqrt{5}dt=\sqrt{5}\pi$$

One can define things like density with respect to arc length in the usual way. As just explained, a little chunk of length is $dl=|\mathbf{r}'(t)|\,dt$. The density is a function $\delta(x,y,z)$ which has the property that a little chunk of mass is given by $dm=\delta(\mathbf{r}(t))\,dl$.

Definition 25.0.7 *Let* δ *be the density with respect to arc length. Then the total mass of a smooth curve,* C *having parameterization* $\mathbf{r}:[a,b]\to\mathbb{R}^3$ *is*

$$\int_C \delta(x,y,z)\,dl=\int_a^b \delta(\mathbf{r}(t))\,|\mathbf{r}'(t)|\,dt$$

the center of mass can be given in a similar manner as before. Thus

$$\begin{aligned} x_c &\equiv \frac{\int_C x\delta(x,y,z)\,dl}{\int_C \delta(x,y,z)\,dl}, y_c=\frac{\int_C y\delta(x,y,z)\,dl}{\int_C \delta(x,y,z)\,dl} \\ z_c &= \frac{\int_C z\delta(x,y,z)\,dl}{\int_C \delta(x,y,z)\,dl} \end{aligned}$$

and the only thing you need to do is to evaluate the integrals after changing everything to give a one dimensional integral with respect to the parameter t.

Example 25.0.8 *Let a smooth curve be given by the parameterization,* $\mathbf{r}(t)=(\cos t,\sin t,t)$: $t\in[0,10]$. *This is a helix in case you are interested. Suppose the density is given by* $\delta(x,y,z)=x^2$. *Find the total mass and the center of mass.*

The increment of arc length is $dl=\sqrt{\sin^2(t)+\cos^2(t)+1}dt=\sqrt{2}dt$. Then the total mass is

$$\int_0^{10}\cos^2(t)\sqrt{2}dt=\frac{1}{2}\sqrt{2}\cos(10)\sin(10)+5\sqrt{2}$$

The center of mass is given by

$$x_c=\frac{\int_0^{10}(\cos(t))\cos^2(t)\sqrt{2}dt}{\frac{1}{2}\sqrt{2}\cos(10)\sin(10)+5\sqrt{2}}=\frac{\frac{1}{3}\sqrt{2}\sin 10\cos^2 10+\frac{2}{3}\sqrt{2}\sin 10}{\frac{1}{2}\sqrt{2}\cos 10\sin 10+5\sqrt{2}}$$

$$y_c=\frac{\int_0^{10}(\sin(t))\cos^2(t)\sqrt{2}dt}{\frac{1}{2}\sqrt{2}\cos(10)\sin(10)+5\sqrt{2}}=\frac{-\frac{1}{3}\left(\cos^3 10\right)\sqrt{2}+\frac{1}{3}\sqrt{2}}{\frac{1}{2}\sqrt{2}\cos 10\sin 10+5\sqrt{2}}$$

$$z_c=\frac{\int_0^{10}t\cos^2(t)\sqrt{2}dt}{\frac{1}{2}\sqrt{2}\cos(10)\sin(10)+5\sqrt{2}}=\frac{5\sqrt{2}\cos 10\sin 10+\frac{99}{4}\sqrt{2}+\frac{1}{4}\sqrt{2}\cos^2 10}{\frac{1}{2}\sqrt{2}\cos 10\sin 10+5\sqrt{2}}$$

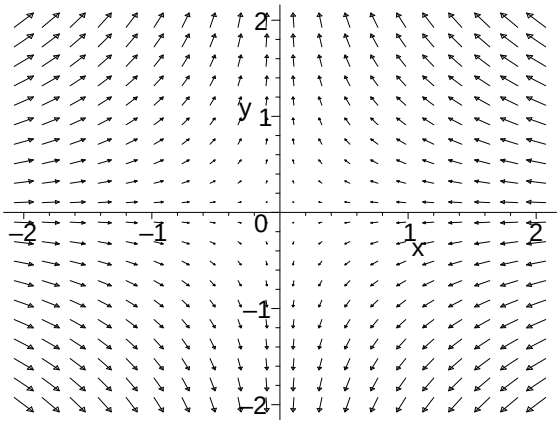
25.0.3 Vector Fields

A **vector field** is nothing but a function which has values which are vectors. For example, consider the force acting on a unit mass by the sun. This determines a force vector which depends on the location of the point. Thus each point in space has associated with it a vector which is the force which the sun exerts on a particle of mass 1 which is placed at that point.

Some people find it useful to try and draw pictures to illustrate a vector valued function or vector field. 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 $\mathbb{R}^2$.

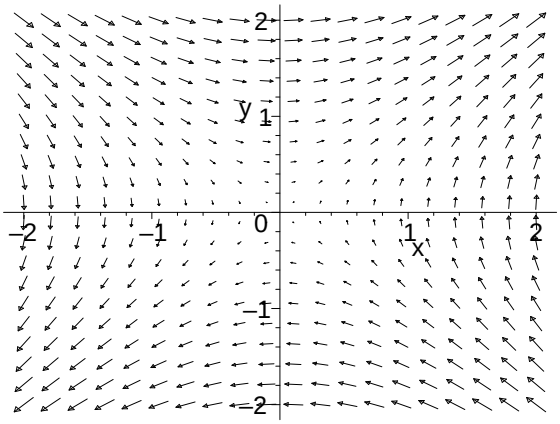
For many points, $(x, y) \in D$, you draw an arrow of the appropriate length and direction with its tail at (x, y). The picture of all these arrows can give you an understanding of what is happening. For example if the vector valued function gives the velocity of a fluid at the point, (x, y) , the picture 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 25.0.9 *Draw a picture of the vector field, $(-x, y)$ which gives the velocity of a fluid flowing in two dimensions.*



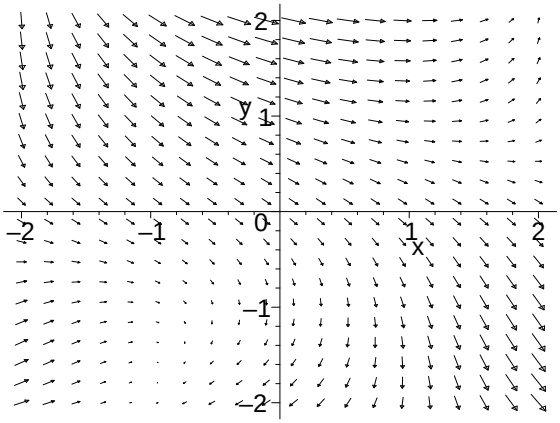
In this example, drawn by Maple, you can see how the arrows indicate the motion of this fluid.

Example 25.0.10 *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 25.0.11 *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.*



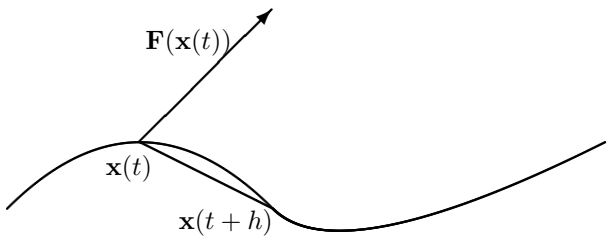
25.0.4 Line Integrals And Work

The interesting concept of line integral has to do with integrals which involve vector fields, not scalar valued functions as above. The most significant application is to work.

First, it is necessary to give some discussion of the concept of orientation. 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 discussed in Definition 25.0.2, 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 25.0.12 *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 over a curve. In order to define what is meant by the work, consider the following picture.



In this picture, the work done by a force, $\mathbf{F}$ on an object which moves from the point $\mathbf{x}(t)$ to the point $\mathbf{x}(t+h)$ along the straight line shown would equal $\mathbf{F}\cdot(\mathbf{x}(t+h)-\mathbf{x}(t))$. It is reasonable to assume this would be a good approximation to the work done in moving along the curve joining $\mathbf{x}(t)$ and $\mathbf{x}(t+h)$ provided h is small enough. Also, provided h is small,

$$\mathbf{x}(t+h)-\mathbf{x}(t)\approx\mathbf{x}'(t)\,h$$

where the wriggly equal sign indicates the two quantities are close. In the notation of Leibniz, one writes dt for h and

$$dW=\mathbf{F}(\mathbf{x}(t))\cdot\mathbf{x}'(t)\,dt$$

Thus the total work along the whole curve should be given by the integral,

$$\int_a^b \mathbf{F}(\mathbf{x}(t))\cdot\mathbf{x}'(t)\,dt$$

This motivates the following definition of work.

Definition 25.0.13 *Let $\mathbf{F}(\mathbf{x})$ be given above. Then the work done by this force field on an object moving over the curve C in the direction determined by the specified orientation is defined as*

$$\int_C \mathbf{F}\cdot d\mathbf{R}\equiv\int_a^b \mathbf{F}(\mathbf{x}(t))\cdot\mathbf{x}'(t)\,dt$$

where the function, $\mathbf{x}$ is one of the allowed parameterizations of C in the given orientation of C. In other words, there is an interval, $[a,b]$ and as t goes from a to b, $\mathbf{x}(t)$ moves in the direction determined from the given orientation of the curve.

Theorem 25.0.14 *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. Letting $s=\phi(t)$ and changing variables, and using the fact 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 negative work.

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 25.0.15 *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

$$\int_C \mathbf{F} \cdot d\mathbf{R} = \int_0^\pi \left(2t\left(\cos(2t)\right), t^2, 1\right) \cdot \left(1, -2\sin(2t), 2\cos(2t)\right) dt$$

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 25.0.16 *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.

$$\begin{aligned}
\int_C \mathbf{F} \cdot d\mathbf{R} &= \int_0^2 \left(t, \sin(t), (t+1)\, t^3\right) \cdot \left(1, \cos(t), 3t^2\right) dt \\
&= \int_0^2 \left(t + \sin(t)\cos(t) + 3(t+1)\, t^5\right) dt \\
&= \frac{1251}{14} - \frac{1}{2}\cos^2(2).
\end{aligned}$$

25.0.5 Another Notation For Line Integrals

Definition 25.0.17 *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.

25.0.6 Exercises With Answers

1. Suppose for $t \in [0, 2\pi]$ the position of an object is given by $\mathbf{r}(t) = 2t\mathbf{i} + \cos(t)\,\mathbf{j} + \sin(t)\,\mathbf{k}$. 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.

You might think of $d\mathbf{R} = \mathbf{r}'(t)\,dt$ to help remember what to do. Then from the definition,

$$\int_C \mathbf{F} \cdot d\mathbf{R} =$$

$$\begin{aligned} & \int_0^{2\pi} \left(2(2t)(\sin t), 4t^2 + 2\sin(t)\cos(t), \sin^2(t)\right) \cdot (2, -\sin(t), \cos(t))\,dt \\ = & \int_0^{2\pi} \left(8t\sin t - \left(2\sin t\cos t + 4t^2\right)\sin t + \sin^2 t\cos t\right)dt = 16\pi^2 - 16\pi \end{aligned}$$

2. Here is a vector field, $\left(y, x^2 + z, 2yz\right)$ 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 $\int_C \mathbf{F} \cdot d\mathbf{R}$.

$$d\mathbf{R} = (-2\sin(2t), 4\cos(2t), 1)\,dt.$$

Then by the definition,

$$\int_C \mathbf{F} \cdot d\mathbf{R} =$$

$$\begin{aligned} & \int_0^{\pi/4} \left(2\sin(2t), \cos^2(2t) + t, 4t\sin(2t)\right) \cdot (-2\sin(2t), 4\cos(2t), 1)\,dt \\ = & \int_0^{\pi/4} \left(-4\sin^2 2t + 4\left(\cos^2 2t + t\right)\cos 2t + 4t\sin 2t\right)dt = \frac{4}{3} \end{aligned}$$

3. 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 $\mathbb{R}^3$,

$$\mathbf{F}(x, y, z) \equiv yz\mathbf{i} + xz\mathbf{j} + xy\mathbf{k}.$$

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 $\mathbf{r}(t) = t\mathbf{i} + t^2\mathbf{j} + t\mathbf{k}$.

You should get the same answer in this case. This is because the vector field happens to be conservative. (More on this later.)

25.1 Path Independent Line Integrals 15 Nov.

Sometimes the line integral giving the work done by a force field depends only on the endpoints of the curve. This is very nice when it happens because it makes the line integral very easy to compute. It also has great physical significance.

Definition 25.1.1 *A vector field,* $\mathbf{F}$ *defined in a three dimensional region is said to be* ***conservative***[1] *if for every piecewise smooth closed curve,* C*, it follows* $\int_C \mathbf{F}\cdot d\mathbf{R} = 0$*.*

Definition 25.1.2 *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 25.1.3 *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*

$$\int_C \mathbf{F}\cdot\, d\mathbf{R} = \phi(\mathbf{y}) - \phi(\mathbf{x}). \tag{25.5}$$

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

$$\phi(\mathbf{x}) \equiv \int_{\mathbf{p}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R}.$$

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

$$\int_{\mathbf{p}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R} = \int_{\mathbf{q}(\mathbf{x}_0,\mathbf{x})} \mathbf{F}\cdot d\mathbf{R}$$

[1] There is no such thing as a liberal vector field.

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 25.5 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 25.5 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.

25.1.1 Finding The Scalar Potential, (Recover The Function From Its Gradient)

Example 25.1.4 *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 25.1.5 *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.

Example 25.1.6 *Let* $\mathbf{F}(x,y,z) = \left(2yx + 1 + y, x^2 + x, 1\right)$. *Find a scalar potential for* $\mathbf{F}$ *if it exists.*

You need $\phi_x = 2yx+1+y$ and so $\phi = yx^2+x+yx+\psi(y,z)$. Then you need $\phi_y = x^2+x+\psi_y = x^2+x$ which shows $\psi_y = 0$ and so $\psi = \psi(z)$. Hence $\phi = yx^2+x+yx+\psi(z)$ Now finally, $\phi_z = \psi'(z) = 1$ and so $\psi(z) = z$ will work. A scalar potential is $\phi(x,y,z) = yx^2+x+yx+z$.

Example 25.1.7 *Let* $\mathbf{F}(x,y,z) = \left(1, 2yz + z\cos y, y^2 + \sin y\right)$. *Find a scalar potential for* $\mathbf{F}$ *if it exists.*

You need $\phi_x = 1$ and so $\phi = x + \psi(y,z)$. Then you need $\phi_y = \psi_y = 2yz + z\cos y$ and so $\psi = y^2z + z\sin y + g(z)$. Hence $\phi = x + y^2z + z\sin y + g(z)$ and you still don't know g. But you must have $\phi_z = y^2 + \sin y + g'(z) = y^2 + \sin y$ and so g is a constant. You can take it to equal zero. Hence $\phi = x + y^2z + z\sin y$ is a scalar potential.

When you are finding one of these scalar potentials, be sure to check your work. Take what you think is the answer and find its gradient. If you get the given vector field, rejoice. If not, it is wrong. Start over again.

Example 25.1.8 *Let the vector field,* $\mathbf{F}$ *be given in Example 25.1.7. Find* $\int_C \mathbf{F}\cdot d\mathbf{R}$ *where* C *is an oriented curve which goes from* $(0,\pi,2)$ *to* $(1,\pi/2,2)$.

This is very easy. It is just $\phi(1,\pi/2,2) - \phi(0,\pi,2)$ where ϕ is the scalar potential in this example. Thus it equals

$$\left(1 + \left(\frac{\pi}{2}\right)^2 2 + 2\right) - \left(\pi^2 2\right) = 3 - \frac{3}{2}\pi^2$$

25.1.2 Terminology

For a vector field, $\mathbf{F}(x,y,z) = F_1(x,y,z)\,\mathbf{i} + F_2(x,y,z)\,\mathbf{j} + F_3(x,y,z)\,\mathbf{k}$, $\mathbf{F}$ is called conservative if it is the gradient of a scalar potential. Thus $\mathbf{F}$ is conservative if there exists a scalar function, ϕ such that $\nabla\phi = \mathbf{F}$. This was discussed above. Another way to say this is that the differential form $F_1dx + F_2dy + F_3dz$ is **exact**. This terminology holds with obvious modifications in any number of dimensions.

Part XI

Green's Theorem, Integrals On Surfaces

Outcomes

Green's Theorem

A. Recall and verify Green's Theorem.

B. Apply Green's Theorem to evaluate line integrals.

C. Apply Green's Theorem to find the area of a region.

Reading: Multivariable Calculus 4.4
Outcome Mapping:

A. L1,L2,3,5

B. 1

C. 2

Surface Integrals

A. Determine the area of a given surface using integration.

B. Evaluate the physical characteristics of a surface such as centroid, mass, and center of mass using surface integrals.

C. Find the flux of a vector field through a surface.

Reading: Multivariable Calculus 4.5
Outcome Mapping:

A. 1

B. 1,2

C. 3

Parametric Surfaces

A. Write a parameterization for a given surface.

B. Identify a surface from its parameterization.

C. Describe a surface from its nets. Sketch a parametric surface.

Reading: Multivariable Calculus 4.6
Outcome Mapping:

A. 3,5,9

B. 1,2

C. 6,7

Integrals over Parametric Surfaces

A. Graphically describe a surface in terms of its parameterization.

B. Determine a (unit) normal vector to a surface from a parameterization of the surface.

C. Determine the plane tangent to a surface at a given point.

D. Evaluate the physical characteristics of parameterized surfaces such as centroid, mass, and center of mass.

E. Find the flux of a flow through a parametric surface.

Reading: Multivariable Calculus 4.7
Outcome Mapping:

A. 1,4

B. 1,4

C. 1,4

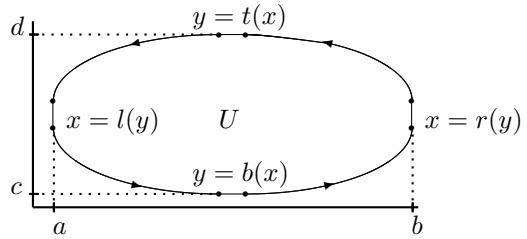
D. 1,4

E. 1,4

Green's Theorem 20 Nov.

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 26.0.9 *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*

$$\int_{\partial U} Pdx + Qdy \equiv$$

$$\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA. \tag{26.1}$$

Proof: First consider the right side of 26.1.

$$\int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA$$

$$\begin{aligned} &= \int_c^d \int_{l(y)}^{r(y)} \frac{\partial Q}{\partial x} dx dy - \int_a^b \int_{b(x)}^{t(x)} \frac{\partial P}{\partial y} dy dx \\ &= \int_c^d (Q(r(y), y) - Q(l(y), y)) dy + \int_a^b (P(x, b(x))) - P(x, t(x)) dx. \end{aligned} \tag{26.2}$$

Now consider the left side of 26.1. Denote by V the vertical parts of ∂U and by H the horizontal parts.

$$\int_{\partial U} \mathbf{F}\cdot d\mathbf{R} =$$

$$\begin{aligned}
&= \int_{\partial U} ((0,Q)+(P,0))\cdot d\mathbf{R} \\
&= \int_c^d (0,Q(r(s),s))\cdot(r'(s),1)\,ds + \int_H (0,Q(r(s),s))\cdot(\pm 1,0)\,ds \\
&\quad - \int_c^d (0,Q(l(s),s))\cdot(l'(s),1)\,ds + \int_a^b (P(s,b(s)),0)\cdot(1,b'(s))\,ds \\
&\quad + \int_V (P(s,b(s)),0)\cdot(0,\pm 1)\,ds - \int_a^b (P(s,t(s)),0)\cdot(1,t'(s))\,ds \\
&= \int_c^d Q(r(s),s)\,ds - \int_c^d Q(l(s),s)\,ds + \int_a^b P(s,b(s))\,ds - \int_a^b P(s,t(s))\,ds
\end{aligned}$$

which coincides with 26.2. This proves the lemma.

Corollary 26.0.10 *Let everything be the same as in Lemma 26.0.9 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.

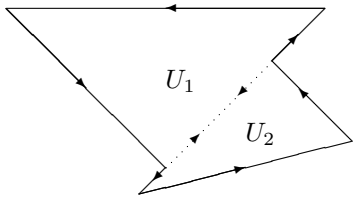
From this corollary, it follows 26.1 is valid for any triangle for example.

Now suppose 26.1 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 26.1 must hold for $U \equiv \cup_{i=1}^m U_i$, the union of these sets. This is because

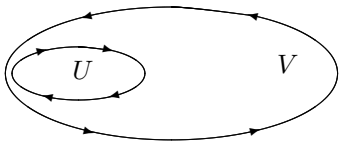
$$\int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA =$$

$$\begin{aligned}
&= \sum_{k=1}^m \int_{U_k} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \\
&= \sum_{k=1}^m \int_{\partial U_k} \mathbf{F}\cdot d\mathbf{R} = \int_{\partial U} \mathbf{F}\cdot d\mathbf{R}
\end{aligned}$$

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 26.1 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 26.1. Roughly speaking, you can drill holes in a region for which 26.1 holds and get another region for which this continues to hold provided 26.1 holds for the holes. To see why this is so, consider the following picture which typifies the situation just described.



Then

$$\int_{\partial V} \mathbf{F}\cdot d\mathbf{R} = \int_V \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA$$

$$\begin{aligned} &= \int_U \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA + \int_{V\setminus U} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \\ &= \int_{\partial U} \mathbf{F}\cdot d\mathbf{R} + \int_{V\setminus U} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \end{aligned}$$

and so

$$\int_{V\setminus U} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = \int_{\partial V} \mathbf{F}\cdot d\mathbf{R} - \int_{\partial U} \mathbf{F}\cdot d\mathbf{R}$$

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 26.1 is valid quite generally. This verifies the following theorem.

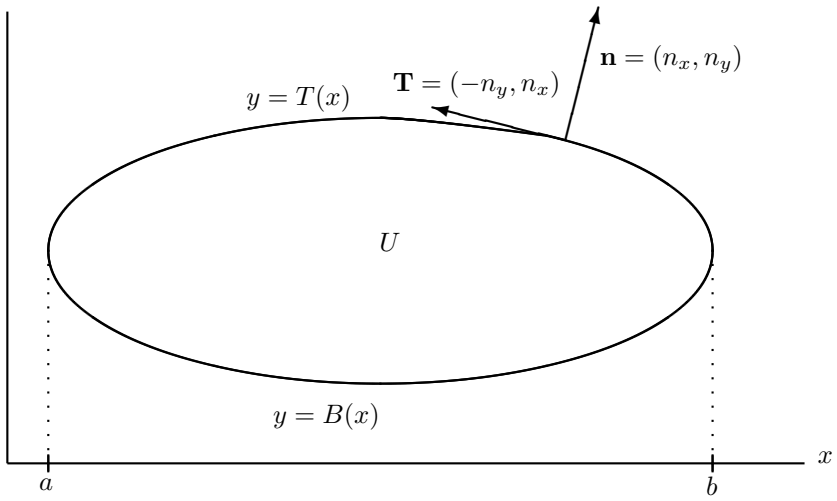
Theorem 26.0.11 *(Green's Theorem) Let U be an open set in the plane and let ∂U be piecewise smooth and let $\mathbf{F}(x,y) = (P(x,y), Q(x,y))$ be a C^1 vector field defined near U. Then it is often[1] the case that*

$$\int_{\partial U} \mathbf{F}\cdot d\mathbf{R} = \int_U \left(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y)\right) dA.$$

26.1 An Alternative Explanation Of Green's Theorem

Consider the following picture.

[1]For a general version see the advanced calculus book by Apostol. The general versions involve the concept of a rectifiable Jordan curve. You need to be able to take the area integral and to take the line integral around the boundary.



In this picture $\mathbf{n}$ is the unit outer normal to U and the vector, $\mathbf{T}$ shown in the picture is the unit tangent vector in the direction of counter clockwise motion around U. To see that it really does point in the correct direction, take the cross product, $(n_x, n_y, 0) \times (-n_y, n_x, 0)$. This equals $\mathbf{k}$. Applying the right hand rule, this shows the vector, $(-n_y, n_x)$ really does point in the direction indicated by the picture.

Next I will establish Gauss' theorem for regions like U. The boundary of U is denoted by ∂U.

Lemma 26.1.1 *(Gauss)Let $(H(x,y), K(x,y))$ be a C^1 vector field defined near U. Then for $\mathbf{n}$ the unit outer normal,*

$$\int_U (H_x + K_y)\, dA = \int_{\partial U} (H, K) \cdot \mathbf{n} dl$$

Proof: A parameterization for the top is $(x, T(x))$ and a parameterization for the bottom is $(x, B(x))$ where in both cases, $x \in [a, b]$. Thus $dl = \sqrt{1 + T'(x)^2}dx$ on the top and $dl = \sqrt{1 + B'(x)^2}dx$ on the bottom. Thus also, on the top, you can find the exterior normal by considering it as the level surface, $y - T(x) = 0$. Thus a unit normal to this surface is

$$\mathbf{n} = \frac{(-T'(x), 1)}{\sqrt{1 + T'(x)^2}} = (n_x, n_y)$$

and you see that since the y component is positive, it is the outer normal, pointing away from U. Similarly, the unit outer normal on the bottom is given by

$$\mathbf{n} = \frac{(B'(x), -1)}{\sqrt{1 + B'(x)^2}} = (n_x, n_y)$$

First consider

$$\begin{aligned}\int_U K_y dA &= \int_a^b \int_{B(x)}^{T(x)} K_y dy dx = \int_a^b \left(K\left(x,T\left(x\right)\right) - K\left(x,B\left(x\right)\right)\right) dx \\ &= \int_a^b K\left(x,T\left(x\right)\right) dx - \int_a^b K\left(x,B\left(x\right)\right) dx \\ &= \int_a^b K\left(x,T\left(x\right)\right) \overbrace{\frac{1}{\sqrt{1+T'\left(x\right)^2}}}^{n_y} \overbrace{\sqrt{1+T'\left(x\right)^2}dx}^{dl} \\ &\quad + \int_a^b K\left(x,B\left(x\right)\right) \overbrace{\left(\frac{-1}{\sqrt{1+B'\left(x\right)^2}}\right)}^{n_y} \overbrace{\sqrt{1+B'\left(x\right)^2}dx}^{dl} \\ &= \int_{\partial U} K n_y dl\end{aligned}$$

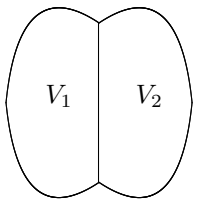
Similar reasoning shows that

$$\int_U H_x dA = \int_{\partial U} H n_x dl$$

Therefore, this proves the Lemma because from the above,

$$\begin{aligned}\int_U \left(H_x + K_y\right) dA &= \int_U H_x dA + \int_U K_y dA = \int_{\partial U} H n_x dl + \int_{\partial U} K n_y dl \\ &= \int_{\partial U} \left(H,K\right) \cdot \mathbf{n} dl\end{aligned}$$

Now this theorem holds for many regions much more general than the one shown. In fact, it holds for any region which is made up by pasting together regions like the above. This is because the area integrals add and the integrals on the parts of the boundary which are shared by two pieces cancel due to the fact they have the exterior normals which are in opposite directions. 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.



This theorem is also called the divergence theorem. This is because the divergence of the vector field, $\left(H\left(x,y\right),K\left(x,y\right)\right)$ is defined as $H_x\left(x,y\right) + K_y\left(x,y\right)$.

Theorem 26.1.2 *(Green's Theorem) Let U be any bounded open set in $\mathbb{R}^2$ for which the above Gauss' theorem holds and let*

$$\mathbf{F}\left(x,y\right) \equiv \left(P\left(x,y\right),Q\left(x,y\right)\right)$$

be a C^1 vector field defined near U. Then

$$\int_U \left(Q_x - P_y\right) dA = \int_{\partial U} \mathbf{F} \cdot d\mathbf{R}$$

where the line integral is oriented in the counter clockwise direction.

Proof: If $\mathbf{r}(t)$ is a parameterization of ∂U near a point on ∂U, then recall the unit tangent vector, $\mathbf{T}$ as shown in the above picture satisfies $|\mathbf{r}'(t)|\,\mathbf{T} = \mathbf{r}'(t)$. Thus $\mathbf{F}\cdot d\mathbf{R}$ is of the form $\mathbf{F}(\mathbf{r}(t))\cdot\mathbf{r}'(t)\,dt = \mathbf{F}\cdot\mathbf{T}\,|\mathbf{r}'(t)|\,dt = \mathbf{F}\cdot\mathbf{T}dl$ because $dl = |\mathbf{r}'(t)|\,dt$. Then using Lemma 26.1.1 and letting $(H,K) = (Q,-P)$

$$\begin{aligned}\int_U (Q_x - P_y)\,dA &= \int_U (H_x + K_y)\,dA \\ &= \int_{\partial U} (H,K)\cdot(n_x,n_y)\,dl \\ &= \int_{\partial U} (Q,-P)\cdot(n_x,n_y)\,dl \\ &= \int_{\partial U} (P,Q)\cdot(-n_y,n_x)\,dl \\ &= \int_{\partial U} \mathbf{F}\cdot\mathbf{T}dl = \int_{\partial U}\mathbf{F}\cdot d\mathbf{R}.\end{aligned}$$

This proves Green's theorem.

26.2 Area And Green's Theorem

Proposition 26.2.1 *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 26.2.2 *Use Proposition 26.2.1 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 Proposition 26.2.1,

$$\begin{aligned}\text{Area equals} &= \frac{1}{2}\int_0^{2\pi} (-b\sin t, a\cos t)\cdot(-a\sin t, b\cos t)\,dt \\ &= \frac{1}{2}\int_0^{2\pi} (ab)\,dt = \pi ab.\end{aligned}$$

Example 26.2.3 *Find $\int_{\partial U}\mathbf{F}\cdot d\mathbf{R}$ where U is the set, $\{(x,y) : x^2+3y^2 \le 9\}$ and $\mathbf{F}(x,y) = (y,-x)$.*

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

$$\int_U ((-1)-1)\,dA = -2\int_U dA.$$

Now U is an ellipse having area equal to $3\sqrt{3}$ and so the answer is $-6\sqrt{3}$.

Example 26.2.4 *Find* $\int_{\partial U} \mathbf{F}\cdot d\mathbf{R}$ *where* U *is the set,* $\{(x,y) : 2 \leq x \leq 4, 0 \leq y \leq 3\}$ *and* $\mathbf{F}(x,y) = \left(x\sin y, y^3 \cos x\right)$.

From Green's theorem this line integral equals

$$\begin{aligned} & \int_2^4 \int_0^3 \left(-y^3 \sin x - x\cos y\right) dydx \\ = \quad & \frac{81}{4}\cos 4 - 6\sin 3 - \frac{81}{4}\cos 2. \end{aligned}$$

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 26.2.5 *Find* $\int_{\partial U} \mathbf{F}\cdot d\mathbf{R}$ *where* U *is the set,* $\{(x,y) : 2 \leq x \leq 4, x \leq y \leq 3\}$ *and* $\mathbf{F}(x,y) = (x\sin y, y\sin x)$.

From Green's theorem this line integral equals

$$\begin{aligned} & \int_2^4 \int_x^3 (y\cos x - x\cos y)\, dydx \\ = \quad & -\frac{3}{2}\sin 4 - 6\sin 3 - 8\cos 4 - \frac{9}{2}\sin 2 + 4\cos 2. \end{aligned}$$

The Integral On Two Dimensional Surfaces In $\mathbb{R}^3$ 27-28 Nov.

27.1 Parametrically Defined Surfaces

Definition 27.1.1 *Let S be a subset of $\mathbb{R}^3$. Then S is a **smooth surface** if there exists an open set, $U \subseteq \mathbb{R}^2$ and a C^1 function, $\mathbf{r}$ defined on U such that $\mathbf{r}(U) = S, \mathbf{r}$ is one to one, and for all $(u,v) \in U$,*

$$\mathbf{r}_u \times \mathbf{r}_v \neq \mathbf{0}. \tag{27.1}$$

This last condition ensures that there is always a well defined normal on S. This function, $\mathbf{r}$ is called a parameterization of the surface. It is just like a parameterization of a curve but here there are two parameters, u, v.

One way to think of this is that there is a piece of rubber occupying U in the plane and then it is taken and stretched in three dimensions. This gives S. Here is an interesting example which is already familiar.

Example 27.1.2 *Let $(\phi,\theta) \in (0,\pi) \times (0,2\pi)$ and for such (ϕ,θ),*

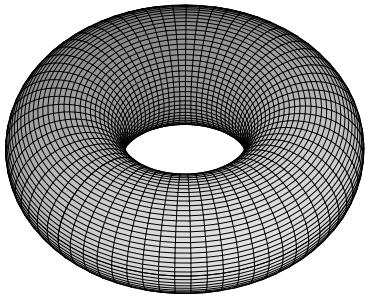
$$\mathbf{r}(\phi,\theta) \equiv \begin{pmatrix} 2\sin(\phi)\cos(\theta) \\ 2\sin(\phi)\sin(\theta) \\ 2\cos(\phi) \end{pmatrix}$$

This gives most of a sphere of radiius 2 for S. You should check condition 27.1. You will find that $|\mathbf{r}_\phi \times \mathbf{r}_\theta| = 4\sin(\phi) \neq 0$.

Example 27.1.3 *Let $R > r$. Consider*

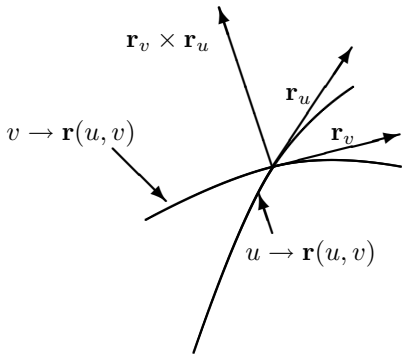
$$\mathbf{r}(u,v) = \begin{pmatrix} (R + r\cos(u))\cos(v) \\ (R + r\cos(u))\sin(v) \\ r\sin(u) \end{pmatrix}$$

where $(u,v) \in (0,2\pi) \times (0,2\pi)$. This surface is most of the surface of a torus (donut) with small radius equal to r. It is obtained by revolving the circle of radius r centered at $(R,0,0)$ about the z axis. Here is a picture.



In the above I have assumed U is open. However, this is generalized later. It is amazing how far this can be generalized in applications to integration.

In general, if you fix u and consider $\mathbf{r}(u,v)$ as a function of v, this yields a smooth curve which lies in the surface, S. By fixing different values of u you obtain many different curves in S. Similarly you can fix v and consider $\mathbf{r}(u,v)$ as a function of u. The curves which result in this way are called a net for the surface. This is the way a computer graphs a surface. It graphs lots of different curves as just described. You can see this in the above picture of a torus. The curves which make up the shape shown correspond to one of the variables in the parameterization being fixed. Now at a point, $\mathbf{r}(u,v)$ of S, there are two vectors tangent to S at this point, $\mathbf{r}_u(u,v)$ and $\mathbf{r}_v(u,v)$. These two vectors determine a plane which can be considered tangent to the surface at the point, $\mathbf{r}(u,v)$. You can find an equation for this plane if you can obtain a normal vector. However, this is easy. You simply take $\mathbf{r}_v \times \mathbf{r}_u$ to obtain a vector which is normal to the tangent plane. Here is a picture. The two curves correspond to $u \to \mathbf{r}(u,v)$ and $v \to \mathbf{r}(u,v)$. The vectors $\mathbf{r}_u$ and $\mathbf{r}_v$ are tangent to the respective curves as shown. Then taking the cross product gives a normal to the surface at that point.



Example 27.1.4 *Let S be the surface defined in Example 27.1.3 in which $R = 2$ and $r = 1$. Find a tangent plane to the point*

$$\mathbf{r}\left(\frac{\pi}{4}, \frac{\pi}{4}\right).$$

This point is $\left(\sqrt{2}+\frac{1}{2}, \sqrt{2}+\frac{1}{2}, \frac{\sqrt{2}}{2}\right)$. I only need to find a normal vector in order to find

the plane.

$$\begin{aligned}\mathbf{r}_u \times \mathbf{r}_v &= \begin{pmatrix} -\sin u \cos v \\ -\sin u \sin v \\ \cos u \end{pmatrix} \times \begin{pmatrix} -(2+\cos u)\sin v \\ (2+\cos u)\cos v \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} -(\cos u)(2+\cos u)\cos v \\ -(\cos u)(2+\cos u)\sin v \\ -\left(\sin u \cos^2 v\right)(2+\cos u) - \left(\sin u \sin^2 v\right)(2+\cos u) \end{pmatrix}\end{aligned}$$

Now plugging in the desired values of u and v, a normal vector is

$$\begin{pmatrix} -1-\frac{1}{4}\sqrt{2} \\ -1-\frac{1}{4}\sqrt{2} \\ -\sqrt{2}-\frac{1}{2} \end{pmatrix}.$$

I don't like the minus signs so the normal vector I will use is

$$\left(1+\frac{1}{4}\sqrt{2}, 1+\frac{1}{4}\sqrt{2}, \sqrt{2}+\frac{1}{2}\right)^T.$$

Now it follows the equation of the tangent plane is

$$\left(1+\frac{1}{4}\sqrt{2}\right)\left(x-\sqrt{2}-\frac{1}{2}\right)+\left(1+\frac{1}{4}\sqrt{2}\right)\left(y-\sqrt{2}-\frac{1}{2}\right)+\left(\sqrt{2}+\frac{1}{2}\right)\left(z-\frac{1}{2}\sqrt{2}\right)=0.$$

You could simplify this if you wanted.

$$\left(1+\frac{1}{4}\sqrt{2}\right)x+\left(1+\frac{1}{4}\sqrt{2}\right)y+\left(\sqrt{2}+\frac{1}{2}\right)z=\frac{5}{2}\sqrt{2}+3.$$

27.2 The Two Dimensional Area In $\mathbb{R}^3$

Consider the boundary of some three dimensional region such that a function, 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. It all depends on the following fundamental definition which is just a review of the fact presented earlier that the area of a parallelogram determined by two vectors in $\mathbb{R}^3$ is the norm of the cross product of the two vectors.

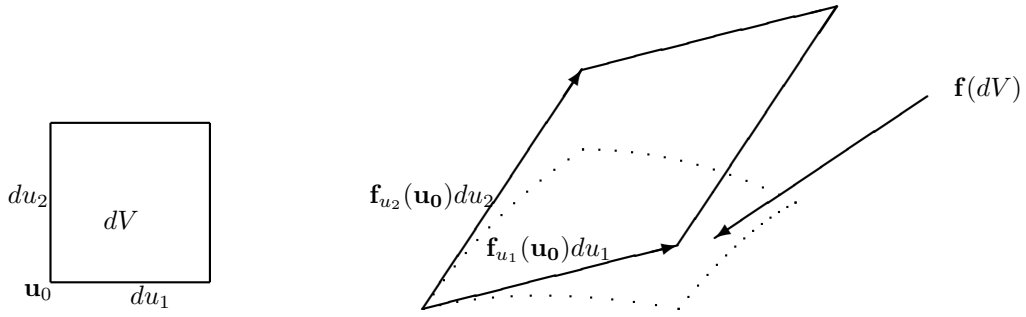
Definition 27.2.1 *Let $\mathbf{u}_1, \mathbf{u}_2$ be vectors in $\mathbb{R}^3$. The 2 dimensional parallelogram determined by these vectors will be denoted by $P(\mathbf{u}_1, \mathbf{u}_2)$ and it is defined as*

$$P(\mathbf{u}_1, \mathbf{u}_2) \equiv \left\{ \sum_{j=1}^{2} s_j \mathbf{u}_j : s_j \in [0,1] \right\}.$$

Then the area of this parallelogram is

$$\text{area } P(\mathbf{u}_1, \mathbf{u}_2) \equiv |\mathbf{u}_1 \times \mathbf{u}_2|.$$

Suppose then that $\mathbf{x} = \mathbf{f}(\mathbf{u})$ where $\mathbf{u} \in U$, a subset of $\mathbb{R}^2$ and $\mathbf{x}$ is a point in V, a subset of 3 dimensional space. Thus, letting the Cartesian coordinates of $\mathbf{x}$ be given by $\mathbf{x} = (x_1, x_2, x_3)^T$, each x_i being a function of $\mathbf{u}$, an infinitesimal rectangle located at $\mathbf{u}_0$ corresponds to an infinitesimal parallelogram located at $\mathbf{f}(\mathbf{u}_0)$ which is determined by the 2 vectors $\left\{ \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_i} du_i \right\}_{i=1}^{2}$, each of which is tangent to the surface defined by $\mathbf{x} = \mathbf{f}(\mathbf{u})$. (No sum on the repeated index.)



From Definition 27.2.1, the volume of this infinitesimal parallelepiped located at $\mathbf{f}(\mathbf{u}_0)$ is given by

$$\left| \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_1} du_1 \times \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_2} du_2 \right| = \left| \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_1} \times \frac{\partial \mathbf{f}(\mathbf{u}_0)}{\partial u_2} \right| du_1 du_2 \tag{27.2}$$

$$= |\mathbf{f}_{u_1} \times \mathbf{f}_{u_2}| \, du_1 du_2 \tag{27.3}$$

It might help to think of a lizard. The infinitesimal parallelogram is 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 If the scales are small enough, their sum would serve as a good approximation to the area of the lizard.

[1].This motivates the following fundamental procedure which I hope is extremely familiar from the earlier material.

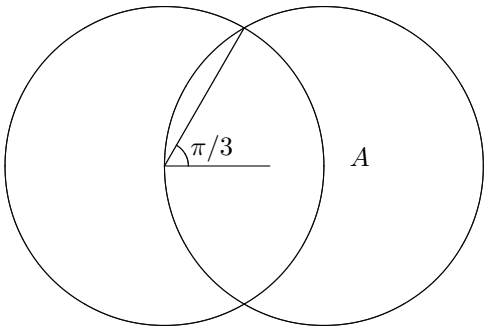
Procedure 27.2.2 *Suppose U is a subset of $\mathbb{R}^2$ and suppose $\mathbf{f} : U \to \mathbf{f}(U) \subseteq \mathbb{R}^3$ is a one to one and C^1 function. Then if $h : \mathbf{f}(U) \to \mathbb{R}$, define the 2 dimensional surface integral, $\int_{\mathbf{f}(U)} h(\mathbf{x})\, dA$ according to the following formula.*

$$\int_{\mathbf{f}(U)} h(\mathbf{x})\, dA \equiv \int_U h(\mathbf{f}(\mathbf{u}))\, |\mathbf{f}_{u_1}(\mathbf{u}) \times \mathbf{f}_{u_2}(\mathbf{u})|\, du_1 du_2.$$

Definition 27.2.3 *It is customary to write $|\mathbf{f}_{u_1}(\mathbf{u}) \times \mathbf{f}_{u_2}(\mathbf{u})| = \frac{\partial(x_1,x_2,x_3)}{\partial(u_1,u_2)}$ because this new notation generalizes to far more general situations for which the cross product is not defined. For example, one can consider three dimensional surfaces in $\mathbb{R}^8$.*

First here is a simple example where the surface is actually in the plane.

Example 27.2.4 *Find the area of the region labelled 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,

$$\int_U 1\, dV = \int_{-\pi/3}^{\pi/3} \int_1^{2\cos\theta} \frac{\partial(x_1,x_2,x_3)}{\partial(\theta,r)}\, dr\, d\theta.$$

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

$$\mathbf{f}(\theta, r) = (r\cos\theta, r\sin\theta)^T.$$

Here $x_3 = 0$ and so

$$\frac{\partial(x_1,x_2,x_3)}{\partial(\theta,r)} = \left\| \begin{array}{ccc} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial x_1}{\partial\theta} & \frac{\partial x_2}{\partial\theta} & \frac{\partial x_3}{\partial\theta} \\ \frac{\partial x_1}{\partial r} & \frac{\partial x_2}{\partial r} & \frac{\partial x_3}{\partial r} \end{array} \right\| = \left\| \begin{array}{ccc} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -r\sin\theta & r\cos\theta & 0 \\ \cos\theta & \sin\theta & 0 \end{array} \right\| = r$$

[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.

Therefore, the area element is $r\,dr\,d\theta$. It follows the desired area is

$$\int_{-\pi/3}^{\pi/3}\int_{1}^{2\cos\theta} r\,dr\,d\theta = \frac{1}{2}\sqrt{3}+\frac{1}{3}\pi.$$

Notice how the area element reduced to the area element for polar coordinates.

Example 27.2.5 *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 $\frac{\partial(x_1,x_2,x_3)}{\partial(x,y)} = |\mathbf{f}_x \times \mathbf{f}_y|$?

$$\mathbf{f}_x = \begin{pmatrix} 1 \\ 0 \\ 2x \end{pmatrix}, \mathbf{f}_y = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

and so

$$|\mathbf{f}_x \times \mathbf{f}_y| = \left| \begin{pmatrix} 1 \\ 0 \\ 2x \end{pmatrix} \times \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right| = \sqrt{1+4x^2}$$

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

$$\int_U dA = \int_0^1\int_0^1 \sqrt{1+4x^2}\,dx\,dy = \frac{1}{2}\sqrt{5} - \frac{1}{4}\ln\left(-2+\sqrt{5}\right)$$

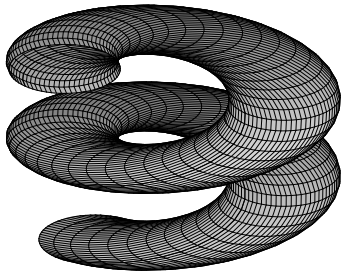
which can be obtained by using the trig. substitution, $2x = \tan\theta$ on the inside integral.

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. The problem is, 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. This is a good example which illustrates the gulf between theory and practice.

Example 27.2.6 *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,

$$\mathbf{f}_t = \begin{pmatrix} -2\sin t \\ 2\cos t \\ 1 \end{pmatrix}, \mathbf{f}_s = \begin{pmatrix} -\sin s \\ \cos s \\ 0 \end{pmatrix}$$

and

$$\mathbf{f}_t \times \mathbf{f}_s = \begin{pmatrix} -\cos s \\ -\sin s \\ -2\sin t\cos s + 2\cos t \sin s \end{pmatrix}$$

and so

$$\frac{\partial(x_1, x_2, x_3)}{\partial(t, s)} = |\mathbf{f}_t \times \mathbf{f}_s| = \sqrt{5 - 4\sin^2 t \sin^2 s - 8\sin t \sin s \cos t \cos s - 4\cos^2 t \cos^2 s}.$$

Therefore, the desired integral giving the area is

$$\int_0^{2\pi} \int_0^{12} \sqrt{5 - 4\sin^2 t \sin^2 s - 8\sin t \sin s \cos t \cos s - 4\cos^2 t \cos^2 s}\, dt\, ds.$$

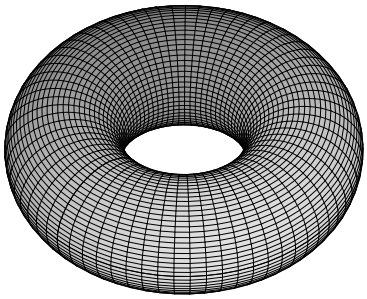
If you really needed to find the number this equals, how would you go about finding it? This is an interesting question and there is no single right answer. You should think about this. It is important in some physical applications to get the number even when you can't find the antiderivative. Here is an example for which you will be able to find the integrals.

Example 27.2.7 *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.

[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.



To find its area,

$$\mathbf{f}_t = \begin{pmatrix} -2\sin t - \sin t \cos s \\ -2\cos t - \cos t \cos s \\ 0 \end{pmatrix}, \mathbf{f}_s = \begin{pmatrix} -\cos t \sin s \\ \sin t \sin s \\ \cos s \end{pmatrix}$$

and so $|\mathbf{f}_t \times \mathbf{f}_s| = (\cos s + 2)$ so the area element is $(\cos s + 2)\, ds\, dt$ and the area is

$$\int_0^{2\pi} \int_0^{2\pi} (\cos s + 2)\, ds\, dt = 8\pi^2$$

Example 27.2.8 *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 $h(x, y, z) = x^2$.

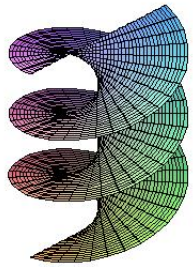
Everything is the same as the preceding example except this time it is an integral of a function. The area element is $(\cos s + 2)\, ds\, dt$ and so the integral called for is

$$\int_{\mathbf{f}(U)} h\, dA = \int_0^{2\pi} \int_0^{2\pi} \left(\overbrace{2\cos t + \cos t \cos s}^{x \text{ on the surface}} \right)^2 (\cos s + 2)\, ds\, dt = 22\pi^2$$

Example 27.2.9 *Let $U = [-5, 5] \times [0, 3\pi]$ and for $(s, t) \in U$, let*

$$\mathbf{f}(s, t) = (3s\cos t, 3s\sin t, 4t).$$

Find a formula for the area of $\mathbf{f}(U)$ in terms of integrals. This is called a helicoid. Here is a picture of it.



The area element is

$$|(3\cos(t), 3\sin(t), 0) \times (-3\sin(t), 3s\cos(t), 4)|\, dsdt = \sqrt{144 + \left(9\left(\cos^2 t\right)s + 9\sin^2 t\right)^2}\, dsdt$$

Therefore, the area is given by the double integral,

$$\int_{-5}^{5}\int_{0}^{3\pi} \sqrt{144 + \left(9\left(\cos^2 t\right)s + 9\sin^2 t\right)^2}\, dsdt$$

You can define the center of mass and density of a surface in exactly the same way as was done before.

Definition 27.2.10 *Let S be a surface with area (volume) element dS. The **density with respect to area** is a function which integrated gives the mass. Thus if $\delta(\mathbf{x})$ is the density, the mass of S is*

$$\int_S \delta(x,y,z)\, dS.$$

*The **center of mass** is defined exactly as before.*

$$\begin{aligned} x_c &\equiv \frac{\int_S \delta(x,y,z)\, x dS}{\int_S \delta(x,y,z)\, dS},\ y_c \equiv \frac{\int_S \delta(x,y,z)\, y dS}{\int_S \delta(x,y,z)\, dS} \\ z_c &\equiv \frac{\int_S \delta(x,y,z)\, z dS}{\int_S \delta(x,y,z)\, dS}. \end{aligned}$$

There is no new thing here. You simply are integrating over a surface rather than a volume. Of course you must put the variables, x, y, z as well as dS in terms of the parameters used to compute the integrals.

Example 27.2.11 *The surface is given by $(x,y,z) = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)$ where $(\phi,\theta) \in (0,\pi)\times(0,2\pi)$. Thus the surface is the surface of a sphere of radius 1. Review spherical coordinates at this time if this is not obvious to you. Suppose the density of a point on this surface corresponding to (ϕ,θ) is $\sin^2\phi$. That is, the density is equal to the square of the distance to the z axis. Find the total mass and the center of mass of this surface.*

First find the area element. This equals

$$\begin{aligned} dS &= |(\cos\phi\cos\theta, \cos\phi\sin\theta, -\sin\phi) \times (-\sin\phi\sin\theta, \sin\phi\cos\theta, 0)|\, d\theta d\phi \\ &= \sqrt{\left(\sin^2\phi\cos\theta\right)^2 + \left(\sin^2\phi\sin\theta\right)^2 + \left(\cos\phi\cos^2\theta\sin\phi + \cos\phi\sin^2\theta\sin\phi\right)^2}\, d\theta d\phi \\ &= \sin(\phi)\, d\theta d\phi \end{aligned}$$

To find the total mass you must integrate this area element times the density. Thus the total mass is

$$\int_0^{\pi}\int_0^{2\pi}\left(\sin^2(\phi)\right)\sin(\phi)\,d\theta d\phi = \frac{8}{3}\pi$$

Next you want to find the center of mass. By symmetry, it should be at the origin. As before, the center of mass does not need to be in the surface just as it did not need to be in the three dimensional shape. Now on the surface, $x = \sin\phi\cos\theta, y = \sin\phi\sin\theta$, and $z = \cos\phi$. Consider the formulas for this.

$$\begin{aligned} x_c &= \frac{\int_0^{\pi}\int_0^{2\pi}(\sin(\phi)\cos(\theta))\left(\sin^2(\phi)\right)\sin(\phi)\,d\theta d\phi}{\left(\frac{8}{3}\pi\right)} = 0,\\ y_c &= \frac{\int_0^{\pi}\int_0^{2\pi}(\sin(\phi)\sin(\theta))\left(\sin^2(\phi)\right)\sin(\phi)\,d\theta d\phi}{\left(\frac{8}{3}\pi\right)} = 0,\\ z_c &= \frac{\int_0^{\pi}\int_0^{2\pi}(\cos(\phi))\left(\sin^2(\phi)\right)\sin(\phi)\,d\theta d\phi}{\left(\frac{8}{3}\pi\right)} = 0. \end{aligned}$$

Example 27.2.12 *In the above example suppose $\delta(x,y,z) = z+1$. What is the mass and center of mass?*

The total mass is

$$\int_0^{\pi}\int_0^{2\pi}(1+\cos(\phi))\sin(\phi)\,d\theta d\phi = 4\pi.$$

Next, the center of mass is given by

$$\begin{aligned} x_c &= \frac{\int_0^{\pi}\int_0^{2\pi}(\sin(\phi)\cos(\theta))(1+\cos(\phi))\sin(\phi)\,d\theta d\phi}{4\pi} = 0,\\ y_c &= \frac{\int_0^{\pi}\int_0^{2\pi}(\sin(\phi)\sin(\theta))(1+\cos(\phi))\sin(\phi)\,d\theta d\phi}{4\pi} = 0,\\ z_c &= \frac{\int_0^{\pi}\int_0^{2\pi}(\cos(\phi))(1+\cos(\phi))\sin(\phi)\,d\theta d\phi}{4\pi} = \frac{1}{3} \end{aligned}$$

27.2.1 Surfaces Of The Form $z = f(x,y)$

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 $\mathbf{f}(x,y) = (x,y,f(x,y))^T$ such that $(x,y) \in U$. Then

$$\mathbf{f}_x = \begin{pmatrix} 1\\ 0\\ f_x \end{pmatrix}, \mathbf{f}_y = \begin{pmatrix} 0\\ 1\\ f_y \end{pmatrix}$$

and

$$|\mathbf{f}_x \times \mathbf{f}_y| = \sqrt{1+f_y^2+f_x^2}$$

so the area element is

$$\sqrt{1+f_y^2+f_x^2}\,dx\,dy.$$

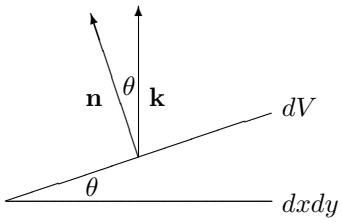
When the surface of interest comes in this simple form, people generally use this area element directly rather than worrying about a parameterization and taking cross products.

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, $dA = \sqrt{1+f_x^2+f_y^2}\,dx\,dy$ 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 27.2.13 *Let $z = \sqrt{x^2+y^2}$ where $(x,y) \in U$ for $U = \{(x,y) : x^2+y^2 \le 4\}$ Find*

$$\int_S h\,dS$$

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,

$$\begin{aligned}\int_S h\,dS &= \int_U \left(x+\sqrt{x^2+y^2}\right)\sqrt{2}\,dA \\ &= \sqrt{2}\int_0^{2\pi}\int_0^2 (r\cos\theta + r)\,r\,dr\,d\theta \\ &= \frac{16}{3}\sqrt{2}\pi.\end{aligned}$$

One other issue is worth mentioning. Suppose $\mathbf{f}_i : U_i \to \mathbb{R}^3$ where U_i are sets in $\mathbb{R}^2$ and suppose $\mathbf{f}_1(U_1)$ intersects $\mathbf{f}_2(U_2)$ along C where $C = \mathbf{h}(V)$ for $V \subseteq \mathbb{R}^1$. Then define

integrals and areas over $\mathbf{f}_1(U_1) \cup \mathbf{f}_2(U_2)$ as follows.

$$\int_{\mathbf{f}_1(U_1)\cup\mathbf{f}_2(U_2)} g\,dA \equiv \int_{\mathbf{f}_1(U_1)} g\,dA + \int_{\mathbf{f}_2(U_2)} g\,dA.$$

Admittedly, the set C gets added in twice but this doesn't matter because its 2 dimensional volume equals zero and therefore, the integrals over this set will also be zero.

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. There are many very interesting issues here which can keep you fascinated for years.

27.3 Flux

Imagine a surface, S which is fixed in space and let $\mathbf{v}$ be a vector field representing the velocity of a fluid flowing through this surface. It is reasonable to ask how fast the fluid crosses the surface in terms of units of mass per units of time. This is expressed in terms of the surface integral,

$$\int_S \rho\mathbf{v}\cdot\mathbf{n}dA$$

where ρ is the density and $\mathbf{n}$ is the normal vector to the surface in the direction in which the crossing is taking place. The vector field, $\rho\mathbf{v}$ is called the flux. To get the rate of transfer of mass across the surface, you take the dot product of the flux with the appropriate unit normal vector and integrate this over the surface. People also speak of heat flux. In general, when they speak of flux, they mean the thing you dot with a unit normal vector and integrate to find the rate at which something crosses a surface. A little later, this idea will be explored much more when the divergence theorem is established. It is a very important idea. You should think about the physical reasons the flux of such a fluid is given as above. Why do you use the unit normal for example? Why not some normal which has different length? Why do you need to take the dot product with the normal? In general situations, people assume formulas about the flux in terms of other quantities such as temperature or concentration. I will mention some later at a convenient place.

27.3.1 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 $4y + 4z = 2x + 4$.

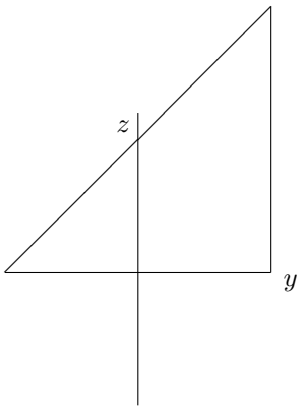
   Answer:

   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 $\int_D \sqrt{1+z_u^2+z_v^2}\,dA = \int_{-ha}^{ha}\int_{-\sqrt{(9h^2-u^2)}}^{\sqrt{(9h^2-u^2)}} \frac{1}{3}\sqrt{10}\,dv\,du = 3\pi h^2\sqrt{10}$

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

$$|\mathbf{f}_u\times\mathbf{f}_v| = \left|\begin{pmatrix}0\\0\\1\end{pmatrix}\times\begin{pmatrix}-2\sin v\\2\cos v\\0\end{pmatrix}\right| = 2.$$

   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 $\int_P \left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| dr\,d\theta = \int_0^{2\pi}\int_0^{2+\cos\theta} r\,dr\,d\theta$ 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 $\int_R \sqrt{1+4x^2h^2+4y^2h^2}\,dA$. Changing to polar coordinates, this becomes
    $\int_0^{2\pi}\int_0^1 \left(\sqrt{1+4h^2r^2}\right) r\,dr\,d\theta = \frac{\pi}{6h^2}\left(\left(1+4h^2\right)^{3/2}-1\right)$.

11. Evaluate $\int_S (1+x)\,dA$ 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 $\int_S (1+x)\,dA$ 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 $\int_S (1+x)\, dA$ 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:

$$\begin{aligned} |\mathbf{f}_\theta \times \mathbf{f}_\alpha| &= \left| \begin{pmatrix} -\sin(\theta)(2+\cos\alpha) \\ -\cos(\theta)(2+\cos\alpha) \\ 0 \end{pmatrix} \times \begin{pmatrix} -\cos\theta\sin\alpha \\ \sin\theta\sin\alpha \\ \cos\alpha \end{pmatrix} \right| \\ &= \left(4+4\cos\alpha+\cos^2\alpha\right)^{1/2} \end{aligned}$$

and so the area element is

$$\left(4+4\cos\alpha+\cos^2\alpha\right)^{1/2}\, d\theta\, d\alpha.$$

Therefore, the area is

$$\int_0^{2\pi}\int_0^{2\pi} \left(4+4\cos\alpha+\cos^2\alpha\right)^{1/2}\, d\theta\, d\alpha = \int_0^{2\pi}\int_0^{2\pi} (2+\cos\alpha)\, d\theta\, d\alpha = 8\pi^2.$$

15. For $(\theta,\alpha)\in[0,2\pi]\times[0,2\pi]$, let

$$\mathbf{f}(\theta,\alpha) \equiv (\cos\theta(4+2\cos\alpha), -\sin\theta(4+2\cos\alpha), 2\sin\alpha)^T.$$

Also let $h(\mathbf{x}) = \cos\alpha$ where α is such that

$$\mathbf{x} = (\cos\theta(4+2\cos\alpha), -\sin\theta(4+2\cos\alpha), 2\sin\alpha)^T.$$

Find $\int_{\mathbf{f}([0,2\pi]\times[0,2\pi])} h\, dA$.

Answer:

$$\begin{aligned} |\mathbf{f}_\theta \times \mathbf{f}_\alpha| &= \left| \begin{pmatrix} -\sin(\theta)(4+2\cos\alpha) \\ -\cos(\theta)(4+2\cos\alpha) \\ 0 \end{pmatrix} \times \begin{pmatrix} -2\cos\theta\sin\alpha \\ 2\sin\theta\sin\alpha \\ 2\cos\alpha \end{pmatrix} \right| \\ &= \left(64+64\cos\alpha+16\cos^2\alpha\right)^{1/2} \end{aligned}$$

and so the area element is

$$\left(64+64\cos\alpha+16\cos^2\alpha\right)^{1/2}\, d\theta\, d\alpha.$$

Therefore, the desired integral is

$$\int_0^{2\pi}\int_0^{2\pi}(\cos\alpha)\left(64+64\cos\alpha+16\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha$$
$$=\int_0^{2\pi}\int_0^{2\pi}(\cos\alpha)\left(8+4\cos\alpha\right)\,d\theta\,d\alpha=8\pi^2$$

16. For $(\theta,\alpha)\in[0,2\pi]\times[0,2\pi]$, let

$$\mathbf{f}(\theta,\alpha)\equiv(\cos\theta(3+\cos\alpha),-\sin\theta(3+\cos\alpha),\sin\alpha)^T.$$

Also let $h(\mathbf{x})=\cos^2\theta$ where θ is such that

$$\mathbf{x}=(\cos\theta(3+\cos\alpha),-\sin\theta(3+\cos\alpha),\sin\alpha)^T.$$

Find $\int_{\mathbf{f}([0,2\pi]\times[0,2\pi])}h\,dV$.

Answer:

The area element is

$$\left(9+6\cos\alpha+\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha.$$

Therefore, the desired integral is

$$\int_0^{2\pi}\int_0^{2\pi}\left(\cos^2\theta\right)\left(9+6\cos\alpha+\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha$$
$$=\int_0^{2\pi}\int_0^{2\pi}\left(\cos^2\theta\right)(3+\cos\alpha)\,d\theta\,d\alpha=6\pi^2$$

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, the area element is

$$\left(68+64\cos\alpha+12\cos^2\alpha\right)^{1/2}\,d\theta\,d\alpha$$

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:

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 $\int_0^{2\pi}\int_0^{2\pi}(2+\cos\alpha)\,d\theta\,d\alpha=8\pi^2$.

Part XII

Divergence Theorem

Outcomes

Flux Density and Divergence

A. Explain what is meant by the flux density and divergence of a vector field.

B. Evaluate the divergence of a vector field.

C. Evaluate the Laplacian of a function.

D. Derive formulas involving divergence, gradient and Laplacian.

Reading: Multivariable Calculus 5.1
Outcome Mapping:

A. M1

B. 1

C. 2,3

D. 4

The Divergence Theorem

A. Recall and verify the Divergence Theorem.

B. Apply the Divergence Theorem to evaluate the flux through a surface.

Reading: Multivariable Calculus 5.2
Outcome Mapping:

A. N1,N2,4,8

B. 1,2

The Divergence Theorem 29-30 Nov.

28.1 Divergence Of A Vector Field

Here the important concepts of divergence is defined.

Definition 28.1.1 *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*. 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 a coordinate free definition of div is presented. For now, everything is defined in terms of a given Cartesian coordinate system. The divergence has profound physical significance and this will be discussed later. For now it is important to understand how to find it. 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.

Example 28.1.2 *Let* $\mathbf{f}(\mathbf{x}) = xy\mathbf{i} + (z-y)\,\mathbf{j} + (\sin(x)+z)\,\mathbf{k}$. *Find* $\operatorname{div}\mathbf{f}$

First the divergence of $\mathbf{f}$ is

$$\frac{\partial(xy)}{\partial x} + \frac{\partial(z-y)}{\partial y} + \frac{\partial(\sin(x)+z)}{\partial z} = y + (-1) + 1 = y.$$

28.2 The Divergence Theorem

Why does anyone care about the divergence of a vector field? The answer is contained in this section. In short, it is because of the divergence theorem which relates the flux over the boundary to a volume integral of the divergence. It is also called Gauss's theorem.

Definition 28.2.1 *A subset,* V *of* $\mathbb{R}^3$ *is called cylindrical in the* x *direction if it is of the form*

$$V = \{(x,y,z) : \phi(y,z) \le x \le \psi(y,z) \text{ for } (y,z) \in D\}$$

where D *is a subset of the* yz *plane.* V *is cylindrical in the* z *direction if*

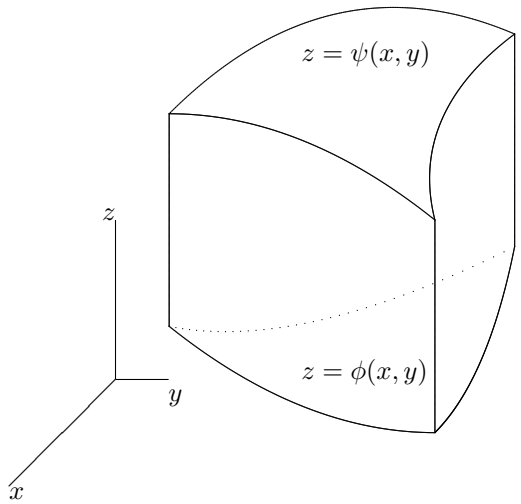
$$V = \{(x,y,z) : \phi(x,y) \le z \le \psi(x,y) \text{ for } (x,y) \in D\}$$

where D *is a subset of the* xy *plane, and* V *is cylindrical in the* y *direction if*

$$V = \{(x,y,z) : \phi(x,z) \le y \le \psi(x,z) \text{ for } (x,z) \in D\}$$

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* $\phi(x,y) \le z \le \psi(x,y)$. *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 28.2.2 *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*

$$\int_V \frac{\partial F}{\partial z}(x,y,z)\, dV = \int_{\partial V} F n_z\, dA.$$

Proof: From the fundamental theorem of calculus,

$$\begin{aligned} \int_V \frac{\partial F}{\partial z}(x,y,z)\, dV &= \int_D \int_{\phi(x,y)}^{\psi(x,y)} \frac{\partial F}{\partial z}(x,y,z)\, dz\, dx\, dy \\ &= \int_D \left[F(x,y,\psi(x,y)) - F(x,y,\phi(x,y))\right] dx\, dy \end{aligned} \tag{28.1}$$

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 28.1 is of the form,

$$\int_D F(x,y,\psi(x,y)) \overbrace{\frac{1}{\sqrt{\psi_x^2 + \psi_y^2 + 1}}}^{n_z} \overbrace{\sqrt{\psi_x^2 + \psi_y^2 + 1}\, dx\, dy}^{dA} +$$

$$\int_D F(x,y,\phi(x,y)) \overbrace{\left(\frac{-1}{\sqrt{\phi_x^2 + \phi_y^2 + 1}}\right)}^{n_z} \overbrace{\sqrt{\phi_x^2 + \phi_y^2 + 1}\, dx\, dy}^{dA}$$

$$+\int_{\text{Lateral surface}} Fn_z\,dA,$$

the last term equaling zero because on the lateral surface, $n_z = 0$. Therefore, this reduces to $\int_{\partial V} Fn_z\,dA$ as claimed.

The following corollary is entirely similar to the above.

Corollary 28.2.3 *If V is cylindrical in the y direction, then*

$$\int_V \frac{\partial F}{\partial y}\,dV = \int_{\partial V} Fn_y\,dA$$

and if V is cylindrical in the x direction, then

$$\int_V \frac{\partial F}{\partial x}\,dV = \int_{\partial V} Fn_x\,dA$$

With this corollary, here is a proof of the divergence theorem.

Theorem 28.2.4 *Let V be cylindrical in each of the coordinate directions and let $\mathbf{F}$ be a C^1 vector field defined on V. Then*

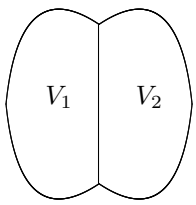
$$\int_V \nabla\cdot\mathbf{F}\,dV = \int_{\partial V} \mathbf{F}\cdot\mathbf{n}\,dA.$$

Proof: From the above lemma and corollary,

$$\begin{aligned}\int_V \nabla\cdot\mathbf{F}\,dV &= \int_V \frac{\partial F_1}{\partial x}+\frac{\partial F_2}{\partial y}+\frac{\partial F_3}{\partial y}\,dV\\ &= \int_{\partial V}(F_1n_x+F_2n_y+F_3n_z)\,dA\\ &= \int_{\partial V}\mathbf{F}\cdot\mathbf{n}\,dA.\end{aligned}$$

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 it can be used with confidence.

Example 28.2.5 *Let $V = [0,1] \times [0,1] \times [0,1]$. That is, V is the cube in the first octant having the lower left corner at $(0,0,0)$ and the sides of length 1. Let $\mathbf{F}(x,y,z) = x\mathbf{i}+y\mathbf{j}+z\mathbf{k}$. Find the flux integral in which $\mathbf{n}$ is the unit exterior normal.*

$$\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS$$

You can certainly inflict much suffering on yourself by breaking the surface up into 6 pieces corresponding to the 6 sides of the cube, finding a parameterization for each face and adding up the appropriate flux integrals. For example, $\mathbf{n} = \mathbf{k}$ on the top face and $\mathbf{n} = -\mathbf{k}$ on the bottom face. On the top face, a parameterization is $(x,y,1) : (x,y) \in [0,1] \times [0,1]$. The area element is just $dxdy$. It isn't really all that hard to do it this way but it is much easier to use the divergence theorem. The above integral equals

$$\int_V \operatorname{div}(\mathbf{F})\, dV = \int_V 3dV = 3.$$

Example 28.2.6 *This time, let V be the unit ball, $\left\{(x,y,z) : x^2+y^2+z^2 \leq 1\right\}$ and let $\mathbf{F}(x,y,z) = x^2\mathbf{i} + y\mathbf{j} + (z-1)\mathbf{k}$. Find*

$$\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS.$$

As in the above you could do this by brute force. A parameterization of the ∂V is obtained as

$$x = \sin\phi\cos\theta,\ y = \sin\phi\sin\theta,\ z = \cos\phi$$

where $(\phi,\theta) \in (0,\pi) \times (0,2\pi]$. Now this does not include all the ball but it includes all but the point at the top and at the bottom. As far as the flux integral is concerned these points contribute nothing to the integral so you can neglect them. Then you can grind away and get the flux integral which is desired. However, it is so much easier to use the divergence theorem! Using spherical coordinates,

$$\begin{aligned}\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS &= \int_V \operatorname{div}(\mathbf{F})\, dV = \int_V (2x+1+1)\, dV \\ &= \int_0^{\pi}\int_0^{2\pi}\int_0^1 (2+2\rho\sin(\phi)\cos\theta)\,\rho^2\sin(\phi)\, d\rho d\theta d\phi = \frac{8}{3}\pi\end{aligned}$$

Example 28.2.7 *Suppose V is an open set in $\mathbb{R}^3$ for which the divergence theorem holds. Let $\mathbf{F}(x,y,z) = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$. Then show*

$$\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS = 3 \times \textit{volume(V)}.$$

This follows from the divergenc theorem.

$$\int_{\partial V} \mathbf{F} \cdot \mathbf{n} dS = \int_V \operatorname{div}(\mathbf{F})\, dV = 3\int_V dV = 3 \times \text{volume}(V).$$

The message of the divergence theorem is the relation between the volume integral and an area integral. This is the exciting thing about this marvelous theorem. It is not its utility as a method for evaluations of boring problems. This will be shown in the examples of its use which follow.

28.2.1 Coordinate Free Concept Of Divergence, Flux Density

The divergence theorem also makes possible a coordinate free definition of the divergence.

Theorem 28.2.8 *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

$$\operatorname{div}\mathbf{F}(\mathbf{x}) = \lim_{\delta\to 0+} \frac{1}{v(B(\mathbf{x},\delta))} \int_{\partial B(\mathbf{x},\delta)} \mathbf{F}\cdot\mathbf{n}\, dA. \tag{28.2}$$

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 28.2.

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.

You also see the physical significance of the divergence from this. It measures the tendency of the vector field to "diverge" from a point.

28.3 The Weak Maximum Principle*

There is 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. It is a very important result which ties in many earlier topics. Don't read it if you are not interested.

Theorem 28.3.1 *Let U be a bounded open set in $\mathbb{R}^n$ and suppose $u \in C^2(U) \cap C(\overline{U})$ such that $\nabla^2 u \geq 0$ in U. Then letting $\partial U = \overline{U} \backslash U$, it follows that* $\max\{u(\mathbf{x}) : \mathbf{x} \in \overline{U}\} = \max\{u(\mathbf{x}) : x \in \partial U\}$.

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.

28.4 Some Applications Of The Divergence Theorem*

There are numerous applications of the divergence theorem. Some are listed here. You might want to read this if you are interested in applications. However, it won't be needed for tests.

28.4.1 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, 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

$$\nabla p = \rho \mathbf{g}. \tag{28.3}$$

It turns out that the pressure in a lake at depth z is equal to $62.5z$. This is easy to see from 28.3. 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 28.3 is more interesting because it does not require ρ to be constant.

28.4.2 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 28.3 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 28.3 because in this equation $\mathbf{g} = -g\mathbf{k}$.

28.4.3 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 $\int_V \rho(\mathbf{x},t)\, c(\mathbf{x},t)\, u(\mathbf{x},t)\ dV$. 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.

$$\frac{\partial}{\partial t}\left(\rho(\mathbf{x},t)\, c(\mathbf{x},t)\, u(\mathbf{x},t)\right) = \nabla\cdot\left(k(u,\mathbf{x},t)\,\nabla u(\mathbf{x},t)\right) + f(\mathbf{x},u,t). \tag{28.4}$$

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, $\frac{d}{dt}\int_V \rho(\mathbf{x},t)\, c(\mathbf{x},t)\, u(\mathbf{x},t)\ dV =$

$$\int_V \frac{\partial\left(\rho(\mathbf{x},t)\, c(\mathbf{x},t)\, u(\mathbf{x},t)\right)}{\partial t}\, dV = \overbrace{\int_{\partial V} -\mathbf{J}\cdot\mathbf{n}\, dA}^{\text{rate at which heat enters}} + \int_V f(\mathbf{x},u,t)\ dV$$

$$= \int_{\partial V} k\nabla(u)\cdot\mathbf{n}\, dA + \int_V f(\mathbf{x},u,t)\ dV = \int_V \left(\nabla\cdot\left(k\nabla(u)\right) + f\right)\ dV.$$

Since this holds for every sample volume, V it must be the case that the above reaction diffusion equation, 28.4 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,

$$\frac{\partial}{\partial t}u(\mathbf{x},t) = \frac{1}{\rho c}\nabla\cdot\left(k(\mathbf{x})\,\nabla u(\mathbf{x},t)\right) + f(\mathbf{x},t). \tag{28.5}$$

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.

28.4.4 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.

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = f(\mathbf{y},t) \tag{28.6}$$

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

$$\begin{aligned}\int_V \frac{\partial \rho(\mathbf{y},t)}{\partial t}\, dV &= \overbrace{-\int_{\partial V} \rho(\mathbf{y},t)\,\mathbf{v}(\mathbf{y},t)\cdot \mathbf{n}\, dA}^{\text{rate at which mass enters}} + \int_V f(\mathbf{y},t)\, dV \\ &= -\int_V \left(\nabla \cdot (\rho(\mathbf{y},t)\,\mathbf{v}(\mathbf{y},t)) + f(\mathbf{y},t)\right) dV.\end{aligned}$$

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.

28.4.5 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. One of the most elegant, possibly the most elegant is the book by Gurtin [13]. To read this book, you will need to know what the derivative of a function of many variables is. This is also the case in this section.

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[2]. Thus

$$\mathbf{y}(0,\mathbf{x}) = \mathbf{x}. \tag{28.7}$$

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 28.4.1 *$\rho_0(\mathbf{x}) = \rho(t,\mathbf{y}(t,\mathbf{x}))\det(F)$ 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 28.7. 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

[2]Apparently, the terminology is all mixed up in so far as the names Euler and Lagrange are concerned.

derivation, $dV(\mathbf{y})$ will indicate the integration is taking place with respect to the variable, $\mathbf{y}$. By Lemma 28.4.1 and the change of variables formula for multiple integrals

$$\begin{aligned}
\frac{d}{dt}\left(\int_{B_t}\rho(t,\mathbf{y})\,\mathbf{v}(t,\mathbf{y})\;dV(\mathbf{y})\right) &= \frac{d}{dt}\left(\int_{B_0}\rho(t,\mathbf{y}(t,\mathbf{x}))\,\mathbf{v}(t,\mathbf{y}(t,\mathbf{x}))\det(F)\;dV(\mathbf{x})\right)\\
&= \frac{d}{dt}\left(\int_{B_0}\rho_0(\mathbf{x})\,\mathbf{v}(t,\mathbf{y}(t,\mathbf{x}))\;dV(\mathbf{x})\right)\\
&= \int_{B_0}\rho_0(\mathbf{x})\left[\frac{\partial\mathbf{v}}{\partial t}+\frac{\partial\mathbf{v}}{\partial y_i}\frac{\partial y_i}{\partial t}\right]dV(\mathbf{x})\\
&= \int_{B_t}\rho(t,\mathbf{y})\det(F)\left[\frac{\partial\mathbf{v}}{\partial t}+\frac{\partial\mathbf{v}}{\partial y_i}\frac{\partial y_i}{\partial t}\right]\frac{1}{\det(F)}\,dV(\mathbf{y})\\
&= \int_{B_t}\rho(t,\mathbf{y})\left[\frac{\partial\mathbf{v}}{\partial t}+\frac{\partial\mathbf{v}}{\partial y_i}\frac{\partial y_i}{\partial t}\right]dV(\mathbf{y}).
\end{aligned}$$

This uses the repeated index summation convention. 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, $\mathbf{b}$ 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. It could also be centrifugal force which might result if the body were undergoing some sort of rigid motion. The traction force is of the form

$$\int_{\partial B_t}\mathbf{s}(t,\mathbf{y},\mathbf{n})\;dA$$

where $\mathbf{n}$ 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 $\mathbf{s}$ 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, $T_{ij}=T_{ji}$. (This comes from conservation of angular momentum.) 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 $\mathbf{v}$, written as $\dot{\mathbf{v}}$. 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, $\mathbf{x}$. The spacial coordinates are fine if you are looking for example at the flow of a fluid through some fixed region in space. However, if you are interested in the deformation of a material, then you might want to consider things like traction boundary conditions. To make sense of these, you should be dealing with the reference or material coordinates. 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

$$\int_{B_0} \rho_0(\mathbf{x})\,\mathbf{v}_t(t,\mathbf{x})\; dV = \int_{B_0} \mathbf{b}_0(t,\mathbf{x})\; dV + \int_{\partial B_t} T_{ij}n_j dA, \tag{28.8}$$

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 28.8 is the sum of integrals over small pieces of the form

$$\int_{P_t} T_{ij}n_j dA \tag{28.9}$$

where P_t is parametrized by $\mathbf{y}(s,t)$, $(s,t) \in D$. Thus the integral in 28.9 is of the form

$$\int_D T_{ij}(\mathbf{y}(s,t))\left(\mathbf{y}_s(s,t) \times \mathbf{y}_t(s,t)\right)_j\; ds\, dt.$$

Using the repeated index summation convention, and 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

$$\int_D T_{ij}(\mathbf{x}(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \left(\frac{\partial \mathbf{y}}{\partial x_\alpha} \times \frac{\partial \mathbf{y}}{\partial x_\beta} \right)_j ds\, dt \tag{28.10}$$

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 28.10 equals

$$\int_D T_{ij}(\mathbf{x}(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{jab} \frac{\partial y_a}{\partial x_\alpha} \frac{\partial y_b}{\partial x_\beta} ds\, dt$$

$$= \int_D T_{ij}(\mathbf{x}(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{cab} \delta_{jc} \frac{\partial y_a}{\partial x_\alpha} \frac{\partial y_b}{\partial x_\beta} ds\, dt$$

$$= \int_D T_{ij}(\mathbf{x}(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \varepsilon_{cab} \overbrace{\frac{\partial y_c}{\partial x_p} \frac{\partial x_p}{\partial y_j}}^{=\delta_{jc}} \frac{\partial y_a}{\partial x_\alpha} \frac{\partial y_b}{\partial x_\beta} ds\, dt$$

$$= \int_D T_{ij}(\mathbf{x}(s,t)) \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \frac{\partial x_p}{\partial y_j} \overbrace{\varepsilon_{cab} \frac{\partial y_c}{\partial x_p} \frac{\partial y_a}{\partial x_\alpha} \frac{\partial y_b}{\partial x_\beta}}^{=\varepsilon_{p\alpha\beta} \det(F)} ds\, dt$$

$$= \int_D (\det F)\, T_{ij}(\mathbf{x}(s,t)) \varepsilon_{p\alpha\beta} \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} \frac{\partial x_p}{\partial y_j} ds\, dt.$$

Now $\frac{\partial x_p}{\partial y_j} = F^{-1}_{pj}$ and also

$$\varepsilon_{p\alpha\beta} \frac{\partial x_\alpha}{\partial s} \frac{\partial x_\beta}{\partial t} = (\mathbf{x}_s \times \mathbf{x}_t)_p$$

so the result just obtained is of the form

$$\int_D (\det F)\, F^{-1}_{pj} T_{ij}(\mathbf{x}(s,t)) (\mathbf{x}_s \times \mathbf{x}_t)_p \, ds\, dt =$$

$$\int_D (\det F)\, T_{ij}(\mathbf{x}(s,t)) \left(F^{-T}\right)_{jp} (\mathbf{x}_s \times \mathbf{x}_t)_p \, ds\, dt.$$

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

$$\int_{P_0} \det(F) \left(TF^{-T}\right)_{ip} N_p dA$$

Summing these up over the pieces of ∂B_t and ∂B_0 yields the last integral in 28.8 equals

$$\int_{\partial B_0} \det(F) \left(TF^{-T}\right)_{ip} N_p dA$$

and so the balance of momentum in terms of the material coordinates becomes

$$\int_{B_0} \rho_0(\mathbf{x})\, \mathbf{v}_t(t, \mathbf{x})\, dV = \int_{B_0} \mathbf{b}_0(t, \mathbf{x})\, dV + \int_{\partial B_0} \det(F) \left(TF^{-T}\right)_{ip} N_p dA$$

The matrix, $\det(F)\left(TF^{-T}\right)_{ip}$ is called the first Piola Kirchhoff stress, S. An application of the divergence theorem yields

$$\int_{B_0}\rho_0(\mathbf{x})\,\mathbf{v}_t(t,\mathbf{x})\,dV=\int_{B_0}\mathbf{b}_0(t,\mathbf{x})\,dV+\int_{B_0}\frac{\partial\left(\det(F)\left(TF^{-T}\right)_{ip}\right)}{\partial x_p}\,dV.$$

Since B_0 is arbitrary, a balance law for momentum in terms of the material coordinates is obtained

$$\begin{aligned}\rho_0(\mathbf{x})\,\mathbf{v}_t(t,\mathbf{x}) &= \mathbf{b}_0(t,\mathbf{x})+\frac{\partial\left(\det(F)\left(TF^{-T}\right)_{ip}\right)}{\partial x_p}\\ &= \mathbf{b}_0(t,\mathbf{x})+\operatorname{div}\left(\det(F)\left(TF^{-T}\right)\right)\\ &= \mathbf{b}_0(t,\mathbf{x})+\operatorname{div}S. \qquad (28.11)\end{aligned}$$

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.

28.4.6 Bernoulli's Principle*

Consider a possibly moving fluid with constant density, ρ and let P denote the pressure in this fluid. If B is a part of this fluid the force exerted on B by the rest of the fluid is $\int_{\partial B}-P\mathbf{n}dA$ where $\mathbf{n}$ is the outer normal from B. Assume this is the only force which matters so for example there is no viscosity in the fluid. Thus the Cauchy stress in rectangular coordinates should be

$$T=\begin{pmatrix}-P & 0 & 0\\ 0 & -P & 0\\ 0 & 0 & -P\end{pmatrix}.$$

Then

$$\operatorname{div}T=-\nabla P.$$

Also suppose the only body force is from gravity, a force of the form

$$-\rho g\mathbf{k}$$

and so from the balance of momentum

$$\rho\dot{\mathbf{v}}=-\rho g\mathbf{k}-\nabla P(\mathbf{x}). \qquad (28.12)$$

Now in all this the coordinates are the spacial coordinates and it is assumed they are rectangular. Thus

$$\mathbf{x}=(x,y,z)^T$$

and $\mathbf{v}$ is the velocity while $\dot{\mathbf{v}}$ is the total derivative of $\mathbf{v} = (v_1, v_2, v_3)^T$ given by $\mathbf{v}_t + v_i \mathbf{v}_{,i}$. Take the dot product of both sides of 28.12 with $\mathbf{v}$. This yields

$$(\rho/2) \frac{d}{dt} |\mathbf{v}|^2 = -\rho g \frac{dz}{dt} - \frac{d}{dt} P(\mathbf{x}).$$

Therefore,

$$\frac{d}{dt} \left(\frac{\rho |\mathbf{v}|^2}{2} + \rho g z + P(\mathbf{x}) \right) = 0$$

and so there is a constant, C' such that

$$\frac{\rho |\mathbf{v}|^2}{2} + \rho g z + P(\mathbf{x}) = C'$$

For convenience define γ to be the weight density of this fluid. Thus $\gamma = \rho g$. Divide by γ. Then

$$\frac{|\mathbf{v}|^2}{2g} + z + \frac{P(\mathbf{x})}{\gamma} = C.$$

this is Bernoulli's[3] principle. Note how if you keep the height the same, then if you raise $|\mathbf{v}|$, it follows the pressure drops.

This is often used to explain the lift of an airplane wing. The top surface is curved which forces the air to go faster over the top of the wing causing a drop in pressure which creates lift. It is also used to explain the concept of a venturi tube in which the air loses pressure due to being pinched which causes it to flow faster. In many of these applications, the assumptions used in which ρ is constant and there is no other contribution to the traction force on ∂B than pressure so in particular, there is no viscosity, are not correct. However, it is hoped that the effects of these deviations from the ideal situation above are small enough that the conclusions are still roughly true. You can see how using balance of momentum can be used to consider more difficult situations. For example, you might have a body force which is more involved than gravity.

28.4.7 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 28.11 you can verify the following holds.

$$\rho_0(\mathbf{x}) \mathbf{u}_{tt}(t, \mathbf{x}) = \mathbf{b}_0(t, \mathbf{x}) + k \Delta \mathbf{u}(t, \mathbf{x}) \tag{28.13}$$

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.

[3]There were many Bernoullis. This is Daniel Bernoulli. He seems to have been nicer than some of the others. Daniel was actually a doctor who was interested in mathematics.He lived from 1700-1782.

28.4.8 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. 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. The possibility exists that the reason for this is the numerical methods used to solve the partial differential equations may be better physical models than the ballance laws themselves. It is an area where people still sometimes disagree.

28.4.9 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 $\mathbf{x} \neq \mathbf{x}_0$,

$$\nabla\cdot\nabla\left(qk\frac{1}{\left|\mathbf{x}-\mathbf{x}_0\right|}\right) = \nabla\cdot\mathbf{E} = 0. \tag{28.14}$$

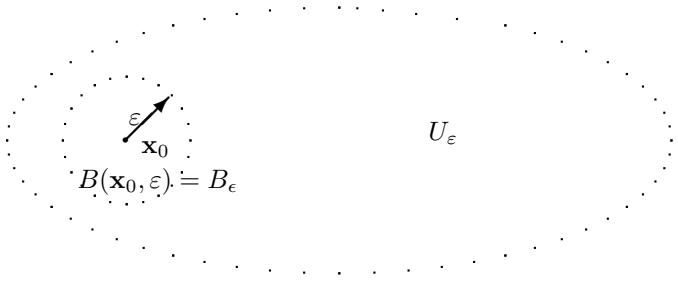
This is left as an exercise for you to verify.

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,

$$\begin{aligned}\int_{\partial B_\varepsilon}\mathbf{E}\cdot\mathbf{n}dS &= -\int_{\partial B_\varepsilon} k\frac{q\left(\mathbf{x}-\mathbf{x}_0\right)}{\left|\mathbf{x}-\mathbf{x}_0\right|^3}\cdot\frac{\mathbf{x}-\mathbf{x}_0}{|\mathbf{x}-\mathbf{x}_0|}dS\\ &= -kq\int_{\partial B_\varepsilon}\frac{1}{\left|\mathbf{x}-\mathbf{x}_0\right|^2}dS = \frac{-kq}{\varepsilon^2}\int_{\partial B_\varepsilon}dS\\ &= \frac{-kq}{\varepsilon^2}4\pi\varepsilon^2 = -4\pi kq.\end{aligned}$$

Therefore, from the divergence theorem and observation 28.14,

$$-4\pi kq+\int_{\partial U}\mathbf{E}\cdot\mathbf{n}dS=\int_{\partial U_\varepsilon}\mathbf{E}\cdot\mathbf{n}dS=\int_{U_\varepsilon}\nabla\cdot\mathbf{E}dV=0.$$

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.

Part XIII

Stoke's Theorem

Outcomes

Circulation Density and Curl

A. Explain what is meant by the circulation density and curl of a vector field.

B. Evaluate the curl of a vector field

C. Derive and apply formulas involving divergence, gradient and curl.

Reading: Multivariable Calculus 5.3
Outcome Mapping:

A. O1,3

B. 1,4,6

C. 2

Stoke's Theorem

A. Recall and verify Stoke's theorem.

B. Apply Stoke's theorem to calculate the circulation (or work) of a vector field around a simple closed curve.

C. Recall and apply the divergence and curl tests.

Reading: Multivariable Calculus 5.4
Outcome Mapping:

A. P1,1,12

B. 2,3,9

C. P2,4,5,10

Stoke's Theorem 4-5 Dec.

29.1 Curl Of A Vector Field

Here the important concepts of curl is defined.

Definition 29.1.1 *Let $\mathbf{f} : U \to \mathbb{R}^3$ for $U \subseteq \mathbb{R}^3$ denote a vector field. The **curl** of the vector field yields another vector field and it is defined as follows.*

$$\left(\operatorname{curl}(\mathbf{f})(\mathbf{x})\right)_i \equiv \left(\nabla \times \mathbf{f}(\mathbf{x})\right)_i \equiv \varepsilon_{ijk}\partial_j f_k(\mathbf{x})$$

where here ∂_j means the partial derivative with respect to x_j and the subscript of i in $\left(\operatorname{curl}(\mathbf{f})(\mathbf{x})\right)_i$ 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.)

This definition is in terms of a given coordinate system but later coordinate free definitions of the curl is presented. For now, everything is defined in terms of a given Cartesian coordinate system. The curl has profound physical significance and this will be discussed later. For now it is important to understand how to find it. Be sure you understand that for $\mathbf{f}$ a vector field, $\operatorname{curl}\mathbf{f}$ is another vector field.

Example 29.1.2 *Let $\mathbf{f}(\mathbf{x}) = xy\mathbf{i} + (z-y)\mathbf{j} + (\sin(x)+z)\mathbf{k}$. Find $\operatorname{curl}\mathbf{f}$.*

$\operatorname{curl}\mathbf{f}$ is obtained by evaluating

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ xy & z-y & \sin(x)+z \end{vmatrix} =$$

$$\mathbf{i}\left(\frac{\partial}{\partial y}(\sin(x)+z) - \frac{\partial}{\partial z}(z-y)\right) - \mathbf{j}\left(\frac{\partial}{\partial x}(\sin(x)+z) - \frac{\partial}{\partial z}(xy)\right) +$$

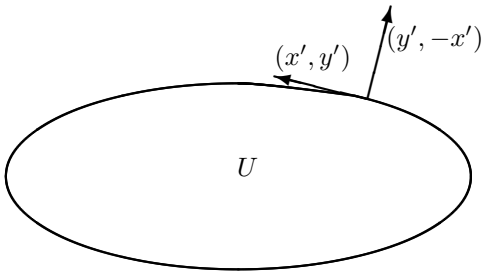
$$\mathbf{k}\left(\frac{\partial}{\partial x}(z-y) - \frac{\partial}{\partial y}(xy)\right) = -\mathbf{i} - \cos(x)\mathbf{j} - x\mathbf{k}.$$

29.2 Green's Theorem, A Review

Theorem 29.2.1 *(Green's Theorem) Let U be an open set in the plane and let ∂U be piecewise smooth and let $\mathbf{F}(x,y) = (P(x,y), Q(x,y))$ be a C^1 vector field defined near U. Then it is often[1] the case that*

$$\int_{\partial U} \mathbf{F} \cdot d\mathbf{R} = \int_U \left(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y) \right) dA.$$

Proof: Suppose the divergence theorem holds for U. Consider the following picture.



Since it is assumed that motion around U is counter clockwise, the tangent vector, (x', y') is as shown. Now the unit exterior normal is either

$$\frac{1}{\sqrt{(x')^2 + (y')^2}} (-y', x')$$

or

$$\frac{1}{\sqrt{(x')^2 + (y')^2}} (y', -x')$$

Again, the counter clockwise motion shows the correct unit exterior normal is the second of the above. To see this note that since the area should be on the left as you walk around the edge, you need to have the unit normal point in the direction of $(x', y', 0) \times \mathbf{k}$ which equals $(y', -x', 0)$. Now let $\mathbf{F}(x,y) = (Q(x,y), -P(x,y))$. Also note the area element on ∂U is $\sqrt{(x')^2 + (y')^2}dt$. Suppose the boundary of U consists of m smooth curves, the i^{th} of which is parameterized by (x_i, y_i) with the parameter, $t \in [a_i, b_i]$. Then by the divergence theorem,

$$\int_U (Q_x - P_y)\, dA = \int_U \operatorname{div}(\mathbf{F})\, dA = \int_{\partial U} \mathbf{F} \cdot \mathbf{n} dS$$

$$= \sum_{i=1}^m \int_{a_i}^{b_i} (Q(x_i(t), y_i(t)), -P(x_i(t), y_i(t))) \cdot \frac{1}{\sqrt{(x_i')^2 + (y_i')^2}} (y_i', -x_i') \overbrace{\sqrt{(x_i')^2 + (y_i')^2} dt}^{dS}$$

$$= \sum_{i=1}^m \int_{a_i}^{b_i} (Q(x_i(t), y_i(t)), -P(x_i(t), y_i(t))) \cdot (y_i', -x_i')\, dt$$

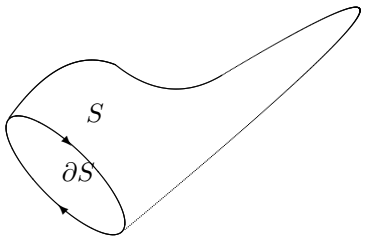
$$= \sum_{i=1}^m \int_{a_i}^{b_i} Q(x_i(t), y_i(t)) y_i'(t) + P(x_i(t), y_i(t)) x_i'(t)\, dt \equiv \int_{\partial U} P dx + Q dy$$

This proves Green's theorem from the divergence theorem.

[1] For a general version see the advanced calculus book by Apostol. The general versions involve the concept of a rectifiable Jordan curve.

29.3 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 29.3.1 *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 29.3.2 *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.*

$$(\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times \mathbf{F})(\mathbf{R}(u,v)) = ((\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u)(u,v). \tag{29.1}$$

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 29.1.

For those of you who do not know the permutation symbol and the reduction identities used in the above, it is possible, but more trouble, to establish the identity by brute force. Letting x, y, z denote the components of $\mathbf{R}(\mathbf{u})$ and f_1, f_2, f_3 denote the components of $\mathbf{F}$, and letting a subscripted variable denote the partial derivative with respect to that variable, the left side of 29.1 equals

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_u & y_u & z_u \\ x_v & y_v & z_v \end{vmatrix} \cdot \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial_x & \partial_y & \partial_z \\ f_1 & f_2 & f_3 \end{vmatrix}$$

$$= (f_{3y} - f_{2z})(y_u z_v - z_u y_v) + (f_{1z} - f_{3x})(z_u x_v - x_u z_v) + (f_{2x} - f_{1y})(x_u y_v - y_u x_v)$$

$$\begin{aligned} &= f_{3y} y_u z_v + f_{2z} z_u y_v + f_{1z} z_u x_v + f_{3x} x_u z_v + f_{2x} x_u y_v + f_{1y} y_u x_v \\ &- (f_{2z} y_u z_v + f_{3y} z_u y_v + f_{1z} x_u z_v + f_{3x} z_u x_v + f_{2x} y_u x_v + f_{1y} x_u y_v) \end{aligned}$$

$$\begin{aligned} &= f_{1y} y_u x_v + f_{1z} z_u x_v + f_{2x} x_u y_v + f_{2z} z_u y_v + f_{3x} x_u z_v + f_{3y} y_u z_v \\ &- (f_{1y} y_v x_u + f_{1z} z_v x_u + f_{2x} x_v y_u + f_{2z} z_v y_u + f_{3x} x_v z_u + f_{3y} y_v z_u) \end{aligned}$$

At this point I become clever and add in and subtract off certain terms. Then

$$\begin{aligned} &= f_{1x} x_u x_v + f_{1y} y_u x_v + f_{1z} z_u x_v + f_{2x} x_u y_v + f_{2x} y_u y_v \\ &\quad + f_{2z} z_u y_v + f_{3x} x_u z_v + f_{3y} y_u z_v + f_{3z} z_u z_v \\ &- \begin{pmatrix} f_{1x} x_v x_u + f_{1y} y_v x_u + f_{1z} z_v x_u + f_{2x} x_v y_u + f_{2x} y_v y_u \\ + f_{2z} z_v y_u + f_{3x} x_v z_u + f_{3y} y_v z_u + f_{3z} z_v z_u \end{pmatrix} \\ &= \frac{\partial f_1 \circ \mathbf{R}(u,v)}{\partial u} x_v + \frac{\partial f_2 \circ \mathbf{R}(u,v)}{\partial u} y_v + \frac{\partial f_3 \circ \mathbf{R}(u,v)}{\partial u} z_v \\ &- \left(\frac{\partial f_1 \circ \mathbf{R}(u,v)}{\partial v} x_u + \frac{\partial f_2 \circ \mathbf{R}(u,v)}{\partial v} y_u + \frac{\partial f_3 \circ \mathbf{R}(u,v)}{\partial v} z_u \right) \\ &= ((\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u)(u,v). \end{aligned}$$

This proves the lemma. Not how much trouble this was and how I had to be clever by adding in and subtracting off the appropriate terms. With the reduction identities for the permutation symbol no cleverness at all was required. The desired identity just fell out of completely routine manipulations. This is a good example which illustrates the utility of good notation.

The proof of Stoke's theorem given next follows [7]. First, it is convenient to give a definition.

Definition 29.3.3 *A vector valued function, $\mathbf{R}: U \subseteq \mathbb{R}^m \to \mathbb{R}^n$ is said to be in $C^k\left(\overline{U}, \mathbb{R}^n\right)$ 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 29.3.4 *(Stoke's Theorem) Let U be any region in $\mathbb{R}^2$ for which the conclusion of Green's theorem holds and let $\mathbf{R} \in C^2\left(\overline{U}, \mathbb{R}^3\right)$ 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,

$$\int_{\partial S} \mathbf{F} \cdot d\mathbf{R} = \int_S \operatorname{curl}(\mathbf{F}) \cdot \mathbf{n} dS$$

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,

$$\int_{\mathbf{R}(C)} \mathbf{F} \cdot d\mathbf{R} =$$

$$\begin{aligned}
&= \int_\alpha^\beta \mathbf{F}(\mathbf{R}(u(t), v(t))) \cdot (\mathbf{R}_u u'(t) + \mathbf{R}_v v'(t))\, dt \\
&= \int_\alpha^\beta \mathbf{F}(\mathbf{R}(u(t), v(t)))\, \mathbf{R}_u(u(t), v(t))\, u'(t)\, dt \\
&\quad + \int_\alpha^\beta \mathbf{F}(\mathbf{R}(u(t), v(t)))\, \mathbf{R}_v(u(t), v(t))\, v'(t)\, dt \\
&= \int_C ((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u, (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v) \cdot d\mathbf{r}.
\end{aligned}$$

Since this holds for each such piece of ∂U, it follows

$$\int_{\partial S} \mathbf{F} \cdot d\mathbf{R} = \int_{\partial U} ((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u, (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v) \cdot d\mathbf{r}.$$

By the assumption that the conclusion of Green's theorem holds for U, this equals

$$\begin{aligned}
&\int_U [((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_v)_u - ((\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_u)_v]\, dA \\
&= \int_U [(\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v + (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_{vu} - (\mathbf{F} \circ \mathbf{R}) \cdot \mathbf{R}_{uv} - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u]\, dA \\
&= \int_U [(\mathbf{F} \circ \mathbf{R})_u \cdot \mathbf{R}_v - (\mathbf{F} \circ \mathbf{R})_v \cdot \mathbf{R}_u]\, dA
\end{aligned}$$

the last step holding by equality of mixed partial derivatives, a result of the assumption that $\mathbf{R}$ is C^2. Now by Lemma 29.3.2, this equals

$$\begin{aligned}
&\int_U (\mathbf{R}_u \times \mathbf{R}_v) \cdot (\nabla \times \mathbf{F})\, dA \\
&= \int_U \nabla \times \mathbf{F} \cdot (\mathbf{R}_u \times \mathbf{R}_v)\, dA \\
&= \int_S \nabla \times \mathbf{F} \cdot \mathbf{n} dS
\end{aligned}$$

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

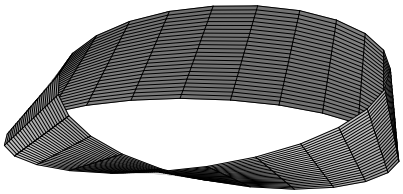
$$\begin{aligned}
(\mathbf{R}_u \times \mathbf{R}_v)\, dA &= \frac{(\mathbf{R}_u \times \mathbf{R}_v)}{|(\mathbf{R}_u \times \mathbf{R}_v)|} |(\mathbf{R}_u \times \mathbf{R}_v)|\, dA \\
&= \mathbf{n} dS.
\end{aligned}$$

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.

29.3.1 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

$$\frac{\left(4\sin\frac{1}{2}\theta\cos\theta - \frac{1}{2}v, 4\sin\frac{1}{2}\theta\sin\theta + \frac{1}{2}v, -8\cos^2\frac{1}{2}\theta\sin\frac{1}{2}\theta - 8\cos^3\frac{1}{2}\theta + 4\cos\frac{1}{2}\theta\right)}{\sqrt{16\sin^2\left(\frac{\theta}{2}\right) + \frac{v^2}{2} + 4\sin\left(\frac{\theta}{2}\right)v(\sin\theta - \cos\theta) + \left(-8\cos^2\frac{1}{2}\theta\sin\frac{1}{2}\theta - 8\cos^3\frac{1}{2}\theta + 4\cos\frac{1}{2}\theta\right)^2}}$$

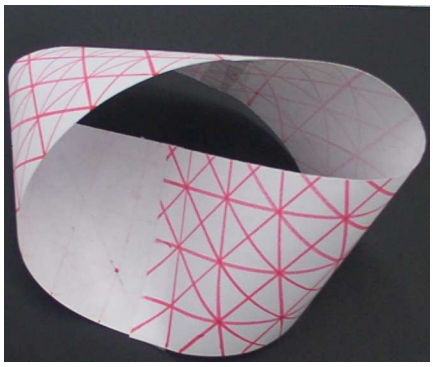
and you can verify that the denominator will not vanish. Letting $v = 0$ and $\theta = 0$ and 2π yields the two vectors,

$$(0, 0, -1), (0, 0, 1)$$

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.



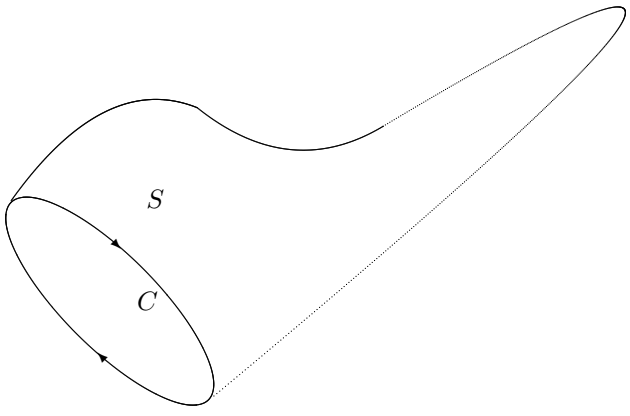
29.3.2 Conservative Vector Fields And Stoke's Theorem

Recall the following definition.

Definition 29.3.5 *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* $\int_C \mathbf{F}\cdot d\mathbf{R} = 0$.

Stokes theorem provides an easy to use criterion for determining whether a given vector field is conservative.

Definition 29.3.6 *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.



[2]There is no such thing as a liberal vector field.

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 29.3.7 *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 29.3.8 *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 $\phi(x,y,z) = y + x^4 + \sin\left(x^2+z^2\right)$.

29.3.3 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 29.3.8, 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.

29.3.4 Vector Identities*

There are many interesting identities which relate the gradient, divergence and curl.

Theorem 29.3.9 *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 60.

$$\begin{aligned}
\nabla \cdot (\nabla \times \mathbf{f}) &= \partial_i (\nabla \times \mathbf{f})_i \\
&= \partial_i (\varepsilon_{ijk} \partial_j f_k) \\
&= \varepsilon_{ijk} \partial_i (\partial_j f_k) \\
&= \varepsilon_{jik} \partial_j (\partial_i f_k) \\
&= -\varepsilon_{ijk} \partial_j (\partial_i f_k) \\
&= -\varepsilon_{ijk} \partial_i (\partial_j f_k) \\
&= -\nabla \cdot (\nabla \times \mathbf{f}) .
\end{aligned}$$

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 60 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.

29.3.5 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.

$$\mathbf{f}(x,y,z) = \left(\int_0^z g_2(x,y,t)\, dt, -\int_0^z g_1(x,y,t)\, dt + \int_0^x g_3(t,y,0)\, dt, 0 \right)^T. \tag{29.2}$$

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.

$$\frac{\partial}{\partial x} \int_a^b h(x,t)\, dt = \int_a^b \frac{\partial h}{\partial x}(x,t)\, dt. \tag{29.3}$$

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 29.3.10 *Suppose h and $\frac{\partial h}{\partial x}$ are continuous on the rectangle $R = [c,d] \times [a,b]$. Then 29.3 holds.*

29.3.6 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

$$\nabla \times \mathbf{E} + \frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} = \mathbf{0} \tag{29.4}$$

$$\nabla \cdot \mathbf{E} = 4\pi\rho \tag{29.5}$$

$$\nabla \times \mathbf{B} - \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} = \frac{4\pi}{c}\mathbf{f} \tag{29.6}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{29.7}$$

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 29.7 $\mathbf{B}$ has a vector potential, $\mathbf{A}_1$ such that $\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A}_1$. Now go to 29.4 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

$$\nabla \psi_1 = \mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{A}_1}{\partial t}. \tag{29.8}$$

Now suppose ϕ is a time dependent scalar field satisfying

$$\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = \frac{1}{c} \frac{\partial \psi_1}{\partial t} - \nabla \cdot \mathbf{A}_1. \tag{29.9}$$

Next define

$$\mathbf{A} \equiv \mathbf{A}_1 + \nabla \phi, \quad \psi \equiv \psi_1 + \frac{1}{c} \frac{\partial \phi}{\partial t}. \tag{29.10}$$

Therefore, in terms of the new variables, 29.9 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

$$0 = \frac{\partial \psi}{\partial t} - c \nabla \cdot A. \tag{29.11}$$

Then it follows from Theorem 29.3.9 on Page 534 that $\mathbf{A}$ is also a vector potential for $\mathbf{B}$. That is

$$\nabla \times \mathbf{A} = \mathbf{B}. \tag{29.12}$$

From 29.8

$$\nabla \left(\psi - \frac{1}{c} \frac{\partial \phi}{\partial t} \right) = \mathbf{E} + \frac{1}{c} \left(\frac{\partial A}{\partial t} - \nabla \frac{\partial \phi}{\partial t} \right)$$

and so

$$\nabla \psi = \mathbf{E} + \frac{1}{c} \frac{\partial \mathbf{A}}{\partial t}. \tag{29.13}$$

Using 29.6 and 29.13,

$$\nabla \times (\nabla \times \mathbf{A}) - \frac{1}{c} \frac{\partial}{\partial t} \left(\nabla \psi - \frac{1}{c} \frac{\partial \mathbf{A}}{\partial t} \right) = \frac{4\pi}{c} \mathbf{f}. \tag{29.14}$$

Now from Theorem 29.3.9 on Page 534 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 29.11, this gives

$$\frac{1}{c^2} \frac{\partial^2 \mathbf{A}}{\partial t^2} - \nabla^2 \mathbf{A} = \frac{4\pi}{c} \mathbf{f}. \tag{29.15}$$

Also from 29.13, 29.5, and 29.11,

$$\begin{aligned} \nabla^2 \psi &= \nabla \cdot \mathbf{E} + \frac{1}{c} \frac{\partial}{\partial t} (\nabla \cdot \mathbf{A}) \\ &= 4\pi \rho + \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} \end{aligned}$$

and so

$$\frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} - \nabla^2\psi = -4\pi\rho. \tag{29.16}$$

This is very interesting. If a solution to the wave equations, 29.16, and 29.15 can be found along with a solution to 29.11, then letting the magnetic field be given by 29.12 and letting $\mathbf{E}$ be given by 29.13 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 29.11 to hold.

Part XIV

Some Iterative Techniques For Linear Algebra

Iterative Methods For Linear Systems

Consider the problem of solving the equation

$$A\mathbf{x} = \mathbf{b} \tag{30.1}$$

where A is an $n \times n$ matrix. In many applications, the matrix A is huge and composed mainly of zeros. For such matrices, the method of Gauss elimination (row operations) is not a good way to solve the system because the row operations can destroy the zeros and storing all those zeros takes a lot of room in a computer. These systems are called sparse. To solve them it is common to use an iterative technique. I am following the treatment given to this subject by Nobel and Daniel [20].

There are two main methods which are used to obtain solutions iteratively, the Jacobi method and the Gauss Seidel method. I will illustrate with an example and then describe the method precisely.

30.1 Jacobi Method

Example 30.1.1 *Use the Jacobi method to find the solutions to the following system of equations.*

$$\begin{aligned} 7x + y &= 11 \\ x - 5y &= 7 \end{aligned}$$

It is profoundly stupid to use the Jacobi method on such a 2×2 system. You should simply use row operations. If you do, the solution is $\left\{y = -\frac{19}{18}, x = \frac{31}{18}\right\}$. In terms of decimals this is $\{y = -1.055\,555\,56, x = 1.722\,222\,22\}$. Now I will proceed to show how to use the Jacobi method to also find this solution.

Here are steps which describe the Jacobi method. You write the system as

$$\begin{pmatrix} 7 & 1 \\ 1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 11 \\ 7 \end{pmatrix}$$

Next you split the matrix as follows

$$\begin{pmatrix} 7 & 0 \\ 0 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 11 \\ 7 \end{pmatrix}$$

That is you write the matrix as the sum of a diagonal matrix plus the off diagonal terms. Then if you have a solution, you would need

$$\begin{pmatrix} 7 & 0 \\ 0 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = - \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 11 \\ 7 \end{pmatrix}$$

You could write this as

$$\begin{aligned}\begin{pmatrix} x \\ y \end{pmatrix} &= -\begin{pmatrix} 7 & 0 \\ 0 & -5 \end{pmatrix}^{-1}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}+\begin{pmatrix} 7 & 0 \\ 0 & -5 \end{pmatrix}^{-1}\begin{pmatrix} 11 \\ 7 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -\frac{1}{7} \\ \frac{1}{5} & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}+\begin{pmatrix} \frac{11}{7} \\ -\frac{7}{5} \end{pmatrix}\end{aligned}$$

This suggests a way to approach the problem through a process of iterations. You pick an initial guess for (x, y) say $(0, 0)$. (It really doesn't matter what you pick. When the method works it will do so for any initial choice.) Call this initial guess

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

and then you obtain the next guess, $\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$ as follows

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{7} \\ \frac{1}{5} & 0 \end{pmatrix}\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}+\begin{pmatrix} \frac{11}{7} \\ -\frac{7}{5} \end{pmatrix}$$

Then to get the next guess you do the same thing.

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{7} \\ \frac{1}{5} & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}+\begin{pmatrix} \frac{11}{7} \\ -\frac{7}{5} \end{pmatrix}$$

Continuing this way, this method hopefully will give guesses which are increasingly close to the true solution. Lets apply this to this example.

$$\begin{aligned}\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} &= \begin{pmatrix} 0 & -\frac{1}{7} \\ \frac{1}{5} & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix}+\begin{pmatrix} \frac{11}{7} \\ -\frac{7}{5} \end{pmatrix} \\ &= \begin{pmatrix} 1.571\,428\,57 \\ -1.4 \end{pmatrix}\end{aligned}$$

Now you find the next guess.

$$\begin{aligned}\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} &= \begin{pmatrix} 0 & -\frac{1}{7} \\ \frac{1}{5} & 0 \end{pmatrix}\begin{pmatrix} 1.571\,428\,57 \\ -1.4 \end{pmatrix}+\begin{pmatrix} \frac{11}{7} \\ -\frac{7}{5} \end{pmatrix} \\ &= \begin{pmatrix} 1.771\,428\,57 \\ -1.085\,714\,29 \end{pmatrix}\end{aligned}$$

Things are still changing so I will try the next guess.

$$\begin{aligned}\begin{pmatrix} x_3 \\ y_3 \end{pmatrix} &= \begin{pmatrix} 0 & -\frac{1}{7} \\ \frac{1}{5} & 0 \end{pmatrix}\begin{pmatrix} 1.771\,428\,57 \\ -1.085\,714\,29 \end{pmatrix}+\begin{pmatrix} \frac{11}{7} \\ -\frac{7}{5} \end{pmatrix} \\ &= \begin{pmatrix} 1.726\,530\,61 \\ -1.045\,714\,29 \end{pmatrix}\end{aligned}$$

Lets do another iteration.

$$\begin{aligned}\begin{pmatrix} x_4 \\ y_4 \end{pmatrix} &= \begin{pmatrix} 0 & -\frac{1}{7} \\ \frac{1}{5} & 0 \end{pmatrix}\begin{pmatrix} 1.726\,530\,61 \\ -1.045\,714\,29 \end{pmatrix}+\begin{pmatrix} \frac{11}{7} \\ -\frac{7}{5} \end{pmatrix} \\ &= \begin{pmatrix} 1.720\,816\,33 \\ -1.054\,693\,88 \end{pmatrix}.\end{aligned}$$

This should be pretty close because the guesses are not changing much from one to the next. The exact solution was

$$\{y = -1.055\,555\,56, x = 1.722\,222\,22\}$$

Actually, you don't do it this way. The following gives the way a computer would do it. You do not invert the matrix as I did. However, for the purposes of illustration and for small systems there is no harm in doing it as I did above, especially since for small systems of equations it is a stupid idea to use an iterative method in the first place.

Definition 30.1.2 *The Jacobi iterative technique, also called the method of simultaneous corrections is defined as follows. Let* $\mathbf{x}^1$ *be an initial vector, say the zero vector or some other vector. The method generates a succession of vectors,* $\mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^4, \cdots$ *and hopefully this sequence of vectors will converge to the solution to 30.1. The vectors in this list are called iterates and they are obtained according to the following procedure. Letting* $A = (a_{ij})$,

$$a_{ii}x_i^{r+1} = -\sum_{j\neq i} a_{ij}x_j^r + b_i. \tag{30.2}$$

In terms of matrices, letting

$$A = \begin{pmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix}$$

The iterates are defined as

$$\begin{aligned} &\begin{pmatrix} * & 0 & \cdots & 0 \\ 0 & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & * \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ \vdots \\ x_n^{r+1} \end{pmatrix} \\ = \quad &-\begin{pmatrix} 0 & * & \cdots & * \\ * & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ \vdots \\ x_n^r \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \end{aligned} \tag{30.3}$$

The matrix on the left in 30.3 is obtained by retaining the main diagonal of A and setting every other entry equal to zero. The matrix on the right in 30.3 is obtained from A by setting every diagonal entry equal to zero and retaining all the other entries unchanged.

Example 30.1.3 *Use the Jacobi method to solve the system*

$$\begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

Of course this is solved most easily using row reductions. The Jacobi method is useful when the matrix is 1000×1000 or larger. This example is just to illustrate how the method works. First lets solve it using row operations. The augmented matrix is

$$\begin{pmatrix} 3 & 1 & 0 & 0 & 1 \\ 1 & 4 & 1 & 0 & 2 \\ 0 & 2 & 5 & 1 & 3 \\ 0 & 0 & 2 & 4 & 4 \end{pmatrix}$$

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \frac{6}{29} \\ 0 & 1 & 0 & 0 & \frac{11}{29} \\ 0 & 0 & 1 & 0 & \frac{8}{29} \\ 0 & 0 & 0 & 1 & \frac{25}{29} \end{pmatrix}$$

which in terms of decimals is approximately equal to

$$\begin{pmatrix} 1.0 & 0 & 0 & 0 & .206 \\ 0 & 1.0 & 0 & 0 & .379 \\ 0 & 0 & 1.0 & 0 & .275 \\ 0 & 0 & 0 & 1.0 & .862 \end{pmatrix}.$$

In terms of the matrices, the Jacobi iteration is of the form

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Multiplying by the invese of the matrix on the left, [1]this iteration reduces to

$$\begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix}. \tag{30.4}$$

Now iterate this starting with

$$\mathbf{x}^1 \equiv \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Thus

$$\mathbf{x}^2 = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix}$$

Then

$$\mathbf{x}^3 = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix}}^{\mathbf{x}_2} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .166 \\ .26 \\ .2 \\ .7 \end{pmatrix}$$

$$\mathbf{x}^4 = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} .166 \\ .26 \\ .2 \\ .7 \end{pmatrix}}^{\mathbf{x}_3} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .24 \\ .408\,5 \\ .356 \\ .9 \end{pmatrix}$$

[1]You certainly would not compute the invese in solving a large system. This is just to show you how the method works for this simple example. You would use the first description in terms of indices.

$$\mathbf{x}^5 = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} .24 \\ .408\,5 \\ .356 \\ .9 \end{pmatrix}}^{\mathbf{x}_4} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .197 \\ .351 \\ .256\,6 \\ .822 \end{pmatrix}$$

$$\mathbf{x}^6 = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} .197 \\ .351 \\ .256\,6 \\ .822 \end{pmatrix}}^{\mathbf{x}^5} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .216 \\ .386 \\ .295 \\ .871 \end{pmatrix}.$$

You can keep going like this. Recall the solution is approximately equal to

$$\begin{pmatrix} .206 \\ .379 \\ .275 \\ .862 \end{pmatrix}$$

so you see that with no care at all and only 6 iterations, an approximate solution has been obtained which is not too far off from the actual solution.

It is important to realize that a computer would use 30.2 directly. Indeed, writing the problem in terms of matrices as I have done above destroys every benefit of the method. However, it makes it a little easier to see what is happening and so this is why I have presented it in this way.

30.2 Gauss Seidel Method

Example 30.2.1 *Solve the following system of equations using the Gauss Seidel method. It is the same example as in Example 30.1.1.*

$$\begin{aligned} 7x + y &= 11 \\ x - 5y &= 7 \end{aligned}$$

The solution to this system is is : $\{y = -1.055\,555\,56, x = 1.722\,222\,22\}$. Now I will use the Gauss Seidel method to get this solution. The system is of the form

$$\begin{pmatrix} 7 & 0 \\ 1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 11 \\ 7 \end{pmatrix}$$

Note the difference! Here you split the matrix differently. Then the iteration scheme is just as before,

$$\begin{pmatrix} 7 & 0 \\ 1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = -\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 11 \\ 7 \end{pmatrix}$$

and so the solution satisfies

$$\begin{aligned} \begin{pmatrix} x \\ y \end{pmatrix} &= -\begin{pmatrix} 7 & 0 \\ 1 & -5 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 7 & 0 \\ 1 & -5 \end{pmatrix}^{-1} \begin{pmatrix} 11 \\ 7 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -\frac{1}{7} \\ 0 & -\frac{1}{35} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \frac{11}{7} \\ -\frac{38}{35} \end{pmatrix} \end{aligned}$$

The corresponding iteration scheme yields

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{7} \\ 0 & -\frac{1}{35} \end{pmatrix} \begin{pmatrix} x_n \\ y_n \end{pmatrix} + \begin{pmatrix} \frac{11}{7} \\ -\frac{38}{35} \end{pmatrix}$$

Starting with an initial guess of $x_0 = y_0 = 0$,consider the following iterations.

$$\begin{aligned}\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} &= \begin{pmatrix} 0 & -\frac{1}{7} \\ 0 & -\frac{1}{35} \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{11}{7} \\ -\frac{38}{35} \end{pmatrix} \\ &= \begin{pmatrix} 1.571\,428\,57 \\ -1.085\,714\,29 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} &= \begin{pmatrix} 0 & -\frac{1}{7} \\ 0 & -\frac{1}{35} \end{pmatrix}\begin{pmatrix} 1.571\,428\,57 \\ -1.085\,714\,29 \end{pmatrix} + \begin{pmatrix} \frac{11}{7} \\ -\frac{38}{35} \end{pmatrix} \\ &= \begin{pmatrix} 1.726\,530\,61 \\ -1.054\,693\,88 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\begin{pmatrix} x_3 \\ y_3 \end{pmatrix} &= \begin{pmatrix} 0 & -\frac{1}{7} \\ 0 & -\frac{1}{35} \end{pmatrix}\begin{pmatrix} 1.726\,530\,61 \\ -1.054\,693\,88 \end{pmatrix} + \begin{pmatrix} \frac{11}{7} \\ -\frac{38}{35} \end{pmatrix} \\ &= \begin{pmatrix} 1.722\,099\,13 \\ -1.055\,580\,17 \end{pmatrix}\end{aligned}$$

These guesses are pretty close so it seems this should be close. Note the exact solution is $\{y = -1.055\,555\,56, x = 1.722\,222\,22\}$. I think you can see this method worked a little better than the Jacobi method although both are pretty good.

The following is the precise description of the method. As before, you don't write out the matrices and invert that matrix like above.

Definition 30.2.2 *The Gauss Seidel method, also called the method of successive corrections is given as follows. For $A = (a_{ij})$, the iterates for the problem $A\mathbf{x} = \mathbf{b}$ are obtained according to the formula*

$$\sum_{j=1}^{i} a_{ij} x_j^{r+1} = -\sum_{j=i+1}^{n} a_{ij} x_j^r + b_i. \tag{30.5}$$

In terms of matrices, letting

$$A = \begin{pmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix}$$

The iterates are defined as

$$\begin{aligned}&\begin{pmatrix} * & 0 & \cdots & 0 \\ * & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & * \end{pmatrix}\begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ \vdots \\ x_n^{r+1} \end{pmatrix} \\ &= -\begin{pmatrix} 0 & * & \cdots & * \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & 0 \end{pmatrix}\begin{pmatrix} x_1^r \\ x_2^r \\ \vdots \\ x_n^r \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}\end{aligned} \tag{30.6}$$

In words, you set every entry in the original matrix which is strictly above the main diagonal equal to zero to obtain the matrix on the left. To get the matrix on the right, you set every entry of A which is on or below the main diagonal equal to zero. Using the iteration procedure of 30.5 directly, the Gauss Seidel method makes use of the very latest information which is available at that stage of the computation.

The following example is the same as the example used to illustrate the Jacobi method.

Example 30.2.3 *Use the Gauss Seidel method to solve the system*

$$\begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

In terms of matrices, this procedure is

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 2 & 5 & 0 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Multiplying by the inverse of the matrix on the left[2] this yields

$$\begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\ 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix}$$

As before, I will be totally unoriginal in the choice of $\mathbf{x}^1$. Let it equal the zero vector. Therefore,

$$\mathbf{x}^2 = \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix}.$$

Now

$$\mathbf{x}^3 = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\ 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \overbrace{\begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix}}^{\mathbf{x}^2} + \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix} = \begin{pmatrix} .194 \\ .343 \\ .306 \\ .846 \end{pmatrix}.$$

It follows

$$\mathbf{x}^4 = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\ 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \overbrace{\begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix}}^{\mathbf{x}^3} + \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix} = \begin{pmatrix} .194 \\ .343 \\ .306 \\ .846 \end{pmatrix}$$

[2] As in the case of the Jacobi iteration, the computer would not do this. It would use the iteration procedure in terms of the entries of the matrix directly. Otherwise all benefit to using this method is lost.

and so

$$\mathbf{x}^5 = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{12} & \frac{1}{4} & 0 \\ 0 & \frac{1}{30} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{60} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \overbrace{\begin{pmatrix} .194 \\ .343 \\ .306 \\ .846 \end{pmatrix}}^{\mathbf{x}^4} + \begin{pmatrix} \frac{1}{3} \\ \frac{5}{12} \\ \frac{13}{30} \\ \frac{47}{60} \end{pmatrix} = \begin{pmatrix} .219 \\ .368\,75 \\ .283\,3 \\ .858\,35 \end{pmatrix}.$$

Recall the answer is

$$\begin{pmatrix} .206 \\ .379 \\ .275 \\ .862 \end{pmatrix}$$

so the iterates are already pretty close to the answer. You could continue doing these iterates and it appears they converge to the solution. Now consider the following example.

Example 30.2.4 *Use the Gauss Seidel method to solve the system*

$$\begin{pmatrix} 1 & 4 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

The exact solution is given by doing row operations on the augmented matrix. When this is done the row echelon form is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 6 \\ 0 & 1 & 0 & 0 & -\frac{5}{4} \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \end{pmatrix}$$

and so the solution is approximately

$$\begin{pmatrix} 6 \\ -\frac{5}{4} \\ 1 \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 6.0 \\ -1.25 \\ 1.0 \\ .5 \end{pmatrix}$$

The Gauss Seidel iterations are of the form

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 2 & 5 & 0 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = -\begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

and so, multiplying by the inverse of the matrix on the left, the iteration reduces to the following in terms of matrix multiplication.

$$\mathbf{x}^{r+1} = -\begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & -1 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \mathbf{x}^r + \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{pmatrix}.$$

This time, I will pick an initial vector close to the answer. Let

$$\mathbf{x}^1 = \begin{pmatrix} 6 \\ -1 \\ 1 \\ \frac{1}{2} \end{pmatrix}$$

This is very close to the answer. Now lets see what the Gauss Seidel iteration does to it.

$$\mathbf{x}^2 = -\begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & -1 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} 6 \\ -1 \\ 1 \\ \frac{1}{2} \end{pmatrix} + \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} 5.0 \\ -1.0 \\ .9 \\ .55 \end{pmatrix}$$

You can't expect to be real close after only one iteration. Lets do another.

$$\mathbf{x}^3 = -\begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & -1 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} 5.0 \\ -1.0 \\ .9 \\ .55 \end{pmatrix} + \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} 5.0 \\ -.975 \\ .88 \\ .56 \end{pmatrix}$$

$$\mathbf{x}^4 = -\begin{pmatrix} 0 & 4 & 0 & 0 \\ 0 & -1 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & -\frac{1}{10} & \frac{1}{5} \\ 0 & -\frac{1}{5} & \frac{1}{20} & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} 5.0 \\ -.975 \\ .88 \\ .56 \end{pmatrix} + \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{1}{2} \\ \frac{3}{4} \end{pmatrix} = \begin{pmatrix} 4.9 \\ -.945 \\ .866 \\ .567 \end{pmatrix}$$

The iterates seem to be getting farther from the actual solution. Why is the process which worked so well in the other examples not working here? A better question might be: Why does either process ever work at all?.

Both iterative procedures for solving

$$A\mathbf{x} = \mathbf{b} \tag{30.7}$$

are of the form

$$B\mathbf{x}^{r+1} = -C\mathbf{x}^r + \mathbf{b}$$

where $A = B + C$. In the Jacobi procedure, the matrix C was obtained by setting the diagonal of A equal to zero and leaving all other entries the same while the matrix, B was obtained by making every entry of A equal to zero other than the diagonal entries which are left unchanged. In the Gauss Seidel procedure, the matrix B was obtained from A by making every entry strictly above the main diagonal equal to zero and leaving the others unchanged and C was obtained from A by making every entry on or below the main diagonal equal to zero and leaving the others unchanged. Thus in the Jacobi procedure, B is a diagonal matrix while in the Gauss Seidel procedure, B is lower triangular. Using matrices to explicitly solve for the iterates, yields

$$\mathbf{x}^{r+1} = -B^{-1}C\mathbf{x}^r + B^{-1}\mathbf{b}. \tag{30.8}$$

This is what you would never have the computer do but this is what will allow the statement of a theorem which gives the condition for convergence of these and all other similar methods. Let $\{\lambda_1, \cdots, \lambda_n\}$ be the eigenvalues.

Definition 30.2.5 *The spectral radius of a matrix, M, denoted as $\rho(M)$ is*

$$\max\{|\lambda_1|, \cdots, |\lambda_n|\}.$$

That is it is the maximum of the absolute values of the eigenvalues of M.

The following gives the condition under which any of these iterates as in 30.8 converge.

Theorem 30.2.6 *Suppose $\rho\left(B^{-1}C\right) < 1$. Then the iterates in 30.8 converge to the unique solution of 30.7.*

The following definition is useful.

Definition 30.2.7 *Suppose A is an $n \times n$ matrix. Then A is said to be strictly diagonally dominant if for every i,*

$$|A_{ii}| > \sum_{j \pm i} |A_{ij}| .$$

That is, the absolute value of the entry in the ii^{th} position is larger than the sum of the absolute values of all the other entries on the i^{th} row.

Theorem 30.2.8 *In either the Jacobi or the Gauss Seidel methods, if the matrix of coefficients which gets split to yield an iteration technique is strictly diagonally dominant, then the method converges. This means the iterates get close to the solution to the original system of equations as the iteration progresses.*

Iterative Methods For Finding Eigenvalues

Quiz

1. Let $\mathbf{F} = (x, y, zx)$ and let S be the surface having parameterization $\mathbf{r}(u, v) = (uv, u + v, v)$ for $(u, v) \in [0, 1] \times [0, 2]$. Find the flux integral,

$$\int_S \mathbf{F} \cdot \mathbf{n} dS$$

   where $\mathbf{n}$ is the unit normal to the surface which has the same direction as the parametric normal, $\mathbf{N}(u, v) = \mathbf{r}_u \times \mathbf{r}_v$.

2. Find the flux integral,

$$\int_S \mathbf{F} \cdot \mathbf{n} dS$$

   of $\mathbf{F} = \nabla \times \mathbf{G}$ where $\mathbf{G}(x, y, z) = \left(\sin\left(x^2yz\right), \ln\left(x^4 + 7z^2 + 1\right), e^{x^2+y^2} z^5 \sin(z)\right)$ on the level surface, S given by the ellipsoid $x^2/6 + y^2/7 + z^2/2 = 1$.

3. Let C be the oriented curve consisting of directed line segments which go from $(0, 0, 0)$ to $(3, 2, 1)$ to $(1, 2, 3)$ to $(33.5, 45.7, 67.23)$ and then to $(1, 1, 1)$. Find the line integral,

$$\int_C (2xy + 1)\, dx + \left(x^2 + 1\right) dy + 2z dz.$$

4. Find the Laplacian of $x^3 - 3xy^2$.

5. Find the circulation density (curl) of the vector field $\left(x^2y, yz, z + x\right)$.

31.1 The Power Method For Eigenvalues

As indicated earlier, the eigenvalue eigenvector problem is extremely difficult. Consider for example what happens if you cannot find the eigenvalues exactly. Then you can't find an eigenvector because there isn't one due to the fact that $A - \lambda I$ is invertible whenever λ is not exactly equal to an eigenvalue. Therefore the straightforward way of solving this problem fails right away, even if you can approximate the eigenvalues. The power method allows you to approximate the largest eigenvalue and also the eigenvector which goes with it. By considering the inverse of the matrix, you can also find the smallest eigenvalue.

The method works in the situation of a nondefective matrix, A which has an eigenvalue of algebraic multiplicity 1, λ_n which has the property that $|\lambda_k| < |\lambda_n|$ for all $k \neq n$. Note that for a real matrix this excludes the case that λ_n could be complex. Why? Such an eigenvalue is called a dominant eigenvalue.

Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a basis of eigenvectors for $\mathbb{F}^n$ such that $A\mathbf{x}_n = \lambda_n \mathbf{x}_n$. Now let $\mathbf{u}_1$ be some nonzero vector. Since $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ is a basis, there exists unique scalars, c_i such that

$$\mathbf{u}_1 = \sum_{k=1}^{n} c_k \mathbf{x}_k.$$

Assume you have not been so unlucky as to pick $\mathbf{u}_1$ in such a way that $c_n = 0$. Then let $A\mathbf{u}_k = \mathbf{u}_{k+1}$ so that

$$\mathbf{u}_m = A^m \mathbf{u}_1 = \sum_{k=1}^{n-1} c_k \lambda_k^m \mathbf{x}_k + \lambda_n^m c_n \mathbf{x}_n. \tag{31.1}$$

For large m the last term, $\lambda_n^m c_n \mathbf{x}_n$, determines quite well the direction of the vector on the right. This is because $|\lambda_n|$ is larger than $|\lambda_k|$ and so for a large m, the sum, $\sum_{k=1}^{n-1} c_k \lambda_k^m \mathbf{x}_k$, on the right is fairly insignificant. Therefore, for large m, $\mathbf{u}_m$ is essentially a multiple of the eigenvector, $\mathbf{x}_n$, the one which goes with λ_n. The only problem is that there is no control of the size of the vectors $\mathbf{u}_m$. You can fix this by scaling. Let S_2 denote the entry of $A\mathbf{u}_1$ which is largest in absolute value. We call this a **scaling factor**. Then $\mathbf{u}_2$ will not be just $A\mathbf{u}_1$ but $A\mathbf{u}_1/S_2$. Next let S_3 denote the entry of $A\mathbf{u}_2$ which has largest absolute value and define $\mathbf{u}_3 \equiv A\mathbf{u}_2/S_3$. Continue this way. The scaling just described does not destroy the relative insignificance of the term involving a sum in 31.1. Indeed it amounts to nothing more than changing the units of length. Also note that from this scaling procedure, the absolute value of the largest element of $\mathbf{u}_k$ is always equal to 1. Therefore, for large m,

$$\mathbf{u}_m = \frac{\lambda_n^m c_n \mathbf{x}_n}{S_2 S_3 \cdots S_m} + (\text{relatively insignificant term}).$$

Therefore, the entry of $A\mathbf{u}_m$ which has the largest absolute value is essentially equal to the entry having largest absolute value of

$$A\left(\frac{\lambda_n^m c_n \mathbf{x}_n}{S_2 S_3 \cdots S_m}\right) = \frac{\lambda_n^{m+1} c_n \mathbf{x}_n}{S_2 S_3 \cdots S_m} \approx \lambda_n \mathbf{u}_m$$

and so for large m, it must be the case that $\lambda_n \approx S_{m+1}$. This suggests the following procedure.

Finding the largest eigenvalue with its eigenvector.

1. Start with a vector, $\mathbf{u}_1$ which you hope has a component in the direction of $\mathbf{x}_n$. The vector, $(1, \cdots, 1)^T$ is usually a pretty good choice.

2. If $\mathbf{u}_k$ is known,
$$\mathbf{u}_{k+1} = \frac{A\mathbf{u}_k}{S_{k+1}}$$
where S_{k+1} is the entry of $A\mathbf{u}_k$ which has largest absolute value.

3. When the scaling factors, S_k are not changing much, S_{k+1} will be close to the eigenvalue and $\mathbf{u}_{k+1}$ will be close to an eigenvector.

4. Check your answer to see if it worked well.

Example 31.1.1 *Find the largest eigenvalue of* $A = \begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}$.

The power method will now be applied to find the largest eigenvalue for the above matrix. Letting $\mathbf{u}_1 = (1, \cdots, 1)^T$, we will consider $A\mathbf{u}_1$ and scale it.

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ -4 \\ 6 \end{pmatrix}.$$

Scaling this vector by dividing by the largest entry gives

$$\frac{1}{6} \begin{pmatrix} 2 \\ -4 \\ 6 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ 1 \end{pmatrix} = \mathbf{u}_2$$

Now lets do it again.

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ 1 \end{pmatrix} = \begin{pmatrix} 22 \\ -8 \\ -6 \end{pmatrix}$$

Then

$$\mathbf{u}_3 = \frac{1}{22} \begin{pmatrix} 22 \\ -8 \\ -6 \end{pmatrix} = \begin{pmatrix} 1 \\ -\frac{4}{11} \\ -\frac{3}{11} \end{pmatrix} = \begin{pmatrix} 1.0 \\ -.363\,636\,36 \\ -.272\,727\,27 \end{pmatrix}.$$

Continue doing this

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.363\,636\,36 \\ -.272\,727\,27 \end{pmatrix} = \begin{pmatrix} 7.090\,909\,1 \\ -4.363\,636\,4 \\ 1.636\,363\,7 \end{pmatrix}$$

Then

$$\mathbf{u}_4 = \begin{pmatrix} 1.0 \\ -.615\,38 \\ .230\,77 \end{pmatrix}$$

So far the scaling factors are changing fairly noticeably so continue.

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.615\,38 \\ .230\,77 \end{pmatrix} = \begin{pmatrix} 16.154 \\ -7.384\,6 \\ -1.384\,6 \end{pmatrix}$$

$$\mathbf{u}_5 = \begin{pmatrix} 1.0 \\ -.457\,14 \\ -8.571\,3 \times 10^{-2} \end{pmatrix}$$

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.457\,14 \\ -8.571\,3 \times 10^{-2} \end{pmatrix} = \begin{pmatrix} 10.457 \\ -5.485\,7 \\ .514\,3 \end{pmatrix}$$

$$\mathbf{u}_6 = \begin{pmatrix} 1.0 \\ -.524\,6 \\ 4.918\,2 \times 10^{-2} \end{pmatrix}$$

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.524\,6 \\ 4.918\,2 \times 10^{-2} \end{pmatrix} = \begin{pmatrix} 12.885 \\ -6.295\,1 \\ -.295\,15 \end{pmatrix}$$

$$\mathbf{u}_7 = \begin{pmatrix} 1.0 \\ -.488\,56 \\ -2.290\,6 \times 10^{-2} \end{pmatrix}$$

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.488\,56 \\ -2.290\,6 \times 10^{-2} \end{pmatrix} = \begin{pmatrix} 11.588 \\ -5.862\,6 \\ .137\,36 \end{pmatrix}$$

$$\mathbf{u}_8 = \begin{pmatrix} 1.0 \\ -.505\,92 \\ 1.185\,4 \times 10^{-2} \end{pmatrix}$$

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.505\,92 \\ 1.185\,4 \times 10^{-2} \end{pmatrix} = \begin{pmatrix} 12.213 \\ -6.071\,1 \\ -7.108\,2 \times 10^{-2} \end{pmatrix}$$

$$\mathbf{u}_9 = \begin{pmatrix} 1.0 \\ -.497\,1 \\ -5.820\,2 \times 10^{-3} \end{pmatrix}$$

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.497\,1 \\ -5.820\,2 \times 10^{-3} \end{pmatrix} = \begin{pmatrix} 11.895 \\ -5.965\,1 \\ 3.486\,1 \times 10^{-2} \end{pmatrix}$$

$$\mathbf{u}_{10} = \begin{pmatrix} 1.0 \\ -.501\,48 \\ 2.930\,7 \times 10^{-3} \end{pmatrix}$$

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.501\,48 \\ 2.930\,7 \times 10^{-3} \end{pmatrix} = \begin{pmatrix} 12.053 \\ -6.017\,6 \\ -1.767\,2 \times 10^{-2} \end{pmatrix}$$

$$\mathbf{u}_{11} = \begin{pmatrix} 1.0 \\ -.499\,26 \\ -1.466\,2 \times 10^{-3} \end{pmatrix}$$

At this point, you could stop because the scaling factors are not changing by much. They went from 11.895 to 12.053. It looks like the eigenvalue is something like 12 which is in fact the case. The eigenvector is approximately $\mathbf{u}_{11}$. The true eigenvector for $\lambda = 12$ is

$$\begin{pmatrix} 1 \\ -.5 \\ 0 \end{pmatrix}$$

and so you see this is pretty close. If you didn't know this, observe

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.499\,26 \\ -1.466\,2 \times 10^{-3} \end{pmatrix} = \begin{pmatrix} 11.974 \\ -5.991\,2 \\ 8.838\,6 \times 10^{-3} \end{pmatrix} \tag{31.2}$$

and

$$12.053\begin{pmatrix} 1.0 \\ -.49926 \\ -1.4662\times10^{-3} \end{pmatrix} = \begin{pmatrix} 12.053 \\ -6.0176 \\ -1.7672\times10^{-2} \end{pmatrix}. \tag{31.3}$$

31.1.1 Rayleigh Quotient

In the above procedure, you can sometimes estimate the eigenvalue a little differently. If $A\mathbf{x} = \lambda\mathbf{x}$ then

$$\lambda = \frac{A\mathbf{x}\cdot\mathbf{x}}{|\mathbf{x}|^2}$$

and so, in the method above, you might get an estimate for the eigenvalue in this way. The above is called the Rayleigh quotient. In 31.2 where an approximate eigenvector has been found, you could estimate the eigenvalue as

$$\frac{\left(\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}\begin{pmatrix} 1.0 \\ -.49926 \\ -1.4662\times10^{-3} \end{pmatrix}\right)\cdot\begin{pmatrix} 1.0 \\ -.49926 \\ -1.4662\times10^{-3} \end{pmatrix}}{1+(-.49926)^2+(-1.4662\times10^{-3})^2}$$

$$= \quad 11.978788$$

The scaling factor was 12.053 and the Rayleigh quotient gave 11.978788. I guess that at least in this case the scaling factor wins. Lets look at a symmetric matrix. The book says the convergence of the Rayleigh quotients is about twice as fast as the scaling factors for symmetric matrices.

Example 31.1.2 *Use the Rayleigh quotient with the power method to estimate the dominant eigenvalue for the matrix,*

$$\begin{pmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 5 \end{pmatrix}$$

It turns out that the eigenvalues of this matrix are $3, \frac{7}{2}+\frac{1}{2}\sqrt{13}, \frac{7}{2}-\frac{1}{2}\sqrt{13}$. *In terms of decimals,* $3, 5.30277564, 1.69722436$, *and so the dominant eigenvalue is* 5.30277564.

Use the power method with an initial approximation $(1,1,1)^T$. Thus

$$\begin{pmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 5 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3.0 \\ 3.0 \\ 6.0 \end{pmatrix}$$

and so

$$\mathbf{u}_1 = \frac{1}{6}\begin{pmatrix} 3.0 \\ 3.0 \\ 6.0 \end{pmatrix} = \begin{pmatrix} .5 \\ .5 \\ 1.0 \end{pmatrix}$$

Next iteration,

$$\begin{pmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 5 \end{pmatrix}\begin{pmatrix} .5 \\ .5 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 2.0 \\ 1.5 \\ 5.5 \end{pmatrix}$$

Then

$$\mathbf{u}_2 = \frac{1}{5.5}\begin{pmatrix} 2.0 \\ 1.5 \\ 5.5 \end{pmatrix} = \begin{pmatrix} .363636364 \\ .272727273 \\ 1.0 \end{pmatrix}$$

The scaling factor, 5.5 is an approximation to the dominant eigenvalue, 5. 302 775 64. Lets see what is obtained from the Rayleigh quotient.

$$\frac{\left(\begin{pmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 5 \end{pmatrix}\begin{pmatrix} .363\,636\,364 \\ .272\,727\,273 \\ 1.0 \end{pmatrix}\right)\cdot\begin{pmatrix} .363\,636\,364 \\ .272\,727\,273 \\ 1.0 \end{pmatrix}}{(.363\,636\,364)^2+(.272\,727\,273)^2+1}$$

$$= 5.\,150\,684\,93$$

This is slightly better than the scaling factor, 5.5. Lets do another iteration. Lets see if we get a dramatic increase in accuracy.

$$\begin{pmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 5 \end{pmatrix}\begin{pmatrix} .363\,636\,364 \\ .272\,727\,273 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 1.\,727\,272\,73 \\ .818\,181\,819 \\ 5.\,363\,636\,36 \end{pmatrix}$$

Now

$$\mathbf{u}_3 = \frac{1}{5.\,363\,636\,36}\begin{pmatrix} 1.\,727\,272\,73 \\ .818\,181\,819 \\ 5.\,363\,636\,36 \end{pmatrix} = \begin{pmatrix} .322\,033\,899 \\ .152\,542\,373 \\ 1.\,0 \end{pmatrix}$$

The scaling factor, 5. 363 636 36 is an approximation to the dominant eigenvalue, 5. 302 775 64. Lets try the Rayleigh quotient again. This gives

$$\frac{\left(\begin{pmatrix} 2 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 5 \end{pmatrix}\begin{pmatrix} .322\,033\,899 \\ .152\,542\,373 \\ 1.\,0 \end{pmatrix}\right)\cdot\begin{pmatrix} .322\,033\,899 \\ .152\,542\,373 \\ 1.\,0 \end{pmatrix}}{(.322\,033\,899)^2+(.152\,542\,373)^2+1}$$

$$= \quad 5.\,254\,142\,24$$

The Rayleigh quotient is still just a little bit closer.

31.2 The Shifted Inverse Power Method

This method can find various eigenvalues and eigenvectors. It is a significant generalization of the above simple procedure and yields very good results. The situation is this: You have a number, α which is close to λ, some eigenvalue of an $n \times n$ matrix, A. You don't know λ but you know that α is closer to λ than to any other eigenvalue. Your problem is to find both λ and an eigenvector which goes with λ. Another way to look at this is to start with α and seek the eigenvalue, λ, which is closest to α along with an eigenvector associated with λ. If α is an eigenvalue of A, then you have what you want. Therefore, we will always assume α is not an eigenvalue of A and so $(A-\alpha I)^{-1}$ exists. The method is based on the following lemma. When using this method it is nice to choose α fairly close to an eigenvalue. Otherwise, the method will converge slowly. In order to get some idea where to start, you could use Gerschgorin's theorem but this theorem will only give a rough idea where to look. There isn't a really good way to know how to choose α for general cases. As we mentioned earlier, the eigenvalue problem is very difficult to solve in general.

Lemma 31.2.1 *Let $\{\lambda_k\}_{k=1}^n$ be the eigenvalues of A. If $\mathbf{x}_k$ is an eigenvector of A for the eigenvalue λ_k, then $\mathbf{x}_k$ is an eigenvector for $(A-\alpha I)^{-1}$ corresponding to the eigenvalue $\frac{1}{\lambda_k-\alpha}$.*

Proof: Let λ_k and $\mathbf{x}_k$ be as described in the statement of the lemma. Then

$$(A-\alpha I)\,\mathbf{x}_k = (\lambda_k-\alpha)\,\mathbf{x}_k$$

and so

$$\frac{1}{\lambda_k-\alpha}\mathbf{x}_k = (A-\alpha I)^{-1}\,\mathbf{x}_k.$$

This proves the lemma.

In explaining why the method works, we will assume A is nondefective. **This is not necessary!** This method is much better than it might seem from the explanation we are about to give. Pick $\mathbf{u}_1$, an initial vector and let $A\mathbf{x}_k = \lambda_k\mathbf{x}_k$, where $\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ is a basis of eigenvectors which exists from the assumption that A is nondefective. Assume α is closer to λ_n than to any other eigenvalue. Since A is nondefective, there exist constants, a_k such that

$$\mathbf{u}_1 = \sum_{k=1}^{n} a_k\mathbf{x}_k.$$

Possibly λ_n is a repeated eigenvalue. Then combining the terms in the sum which involve eigenvectors for λ_n, a simpler description of $\mathbf{u}_1$ is

$$\mathbf{u}_1 = \sum_{j=1}^{m} a_j\mathbf{x}_j + \mathbf{y}$$

where $\mathbf{y}$ is an eigenvector for λ_n which is assumed not equal to $\mathbf{0}$. (If you are unlucky in your choice for $\mathbf{u}_1$, this might not happen and things won't work.) Now the iteration procedure is defined as

$$\mathbf{u}_{k+1} \equiv \frac{(A-\alpha I)^{-1}\,\mathbf{u}_k}{S_k}$$

where S_k is the element of $(A-\alpha I)^{-1}\,\mathbf{u}_k$ which has largest absolute value. From Lemma 31.2.1,

$$\begin{aligned}\mathbf{u}_{k+1} &= \frac{\sum_{j=1}^{m} a_j\left(\frac{1}{\lambda_j-\alpha}\right)^k\mathbf{x}_j + \left(\frac{1}{\lambda_n-\alpha}\right)^k\mathbf{y}}{S_2\cdots S_k}\\ &= \frac{\left(\frac{1}{\lambda_n-\alpha}\right)^k}{S_2\cdots S_k}\left(\sum_{j=1}^{m} a_j\left(\frac{\lambda_n-\alpha}{\lambda_j-\alpha}\right)^k\mathbf{x}_j + \mathbf{y}\right).\end{aligned}$$

Now it is being assumed that λ_n is the eigenvalue which is closest to α and so for large k, the term,

$$\sum_{j=1}^{m} a_j\left(\frac{\lambda_n-\alpha}{\lambda_j-\alpha}\right)^k\mathbf{x}_j \equiv \mathbf{E}_k$$

is very small while for every $k\geq 1$, $\mathbf{u}_k$ is a moderate sized vector because every entry has absolute value less than or equal to 1. Thus

$$\mathbf{u}_{k+1} = \frac{\left(\frac{1}{\lambda_n-\alpha}\right)^k}{S_2\cdots S_k}(\mathbf{E}_k+\mathbf{y}) \equiv C_k(\mathbf{E}_k+\mathbf{y})$$

where $\mathbf{E}_k\to\mathbf{0}$, $\mathbf{y}$ is some eigenvector for λ_n, and C_k is of moderate size, remaining bounded as $k\to\infty$. Therefore, for large k,

$$\mathbf{u}_{k+1} - C_k\mathbf{y} = C_k\mathbf{E}_k \approx \mathbf{0}$$

and multiplying by $(A-\alpha I)^{-1}$ yields

$$\begin{aligned}(A-\alpha I)^{-1}\mathbf{u}_{k+1}-(A-\alpha I)^{-1}C_k\mathbf{y} &= (A-\alpha I)^{-1}\mathbf{u}_{k+1}-C_k\left(\frac{1}{\lambda_n-\alpha}\right)\mathbf{y}\\ &\approx (A-\alpha I)^{-1}\mathbf{u}_{k+1}-\left(\frac{1}{\lambda_n-\alpha}\right)\mathbf{u}_{k+1}\approx \mathbf{0}.\end{aligned}$$

Therefore, for large k, $\mathbf{u}_k$ is approximately equal to an eigenvector of $(A-\alpha I)^{-1}$. Therefore,

$$(A-\alpha I)^{-1}\mathbf{u}_k \approx \frac{1}{\lambda_n-\alpha}\mathbf{u}_k$$

and so you could take the dot product of both sides with $\mathbf{u}_k$ and approximate λ_n by solving the following for λ_n.

$$\frac{(A-\alpha I)^{-1}\mathbf{u}_k\cdot\mathbf{u}_k}{|\mathbf{u}_k|^2}=\frac{1}{\lambda_n-\alpha}$$

How else can you find the eigenvalue from this? Suppose $\mathbf{u}_k=(w_1,\cdots,w_n)^T$ and from the construction $|w_i|\le 1$ and $w_k=1$ for some k. Then

$$S_k\mathbf{u}_{k+1}=(A-\alpha I)^{-1}\mathbf{u}_k\approx(A-\alpha I)^{-1}(C_{k-1}\mathbf{y})=\frac{1}{\lambda_n-\alpha}(C_{k-1}\mathbf{y})\approx\frac{1}{\lambda_n-\alpha}\mathbf{u}_k.$$

Hence the entry of $(A-\alpha I)^{-1}\mathbf{u}_k$ which has largest absolute value is approximately $\frac{1}{\lambda_n-\alpha}$ and so it is likely that you can estimate λ_n using the formula

$$S_k=\frac{1}{\lambda_n-\alpha}.$$

Of course this would fail if $(A-\alpha I)^{-1}\mathbf{u}_k$ had more than one entry having equal absolute value.

Here is how you use the shifted inverse power method to find the eigenvalue and eigenvector closest to α.

1. Find $(A-\alpha I)^{-1}$.
2. Pick $\mathbf{u}_1$. It is important that $\mathbf{u}_1=\sum_{j=1}^m a_j\mathbf{x}_j+\mathbf{y}$ where $\mathbf{y}$ is an eigenvector which goes with the eigenvalue closest to α and the sum is in an "invariant subspace corresponding to the other eigenvalues". Of course you have no way of knowing whether this is so but it typically is so. If things don't work out, just start with a different $\mathbf{u}_1$. You were unlucky in your choice.
3. If $\mathbf{u}_k$ has been obtained,
$$\mathbf{u}_{k+1}=\frac{(A-\alpha I)^{-1}\mathbf{u}_k}{S_k}$$
where S_k is the element of $\mathbf{u}_k$ which has largest absolute value.
4. When the scaling factors, S_k are not changing much and the $\mathbf{u}_k$ are not changing much, find the approximation to the eigenvalue by solving
$$S_k=\frac{1}{\lambda-\alpha}$$
for λ. The eigenvector is approximated by $\mathbf{u}_{k+1}$.

5. Check your work by multiplying by the original matrix to see how well what you have found works.

Example 31.2.2 *Find the eigenvalue of* $A = \begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}$ *which is closest to* -7. *Also find an eigenvector which goes with this eigenvalue.*

In this case the eigenvalues are $-6, 0$, and 12 so the correct answer is -6 for the eigenvalue. Then from the above procedure, we will start with an initial vector,

$$\mathbf{u}_1 \equiv \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

We want the eigenvalue closest to -7. Thus we could use the above method. First we find

$$\left(\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} + 7\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)^{-1} = \begin{pmatrix} \frac{68}{133} & \frac{122}{133} & -\frac{65}{133} \\ \frac{4}{133} & \frac{15}{133} & \frac{4}{133} \\ -\frac{3}{7} & -\frac{6}{7} & \frac{4}{7} \end{pmatrix}$$

Then beginning with $\mathbf{u}_1$ above, the next iterate is

$$\mathbf{u}_2 = \begin{pmatrix} \frac{68}{133} & \frac{122}{133} & -\frac{65}{133} \\ \frac{4}{133} & \frac{15}{133} & \frac{4}{133} \\ -\frac{3}{7} & -\frac{6}{7} & \frac{4}{7} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} .939\,849\,624 \\ .172\,932\,331 \\ -.714\,285\,714 \end{pmatrix}$$

Thus $S_2 = .939\,849\,624$ and

$$\begin{pmatrix} .939\,849\,624 \\ .172\,932\,331 \\ -.714\,285\,714 \end{pmatrix}\frac{1}{.939\,849\,624} = \begin{pmatrix} 1.0 \\ .184 \\ -.76 \end{pmatrix}$$

Then doing another iteration,

$$\mathbf{u}_3 = \begin{pmatrix} \frac{68}{133} & \frac{122}{133} & -\frac{65}{133} \\ \frac{4}{133} & \frac{15}{133} & \frac{4}{133} \\ -\frac{3}{7} & -\frac{6}{7} & \frac{4}{7} \end{pmatrix}\begin{pmatrix} 1.0 \\ .184 \\ -.76 \end{pmatrix} = \begin{pmatrix} 1.051\,488\,72 \\ 2.796\,992\,48 \times 10^{-2} \\ -1.020\,571\,43 \end{pmatrix}$$

Dividing by the largest element, this yields

$$\frac{1}{1.051\,488\,72}\begin{pmatrix} 1.051\,488\,72 \\ 2.796\,992\,48 \times 10^{-2} \\ -1.020\,571\,43 \end{pmatrix} = \begin{pmatrix} 1.0 \\ 2.660\,030\,89 \times 10^{-2} \\ -.970\,596\,651 \end{pmatrix}$$

The next iteration is

$$\mathbf{u}_4 = \begin{pmatrix} \frac{68}{133} & \frac{122}{133} & -\frac{65}{133} \\ \frac{4}{133} & \frac{15}{133} & \frac{4}{133} \\ -\frac{3}{7} & -\frac{6}{7} & \frac{4}{7} \end{pmatrix}\begin{pmatrix} 1.0 \\ 2.660\,030\,89 \times 10^{-2} \\ -.970\,596\,651 \end{pmatrix} = \begin{pmatrix} 1.010\,030\,23 \\ 3.884\,346\,09 \times 10^{-3} \\ -1.005\,998\,35 \end{pmatrix}$$

The scaling factors are not changing by very much so this looks like a good time to stop. Thus you solve the following for λ.

$$\frac{1}{\lambda + 7} = 1.010\,030\,23.$$

This yields $\frac{1}{\lambda+7} = 1.010\,030\,23$ which yields $\lambda = -6.009\,930$. This is pretty close to the true eigenvalue, -6. How well does $\mathbf{u}_4$ work as an eigenvector?

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.010\,030\,23 \\ 3.884\,346\,09 \times 10^{-3} \\ -1.005\,998\,35 \end{pmatrix} = \begin{pmatrix} -6.070\,211\,55 \\ -5.901\,356\,4 \times 10^{-4} \\ 6.071\,391\,82 \end{pmatrix}$$

while

$$-6.009\,930 \begin{pmatrix} 1.010\,030\,23 \\ 3.884\,346\,09 \times 10^{-3} \\ -1.005\,998\,35 \end{pmatrix} = \begin{pmatrix} -6.070\,210\,98 \\ -2.334\,464\,81 \times 10^{-2} \\ 6.045\,979\,66 \end{pmatrix}.$$

Example 31.2.3 *Consider the symmetric matrix,* $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix}$. *Find the middle eigenvalue and an eigenvector which goes with it.*

Since A is symmetric, it follows it has three real eigenvalues which are solutions to

$$\begin{aligned} p(\lambda) &= \det\left(\lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix}\right) \\ &= \lambda^3 - 4\lambda^2 - 24\lambda - 17 = 0 \end{aligned}$$

If you use your graphing calculator to graph this polynomial, you find there is an eigenvalue somewhere between $-.9$ and $-.8$ and that this is the middle eigenvalue. Of course you could zoom in and find it very accurately without much trouble but what about the eigenvector which goes with it? If you try to solve

$$\left((-.8)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

there will be only the zero solution because the matrix on the left will be invertible and the same will be true if you replace $-.8$ with a better approximation like $-.86$ or $-.855$. This is because all these are only approximations to the eigenvalue and so the matrix in the above is nonsingular for all of these. Therefore, you will only get the zero solution and

Eigenvectors are never equal to zero!

However, there exists such an eigenvector and you can find it using the shifted inverse power method. Pick $\alpha = -.855$. You know this is close to the true eigenvalue. Then you find

$$\left(\left(\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} + .855\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)^{-1} = \begin{pmatrix} -367.501\,105 & 215.955\,47 & 83.601\,203\,4 \\ 215.955\,47 & -127.169\,104 & -48.753\,063\,2 \\ 83.601\,203\,4 & -48.753\,063\,2 & -19.191\,368\,6 \end{pmatrix}$$

The first step of the iteration is then

$$\mathbf{u}_1 = \begin{pmatrix} -367.501\,105 & 215.955\,47 & 83.601\,203\,4 \\ 215.955\,47 & -127.169\,104 & -48.753\,063\,2 \\ 83.601\,203\,4 & -48.753\,063\,2 & -19.191\,368\,6 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -67.944\,431\,6 \\ 40.033\,302\,8 \\ 15.656\,771\,6 \end{pmatrix}$$

Dividing by the largest entry to normalize the vector on the right,

$$\begin{pmatrix} -67.944\,431\,6 \\ 40.033\,302\,8 \\ 15.656\,771\,6 \end{pmatrix} \frac{1}{-67.944\,431\,6} = \begin{pmatrix} 1.0 \\ -.589\,206\,53 \\ -.230\,434\,949 \end{pmatrix}$$

Then the next approximation is

$$\begin{aligned} \mathbf{u}_2 &= \begin{pmatrix} -367.501\,105 & 215.955\,47 & 83.601\,203\,4 \\ 215.955\,47 & -127.169\,104 & -48.753\,063\,2 \\ 83.601\,203\,4 & -48.753\,063\,2 & -19.191\,368\,6 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.589\,206\,53 \\ -.230\,434\,949 \end{pmatrix} \\ &= \begin{pmatrix} -514.008\,117 \\ 302.118\,746 \\ 116.749\,189 \end{pmatrix} \end{aligned}$$

Divide this by the largest element.

$$\begin{pmatrix} -514.008\,117 \\ 302.118\,746 \\ 116.749\,189 \end{pmatrix} \frac{1}{-514.008\,117} = \begin{pmatrix} 1.0 \\ -.587\,770\,38 \\ -.227\,134\,913 \end{pmatrix}$$

Clearly these vectors are not changing much. An approximate eigenvector is then

$$\begin{pmatrix} 1.0 \\ -.587\,770\,38 \\ -.227\,134\,913 \end{pmatrix}$$

and to find the eigenvalue you solve $\frac{1}{\lambda+.855} = -514.008\,117$, which yields $\lambda = -.856\,945\,495$. How well does it work?

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.587\,770\,38 \\ -.227\,134\,913 \end{pmatrix} = \begin{pmatrix} -.856\,945\,499 \\ .503\,689\,968 \\ .194\,648\,654 \end{pmatrix}$$

while

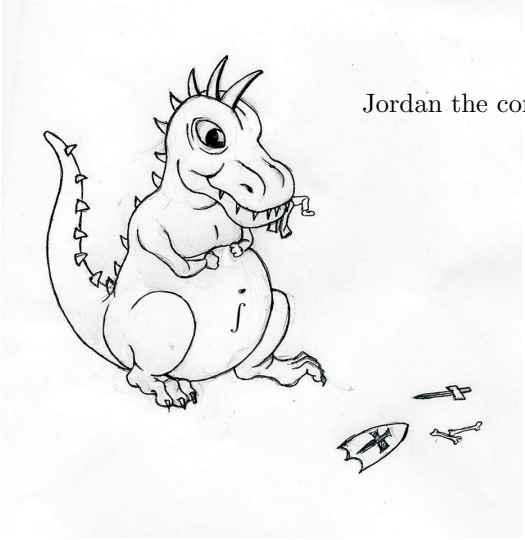
$$-.856\,945\,495 \begin{pmatrix} 1.0 \\ -.587\,770\,38 \\ -.227\,134\,913 \end{pmatrix} = \begin{pmatrix} -.856\,945\,495 \\ .503\,687\,179 \\ .194\,642\,24 \end{pmatrix}$$

I think you can see that for all practical purposes, this has found the eigenvalue and an eigenvector.

Part XV

The Correct Version Of The Riemann Integral *

The Theory Of The Riemann Integral**



Jordan the contented dragon

A.1 An Important Warning

If you read and understand this appendix on the Riemann integral you will become abnormal if you are not already that way. You will laugh at atrocious puns. You will be unpopular with well adjusted confident people. Furthermore, your confidence will be completely shattered. Virtually nothing will be obvious to you ever again. Consider whether it would be better to accept the superficial presentation given earlier than to attempt to acquire deep understanding of the integral, risking your self esteem and confidence, before proceeding further.

A.2 The Definition Of The Riemann Integral

The definition of the Riemann integral of a function of n variables uses the following definition.

Definition A.2.1 *For $i = 1, \cdots, n$, let $\{\alpha_k^i\}_{k=-\infty}^{\infty}$ be points on $\mathbb{R}$ which satisfy*

$$\lim_{k\to\infty} \alpha_k^i = \infty,\ \lim_{k\to-\infty} \alpha_k^i = -\infty,\ \alpha_k^i < \alpha_{k+1}^i. \tag{1.1}$$

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

$$Q = \prod_{i=1}^{n} \left[\alpha_{j_i}^i, \alpha_{j_i+1}^i\right]. \tag{1.2}$$

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 A.2.2 *If $\mathcal{G}$ and $\mathcal{F}$ are two grids, they have a common refinement, denoted here by $\mathcal{G} \vee \mathcal{F}$.*

Proof: Let $\{\alpha_k^i\}_{k=-\infty}^{\infty}$ be the sequences used to construct $\mathcal{G}$ and let $\{\beta_k^i\}_{k=-\infty}^{\infty}$ be the sequence used to construct $\mathcal{F}$. Now let $\{\gamma_k^i\}_{k=-\infty}^{\infty}$ denote the union of $\{\alpha_k^i\}_{k=-\infty}^{\infty}$ and $\{\beta_k^i\}_{k=-\infty}^{\infty}$. 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

$$\{\alpha_k^i\}_{k=-\infty}^{\infty} \cup \{\beta_k^i\}_{k=-\infty}^{\infty} \tag{1.3}$$

which is larger than γ_j^i and let $\gamma_{-(j+1)}^i$ be the last element of 1.3 which is strictly smaller than γ_{-j}^i. The assumption 1.1 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 off some bounded set, D. In what follows f will always be such a function.

Definition A.2.3 *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*

$$M_Q(f) \equiv \sup\{f(\mathbf{x}) : \mathbf{x} \in Q\},\ m_Q(f) \equiv \inf\{f(\mathbf{x}) : \mathbf{x} \in Q\}. \tag{1.4}$$

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 A.2.4 *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 A.2.2 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 A.2.5 *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 A.2.6 $\overline{I}(f) \geq \underline{I}(f)$.

Proof: From Lemma A.2.4 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 A.2.7 *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 A.2.8 *(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 A.2.4 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.

A.3 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 A.3.1 *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 $k = f, g$,

$$\mathcal{U}_{\mathcal{G}}(k) - \mathcal{L}_{\mathcal{G}}(k) < \delta^2. \tag{1.5}$$

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 $k = f, g$,

$$\varepsilon > \delta > \sum_{\mathcal{P}_k} v(Q). \tag{1.6}$$

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 1.6 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 A.3.2 *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,*

$$\int_{\mathbb{R}^n} (af + bg)\, dx = a \int_{\mathbb{R}^n} f\, dx + b \int_{\mathbb{R}^n} g\, dx, \tag{1.7}$$

and

$$\int |f|\, dx \geq \left| \int f\, dx \right|. \tag{1.8}$$

Proof: Each of the combinations of functions described above is Riemann integrable by Theorem A.3.1. 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 $af + bg$,

$$\mathcal{U}_{\mathcal{G}}(k) - \mathcal{L}_{\mathcal{G}}(k) < \varepsilon. \tag{1.9}$$

Consider 1.7. 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 1.7 and shows the integral is linear.

It remains to establish the inequality 1.8. By 1.9, 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 A.3.3 *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 A.3.4 *Let $\mathcal{G}$ be a grid and let f be a step function such that $f = f_Q$ on $\mathrm{int}(Q)$ 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, $|g(\mathbf{x})| \le C$, and $g = 0$ off Q, and $g = 1$ on $\mathrm{int}(Q)$. 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

$$B^k \equiv \left[\alpha^1_{j_1}, \alpha^1_{j_1+1}\right] \times \cdots \times \left[\alpha^{k-1}_{j_{k-1}}, \alpha^{k-1}_{j_{k-1}+1}\right] \times$$

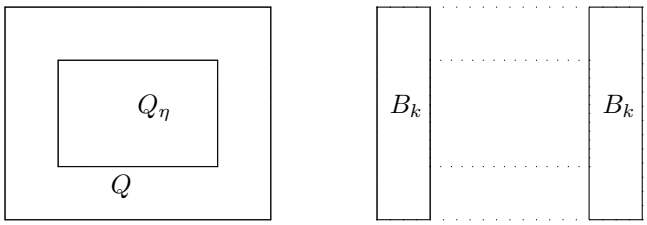
$$\left[\alpha^k_{j_k}, \alpha^k_{j_k} + \eta\right] \times \left[\alpha^{k+1}_{j_{k+1}}, \alpha^{k+1}_{j_{k+1}+1}\right] \times \cdots \times \left[\alpha^n_{j_n}, \alpha^n_{j_n+1}\right]$$

or

$$B^k \equiv \left[\alpha^1_{j_1}, \alpha^1_{j_1+1}\right] \times \cdots \times \left[\alpha^{k-1}_{j_{k-1}}, \alpha^{k-1}_{j_{k-1}+1}\right] \times$$

$$\left[\alpha^k_{j_k} - \eta, \alpha^k_{j_k}\right] \times \left[\alpha^{k+1}_{j_{k+1}}, \alpha^{k+1}_{j_{k+1}+1}\right] \times \cdots \times \left[\alpha^n_{j_n}, \alpha^n_{j_n+1}\right].$$

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,

$$\mathcal{L}_{\mathcal{F}}(g) \ge v(Q_\eta) - \sum_{k=1}^{n} 2Cv(B_k) \ge v(Q_\eta) - 4CL^{n-1}n\eta$$

$$= v(Q_\eta) - K\eta \tag{1.10}$$

where K is a constant which does not depend on η. Similarly,

$$\mathcal{U}_{\mathcal{F}}(g) \leq v(Q_\eta) + K\eta. \tag{1.11}$$

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 1.10 and 1.11 also verify that

$$\begin{aligned} v(Q_\eta) &\in [\mathcal{U}_{\mathcal{F}}(g) - K\eta, \mathcal{L}_{\mathcal{F}}(g) + K\eta] \\ &\subseteq [\mathcal{L}_{\mathcal{F}}(g) - K\eta, \mathcal{U}_{\mathcal{F}}(g) + K\eta]. \end{aligned}$$

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 $\operatorname{int}(Q)$, and let g_Q be a function of the sort just considered which equals 1 on $\operatorname{int}(Q)$. 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 A.3.2, $f \in \mathcal{R}(\mathbb{R}^n)$ and

$$\int f\,dx = \sum_{Q \in \mathcal{G}} f_Q \int g_Q\,dx = \sum_{Q \in \mathcal{G}} f_Q v(Q).$$

This proves the corollary.

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 A.3.5 *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*

$$\sum_{Q \in \mathcal{G}} \left(M_{Q'}(f) - m_{Q'}(f)\right) v(Q) < \varepsilon. \tag{1.12}$$

Proof: Since $Q' \subseteq Q$,

$$M_{Q'}(f) - m_{Q'}(f) \leq M_Q(f) - m_Q(f)$$

and therefore, the only if part of the equivalence is obvious.

Conversely, let $\mathcal{G}$ be a grid such that 1.12 holds with ε replaced with $\frac{\varepsilon}{2}$. It is necessary to show there is a grid such that 1.12 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 $\eta_k \leq \eta$ and is also chosen so small that for each $i = 1, \cdots, n$,

$$\alpha_k^i + \eta_k < \alpha_{k+1}^i.$$

You only need to have $\eta_k > 0$ for the finitely many boxes of $\mathcal{G}$ which intersect the bounded set where f is not zero. 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 1.12 valid because refinements do not increase the difference between upper and lower sums in this context either.) Suppose there are P sets of $\mathcal{G}$ contained in $B(\mathbf{0},R)$ (So these are the only sets of $\mathcal{G}$ which could have nonempty intersection with the set where f is nonzero.) and suppose that for all $\mathbf{x}$, $|f(\mathbf{x})| < C/2$. Then

$$\sum_{Q \in \mathcal{F}} \left(M_Q(f) - m_Q(f)\right) v(Q) \leq \sum_{\widehat{Q} \in \widehat{\mathcal{G}}} \left(M_{\widehat{Q}}(f) - m_{\widehat{Q}}(f)\right) v(Q)$$

$$+ \sum_{Q \in \mathcal{F} \setminus \widehat{\mathcal{G}}} \left(M_Q(f) - m_Q(f)\right) v(Q)$$

The first term on the right of the inequality in the above is no larger than $\varepsilon/2$ because $M_{\widehat{Q}}(f) - m_{\widehat{Q}}(f) \leq M_{Q'}(f) - m_{Q'}(f)$ for each Q. Therefore, the above is dominated by

$$\leq \varepsilon/2 + CPnR^{n-1}\eta < \varepsilon$$

whenever η is small enough. Since ε is arbitrary, $f \in \mathcal{R}(\mathbb{R}^n)$ as claimed.

Definition A.3.6 *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 A.3.7 *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 by the Riemann criterion.

Note that if E is a Jordan set and if $f \in \mathcal{R}(\mathbb{R}^n)$, then by Corollary A.3.2, $\mathcal{X}_E f \in \mathcal{R}(\mathbb{R}^n)$.

Definition A.3.8 *For E a Jordan set and $f\mathcal{X}_E \in \mathcal{R}(\mathbb{R}^n)$.*

$$\int_E f\,dV \equiv \int_{\mathbb{R}^n} \mathcal{X}_E f\,dV.$$

Also, 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 A.3.9 *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 A.3.10 *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

$$\sum_{Q \in \partial_{\mathcal{G}}(E)} v(Q) < \frac{\varepsilon}{2n+1}. \tag{1.13}$$

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,

$$\sum_{Q \in \mathcal{S}(P_i)} v(Q) \leq 2nv(P_i) + v(P_i),$$

the term $v(P_i)$ accounting for those boxes which are contained in P_i. Therefore, 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 1.13. This proves the corollary.

Theorem A.3.11 *If a bounded set, E, has Jordan content 0, then E is a Jordan (contented) 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 A.3.7 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

$$\begin{aligned} \mathcal{U}_{\mathcal{G}}(f\mathcal{X}_E) - \mathcal{L}_{\mathcal{G}}(f\mathcal{X}_E) &\leq \sum_{Q \cap E \neq \emptyset} Mv(Q) - \sum_{Q \cap E \neq \emptyset} mv(Q) \\ &= (M - m) \sum_{Q \cap E \neq \emptyset} v(Q) < \frac{\varepsilon (M - m)}{1 + (M - m)} < \varepsilon. \end{aligned}$$

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 A.3.12 *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: 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 A.3.11 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 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 A.3.13

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 A.3.14 *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 A.3.15 *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 A.3.16 *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 A.3.17 *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*

$$\sum_{Q \in \mathcal{G}_W} v(Q) < \varepsilon \tag{1.14}$$

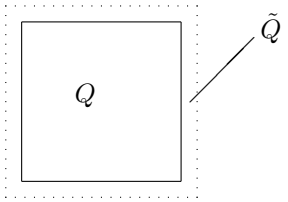
where

$$\mathcal{G}_W \equiv \{Q \in \mathcal{G} : Q \cap W \neq \emptyset\}.$$

Proof: Let W have content zero. Also 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 1.14. 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 1.14 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 $\frac{R}{3}$ and so all boxes of $\mathcal{G}$ which intersect the set where f is nonzero are contained in $B(\mathbf{0},R)$. 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

$$\omega_{f,r_{\mathbf{x}}}(\mathbf{x}) < \varepsilon + \omega_f(\mathbf{x}). \tag{1.15}$$

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 1.15 and the observation that $\mathbf{x} \notin W$,

$$M_Q(f) - m_Q(f) \leq \varepsilon.$$

Therefore,

$$\mathcal{U}_{\mathcal{F}}(f) - \mathcal{L}_{\mathcal{F}}(f) \leq \sum_{Q \in \mathcal{F}_W} Cv(Q) + \sum_{Q \in \mathcal{F}_1 \setminus \mathcal{F}_W} \varepsilon v(Q)$$

$$\leq C\varepsilon + \varepsilon (2R)^n.$$

Since ε is arbitrary, this proves the theorem.[1]

From Theorem A.3.7 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.

[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 $\varepsilon > 0$, 1.14 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. To prove the converse of this theorem would take us too far in that direction and it would not be reasonable to pay any more attention to this inferior integral.

Theorem A.3.18 *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 $k = f, g$,

$$\mathcal{U}_{\mathcal{G}}(k) - \mathcal{L}_{\mathcal{G}}(k) < \varepsilon/4. \tag{1.16}$$

Also let $K \geq \sum_{j=1}^{m} v_n(Q_j)$ for all $\mathbf{x} \in E$. 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

$$M_{Q_j}(f\mathcal{X}_E) + \frac{\varepsilon}{4mK}, M_{Q_j}(f\mathcal{X}_E), M_{Q_j}(g\mathcal{X}_E), m_{Q_j}(f\mathcal{X}_E), m_{Q_j}(g\mathcal{X}_E), m_{Q_j}(g\mathcal{X}_E) - \frac{\varepsilon}{4mK}$$

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 for $P \in \mathcal{G}'$, $\mathcal{X}_P \in \mathcal{R}(\mathbb{R}^{n+1})$. 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.

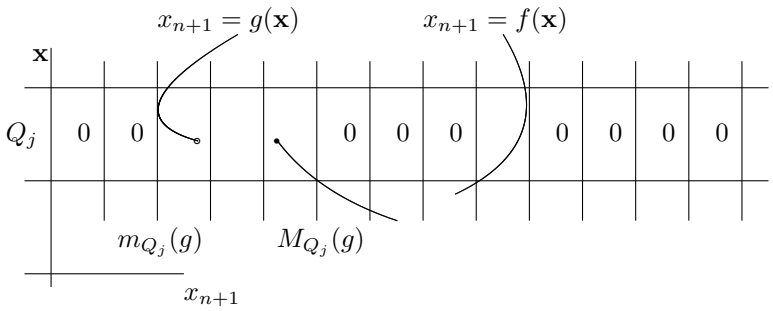
$$\begin{aligned}\mathcal{U}_{\mathcal{G}'}(\mathcal{X}_P) - \mathcal{L}_{\mathcal{G}'}(\mathcal{X}_P) &\equiv \sum_{Q' \in \mathcal{G}'} \left(M_{Q'}(\mathcal{X}_P) - m_{Q'}(\mathcal{X}_P)\right) v_{n+1}(Q') \\ &= \sum_{i=-\infty}^{\infty} \sum_{j=1}^{m} \left(M_{Q_j'}(\mathcal{X}_P) - m_{Q_j'}(\mathcal{X}_P)\right) v_n(Q_j)(a_{i+1} - a_i)\end{aligned}$$

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

$$\mathcal{U}_{\mathcal{G}'}(\mathcal{X}_P) - \mathcal{L}_{\mathcal{G}'}(\mathcal{X}_P) =$$

$$\sum_{j=1}^{m} v_n(Q_j) \sum_{i=-\infty}^{\infty} \left(M_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P) - m_{Q_j \times [a_i, a_{i+1}]}(\mathcal{X}_P)\right)(a_{i+1} - a_i).$$

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 $M_{Q_j}(f\mathcal{X}_E) + \frac{\varepsilon}{4mK}$, $M_{Q_j}(f\mathcal{X}_E), M_{Q_j}(g\mathcal{X}_E), m_{Q_j}(f\mathcal{X}_E), m_{Q_j}(g\mathcal{X}_E)$ for all $j = 1, \cdots, m$,

$$\sum_{i=-\infty}^{\infty} \left(M_{Q_j\times[a_i,a_{i+1}]}(\mathcal{X}_P) - m_{Q_j\times[a_i,a_{i+1}]}(\mathcal{X}_P)\right)(a_{i+1} - a_i) \leq$$

$$\sum_{\{i:m_{Q_j}(g)\leq a_i < M_{Q_j}(g)\}} 1\,(a_{i+1} - a_i) + \sum_{\{i:m_{Q_j}(f)\leq a_i < M_{Q_j}(f)\}} 1\,(a_{i+1} - a_i)$$

$$\left(M_{Q_j}(f\mathcal{X}_E) + \frac{\varepsilon}{4mK} - M_{Q_j}(f\mathcal{X}_E)\right) + \left(m_{Q_j}(g\mathcal{X}_E) - \left(m_{Q_j}(g\mathcal{X}_E) - \frac{\varepsilon}{4mK}\right)\right)$$

$$= \left(M_{Q_j}(g\mathcal{X}_E) - m_{Q_j}(g\mathcal{X}_E)\right) + \left(M_{Q_j}(f\mathcal{X}_E) - m_{Q_j}(f\mathcal{X}_E)\right) + \frac{\varepsilon}{2m}\left(\sum_{j=1}^{m} v(Q_j)\right)^{-1}.$$

(Note the inequality.) The last two terms which add to $\frac{\varepsilon}{2m}\left(\sum_{j=1}^{m} v(Q_j)\right)^{-1}$ come from the case where $a_i = M_{Q_j}(f)$ or $a_{i+1} = m_{Q_j}(f)$. Therefore, by 1.16,

$$\mathcal{U}_{\mathcal{G}'}(\mathcal{X}_P) - \mathcal{L}_{\mathcal{G}'}(\mathcal{X}_P) \leq$$

$$\begin{aligned}
&\sum_{j=1}^{m} v_n(Q_j)\left[\left(M_{Q_j}(g\mathcal{X}_E) - m_{Q_j}(g\mathcal{X}_E)\right) + \left(M_{Q_j}(f\mathcal{X}_E) - m_{Q_j}(f\mathcal{X}_E)\right)\right] \\
&+\sum_{j=1}^{m} v(Q_j)\frac{\varepsilon}{2m}\left(\sum_{j=1}^{m} v(Q_j)\right)^{-1} \\
=\ &\mathcal{U}_{\mathcal{G}}(f) - \mathcal{L}_{\mathcal{G}}(f) + \mathcal{U}_{\mathcal{G}}(g) - \mathcal{L}_{\mathcal{G}}(g) + \frac{\varepsilon}{2} \\
<\ &\frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, this proves the theorem.

Corollary A.3.19 *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 A.3.10 on Page 574 implies this is a set of content 0. Therefore, from Theorem A.3.17, 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 A.3.18. 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 A.3.4 on Page 571, it follows that every box is a contented set.

A.4 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 A.4.1 *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 A.4.2 *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 $\mathcal{G}$. This proves the lemma.

Now let $\mathcal{G}$ be a grid on $\mathbb{R}^{n+m}$ and suppose

$$\phi(\mathbf{z}) = \sum_{Q \in \mathcal{G}} \phi_Q \mathcal{X}_{Q'}(\mathbf{z}) \tag{1.17}$$

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 A.4.2 implies

$$\begin{aligned} \phi(\mathbf{z}) = \phi(\mathbf{x}, \mathbf{y}) &= \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \mathcal{X}_{R' \times P'}(\mathbf{x}, \mathbf{y}) \\ &= \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \mathcal{X}_{R'}(\mathbf{x}) \mathcal{X}_{P'}(\mathbf{y}). \end{aligned} \tag{1.18}$$

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 A.4.3 *Let ϕ be a step function as described in 1.17. Then*

$$\phi(\mathbf{x}, \cdot) \in \mathcal{R}(\mathbb{R}^m), \tag{1.19}$$

$$\int_{\mathbb{R}^m} \phi(\cdot, \mathbf{y}) \, dV_y \in \mathcal{R}(\mathbb{R}^n), \tag{1.20}$$

and

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^m} \phi(\mathbf{x}, \mathbf{y}) \, dV_y \, dV_x = \int_{\mathbb{R}^{n+m}} \phi(\mathbf{z}) \, dV. \tag{1.21}$$

Proof: To verify 1.19, 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 A.3.4, this verifies 1.19. From the description in 1.18 and this corollary,

$$\begin{aligned} \int_{\mathbb{R}^m} \phi(\mathbf{x}, \mathbf{y}) \, dV_y &= \sum_{R \in \mathcal{G}_n} \sum_{P \in \mathcal{G}_m} \phi_{R \times P} \mathcal{X}_{R'}(\mathbf{x}) v(P) \\ &= \sum_{R \in \mathcal{G}_n} \left(\sum_{P \in \mathcal{G}_m} \phi_{R \times P} v(P) \right) \mathcal{X}_{R'}(\mathbf{x}), \end{aligned} \tag{1.22}$$

another step function. Therefore, Corollary A.3.4 applies again to verify 1.20. Finally, 1.22 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 1.22,

$$M_{R_1'}\left(\int_{\mathbb{R}^m} \phi(\cdot, \mathbf{y})\, dV_y\right) \equiv \sup\left\{\sum_{R \in \mathcal{G}_n}\left(\sum_{P \in \mathcal{G}_m} \phi_{R \times P} v(P)\right) \mathcal{X}_{R'}(\mathbf{x}) : \mathbf{x} \in R_1'\right\}$$

$$= \sum_{P \in \mathcal{G}_m} \phi_{R_1 \times P} v(P) \tag{1.23}$$

because $\int_{\mathbb{R}^m} \phi(\cdot, \mathbf{y})\, dV_y$ has the constant value given in 1.23 for $\mathbf{x} \in R_1'$. Similarly,

$$m_{R_1'}\left(\int_{\mathbb{R}^m} \phi(\cdot, \mathbf{y})\, dV_y\right) \equiv \inf\left\{\sum_{R \in \mathcal{G}_n}\left(\sum_{P \in \mathcal{G}_m} \phi_{R \times P} v(P)\right) \mathcal{X}_{R'}(\mathbf{x}) : \mathbf{x} \in R_1'\right\}$$

$$= \sum_{P \in \mathcal{G}_m} \phi_{R_1 \times P} v(P). \tag{1.24}$$

Theorem A.4.4 *(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*

$$\int_{\mathbb{R}^m} f(\cdot, \mathbf{y})\, dV_y \in \mathcal{R}(\mathbb{R}^n) \tag{1.25}$$

and

$$\int_{\mathbb{R}^{n+m}} f(\mathbf{z})\, dV = \int_{\mathbb{R}^n}\int_{\mathbb{R}^m} f(\mathbf{x}, \mathbf{y})\, dV_y\, dV_x. \tag{1.26}$$

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 A.3.4, 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 1.23,

$$M_{R'}\left(\int_{\mathbb{R}^m} f(\cdot, \mathbf{y})\, dV_y\right) \le M_{R'}\left(\int_{\mathbb{R}^m} \phi(\cdot, \mathbf{y})\, dV_y\right) = \sum_{P \in \mathcal{G}_m} M_{R' \times P'}(f)\, v(P)$$

and from 1.24,

$$m_{R'}\left(\int_{\mathbb{R}^m} f(\cdot, \mathbf{y})\, dV_y\right) \ge m_{R'}\left(\int_{\mathbb{R}^m} \psi(\cdot, \mathbf{y})\, dV_y\right) = \sum_{P \in \mathcal{G}_m} m_{R' \times P'}(f)\, v(P).$$

Therefore,

$$\sum_{R \in \mathcal{G}_n}\left[M_{R'}\left(\int_{\mathbb{R}^m} f(\cdot, \mathbf{y})\, dV_y\right) - m_{R'}\left(\int_{\mathbb{R}^m} f(\cdot, \mathbf{y})\, dV_y\right)\right] v(R) \le$$

$$\sum_{R\in\mathcal{G}_n}\sum_{P\in\mathcal{G}_m}\left[M_{R'\times P'}(f)-m_{R'\times P'}(f)\right]v(P)\,v(R)\leq\mathcal{U}_{\mathcal{G}}(f)-\mathcal{L}_{\mathcal{G}}(f)<\varepsilon.$$

This shows, from Lemma A.3.5 and the Riemann criterion, that $\int_{\mathbb{R}^m} f(\cdot,\mathbf{y})\,dV_y\in\mathcal{R}(\mathbb{R}^n)$. It remains to verify 1.26. 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 A.4.3,

$$\mathcal{L}_{\mathcal{G}}(f)\leq\int_{\mathbb{R}^{n+m}}\psi\,dV=\int_{\mathbb{R}^n}\int_{\mathbb{R}^m}\psi\,dV_y\,dV_x\leq\int_{\mathbb{R}^n}\int_{\mathbb{R}^m}f(\mathbf{x},\mathbf{y})\,dV_y\,dV_x$$

$$\leq\int_{\mathbb{R}^n}\int_{\mathbb{R}^m}\phi(\mathbf{x},\mathbf{y})\,dV_y\,dV_x=\int_{\mathbb{R}^{n+m}}\phi\,dV\leq\mathcal{U}_{\mathcal{G}}(f).$$

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 A.4.5 *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 $f\mathcal{X}_P\in\mathcal{R}\left(\mathbb{R}^{n+1}\right)$ 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 $f(\mathbf{x},\cdot)\,\mathcal{X}_{[\psi(\mathbf{x}),\phi(\mathbf{x})]}(\cdot)\in\mathcal{R}\left(\mathbb{R}^1\right)$. Also, $f\mathcal{X}_P\in\mathcal{R}\left(\mathbb{R}^{n+1}\right)$ because P is contented thanks to Corollary A.3.19. 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.

A.5 The Change Of Variables Formula

First recall Theorem 22.2.2 on Page 394 which is listed here for convenience.

[2]Actually, Fubini's theorem usually refers to a much more profound result in the theory of Lebesgue integration.

Theorem A.5.1 *Let $\mathbf{h}: U \to \mathbb{R}^n$ be a C^1 function with $\mathbf{h}(\mathbf{0}) = \mathbf{0}$,$D\mathbf{h}(\mathbf{0})^{-1}$ exists. Then there exists an open set, $V \subseteq U$ containing $\mathbf{0}$, flips, $\mathbf{F}_1, \cdots, \mathbf{F}_{n-1}$, and primitive functions, $\mathbf{G}_n, \mathbf{G}_{n-1}, \cdots, \mathbf{G}_1$ such that for $\mathbf{x} \in V$,*

$$\mathbf{h}(\mathbf{x}) = \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{G}_n \circ \mathbf{G}_{n-1} \circ \cdots \circ \mathbf{G}_1(\mathbf{x}).$$

Also recall Theorem 16.2.13 on Page 295.

Theorem A.5.2 *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 $[a,b]$,*

$$\int_c^d f(s)\, ds = \int_a^b f(\phi(t)) \left|\phi'(t)\right| dt$$

The following is a simple corollary to this theorem.

Corollary A.5.3 *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 $[a,b]$,*

$$\int_{\mathbb{R}} \mathcal{X}_{[a,b]}\left(\phi^{-1}(x)\right) f(x)\, dx = \int_{\mathbb{R}} \mathcal{X}_{[a,b]}(t)\, f(\phi(t)) \left|\phi'(t)\right| dt$$

Lemma A.5.4 *Let $\mathbf{h}: V \to \mathbb{R}^n$ be a C^1 function and suppose H is a compact subset of V. Then there exists a constant, C independent of $\mathbf{x} \in H$ such that*

$$|D\mathbf{h}(\mathbf{x})\,\mathbf{v}| \leq C\,|\mathbf{v}|.$$

Proof: Consider the compact set, $H \times \partial B(\mathbf{0},1) \subseteq \mathbb{R}^{2n}$. Let $f: H \times \partial B(\mathbf{0},1) \to \mathbb{R}$ be given by $f(\mathbf{x},\mathbf{v}) = |D\mathbf{h}(\mathbf{x})\,\mathbf{v}|$. Then let C denote the maximum value of f. It follows that for $\mathbf{v} \in \mathbb{R}^n$,

$$\left|D\mathbf{h}(\mathbf{x}) \frac{\mathbf{v}}{|\mathbf{v}|}\right| \leq C$$

and so the desired formula follows when you multiply both sides by $|\mathbf{v}|$.

Definition A.5.5 *Let A be an open set. Write $C^k(A;\mathbb{R}^n)$ to denote a C^k function whose domain is A and whose range is in $\mathbb{R}^n$. Let U be an open set in $\mathbb{R}^n$. Then $\mathbf{h} \in C^k\left(\overline{U};\mathbb{R}^n\right)$ if there exists an open set, $V \supseteq \overline{U}$ and a function, $\mathbf{g} \in C^1(V;\mathbb{R}^n)$ such that $\mathbf{g} = \mathbf{h}$ on $\overline{U}$. $f \in C^k\left(\overline{U}\right)$ means the same thing except that f has values in $\mathbb{R}$.*

Theorem A.5.6 *Let U be a bounded open set such that ∂U has zero content and let $\mathbf{h} \in C\left(\overline{U};\mathbb{R}^n\right)$ be one to one and $D\mathbf{h}(\mathbf{x})^{-1}$ exists for all $\mathbf{x} \in U$. Then $\mathbf{h}(\partial U) = \partial(\mathbf{h}(U))$ and $\partial(\mathbf{h}(U))$ has zero content.*

Proof: Let $\mathbf{x} \in \partial U$ and let $\mathbf{g} = \mathbf{h}$ where $\mathbf{g}$ is a C^1 function defined on an open set containing $\overline{U}$. By the inverse function theorem, $\mathbf{g}$ is locally one to one and an open mapping near $\mathbf{x}$. Thus $\mathbf{g}(\mathbf{x}) = \mathbf{h}(\mathbf{x})$ and is in an open set containing points of $\mathbf{g}(U)$ and points of $\mathbf{g}\left(U^C\right)$. These points of $\mathbf{g}\left(U^C\right)$ cannot equal any points of $\mathbf{h}(U)$ because $\mathbf{g}$ is one to one locally. Thus $\mathbf{h}(\mathbf{x}) \in \partial(\mathbf{h}(U))$ and so $\mathbf{h}(\partial U) \subseteq \partial(\mathbf{h}(U))$. Now suppose $\mathbf{y} \in \partial(\mathbf{h}(U))$. By the inverse function theorem $\mathbf{y}$ cannot be in the open set $\mathbf{h}(U)$. Since $\mathbf{y} \in \partial(\mathbf{h}(U))$, every ball centered at $\mathbf{y}$ contains points of $\mathbf{h}(U)$ and so $\mathbf{y} \in \overline{\mathbf{h}(U)} \setminus \mathbf{h}(U)$. Thus there exists a sequence, $\{\mathbf{x}_n\} \subseteq U$ such that $\mathbf{h}(\mathbf{x}_n) \to \mathbf{y}$. But then, by the inverse function theorem, $\mathbf{x}_n \to \mathbf{h}^{-1}(\mathbf{y})$ and so $\mathbf{h}^{-1}(\mathbf{y}) \in \partial U$. Therefore, $\mathbf{y} \in \mathbf{h}(\partial U)$ and this proves the two sets are equal. It remains to verify the claim about content.

First let H denote a compact set whose interior contains $\overline{U}$ which is also in the interior of the domain of $\mathbf{g}$. Now since ∂U has content zero, it follows that for $\varepsilon > 0$ given, there exists a grid, $\mathcal{G}$ such that if $\mathcal{G}'$ are those boxes of $\mathcal{G}$ which have nonempty intersection with ∂U, then

$$\sum_{Q \in \mathcal{G}'} v(Q) < \varepsilon.$$

and by refining the grid if necessary, no box of $\mathcal{G}$ has nonempty intersection with both $\overline{U}$ and H^C. Refining this grid still more, you can also assume that for all boxes in $\mathcal{G}'$,

$$\frac{l_i}{l_j} < 2$$

where l_i is the length of the i^{th} side. (Thus the boxes are not too far from being cubes.)

Let C be the constant of Lemma A.5.4 applied to $\mathbf{g}$ on H.

Now consider one of these boxes, $Q \in \mathcal{G}'$. If $\mathbf{x}, \mathbf{y} \in Q$, it follows from the chain rule that

$$\mathbf{g}(\mathbf{y}) - \mathbf{g}(\mathbf{x}) = \int_0^1 D\mathbf{g}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))(\mathbf{y}-\mathbf{x})\,dt$$

By Lemma A.5.4 applied to H

$$\begin{aligned} |\mathbf{g}(\mathbf{y}) - \mathbf{g}(\mathbf{x})| &\leq \int_0^1 |D\mathbf{g}(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))(\mathbf{y}-\mathbf{x})|\,dt \\ &\leq C\int_0^1 |\mathbf{x}-\mathbf{y}|\,dt \leq C\,\mathrm{diam}(Q) \\ &= C\left(\sum_{i=1}^n l_i^2\right)^{1/2} \leq C\sqrt{n}L \end{aligned}$$

where L is the length of the longest side of Q. Thus $\mathrm{diam}(\mathbf{g}(Q)) \leq C\sqrt{n}L$ and so $\mathbf{g}(Q)$ is contained in a cube having sides equal to $C\sqrt{n}L$ and volume equal to

$$C^n n^{n/2} L^n \leq C^n n^{n/2} 2^n l_1 l_2 \cdots l_n = C^n n^{n/2} 2^n v(Q).$$

Denoting by P_Q this cube, it follows

$$\mathbf{h}(\partial U) \subseteq \cup_{Q \in \mathcal{G}'} v(P_Q)$$

and

$$\sum_{Q \in \mathcal{G}'} v(P_Q) \leq C^n n^{n/2} 2^n \sum_{Q \in \mathcal{G}'} v(Q) < \varepsilon C^n n^{n/2} 2^n.$$

Since $\varepsilon > 0$ is arbitrary, this shows $\mathbf{h}(\partial U)$ has content zero as claimed.

Theorem A.5.7 *Suppose $f \in C(\overline{U})$ where U is a bounded open set with ∂U having content 0. Then $f\mathcal{X}_U \in \mathcal{R}(\mathbb{R}^n)$.*

Proof: Let H be a compact set whose interior contains $\overline{U}$ which is also contained in the domain of g where g is a continuous functions whose restriction to U equals f. Consider $g\mathcal{X}_U$, a function whose set of discontinuities has content 0. Then $g\mathcal{X}_U = f\mathcal{X}_U \in \mathcal{R}(\mathbb{R}^n)$ as claimed. This is by the big theorem which tells which functions are Riemann integrable.

The following lemma is obvious from the definition of the integral.

Lemma A.5.8 *Let U be a bounded open set and let $f\mathcal{X}_U \in \mathcal{R}(\mathbb{R}^n)$. Then*

$$\int f(\mathbf{x}+\mathbf{p})\,\mathcal{X}_{U-\mathbf{p}}(\mathbf{x})\,dx = \int f(\mathbf{x})\,\mathcal{X}_U(\mathbf{x})\,dx$$

A few more lemmas are needed.

Lemma A.5.9 *Let S be a nonempty subset of $\mathbb{R}^n$. Define*

$$f(\mathbf{x}) \equiv \operatorname{dist}(\mathbf{x},S) \equiv \inf\{|\mathbf{x}-\mathbf{y}| : \mathbf{y}\in S\}.$$

Then f is continuous.

Proof: Consider $|f(\mathbf{x})-f(\mathbf{x}_1)|$and suppose without loss of generality that $f(\mathbf{x}_1)\geq f(\mathbf{x})$. Then choose $\mathbf{y}\in S$ such that $f(\mathbf{x})+\varepsilon > |\mathbf{x}-\mathbf{y}|$. Then

$$\begin{aligned} |f(\mathbf{x}_1)-f(\mathbf{x})| &= f(\mathbf{x}_1)-f(\mathbf{x}) \leq f(\mathbf{x}_1)-|\mathbf{x}-\mathbf{y}|+\varepsilon \\ &\leq |\mathbf{x}_1-\mathbf{y}|-|\mathbf{x}-\mathbf{y}|+\varepsilon \\ &\leq |\mathbf{x}-\mathbf{x}_1|+|\mathbf{x}-\mathbf{y}|-|\mathbf{x}-\mathbf{y}|+\varepsilon \\ &= |\mathbf{x}-\mathbf{x}_1|+\varepsilon. \end{aligned}$$

Since ε is arbitrary, it follows that $|f(\mathbf{x}_1)-f(\mathbf{x})|\leq|\mathbf{x}-\mathbf{x}_1|$ and this proves the lemma.

Theorem A.5.10 *(Urysohn's lemma for $\mathbb{R}^n$) Let H be a closed subset of an open set, U. Then there exists a continuous function, $g:\mathbb{R}^n\to[0,1]$ such that $g(\mathbf{x})=1$ for all $\mathbf{x}\in H$ and $g(\mathbf{x})=0$ for all $\mathbf{x}\notin U$.*

Proof: If $\mathbf{x}\notin C$, a closed set, then $\operatorname{dist}(\mathbf{x},C)>0$ because if not, there would exist a sequence of points of C converging to $\mathbf{x}$ and it would follow that $\mathbf{x}\in C$. Therefore, $\operatorname{dist}(\mathbf{x},H)+\operatorname{dist}\left(\mathbf{x},U^C\right)>0$ for all $\mathbf{x}\in\mathbb{R}^n$. Now define a continuous function, g as

$$g(\mathbf{x}) \equiv \frac{\operatorname{dist}\left(\mathbf{x},U^C\right)}{\operatorname{dist}(\mathbf{x},H)+\operatorname{dist}(\mathbf{x},U^C)}.$$

It is easy to see this verifies the conclusions of the theorem and this proves the theorem.

Definition A.5.11 *Define $\operatorname{spt}(f)$ (support of f) to be the closure of the set $\{x : f(x)\neq 0\}$. If V is an open set, $C_c(V)$ will be the set of continuous functions f, defined on $\mathbb{R}^n$ having $\operatorname{spt}(f)\subseteq V$.*

Definition A.5.12 *If K is a compact subset of an open set, V, then $K\prec\phi\prec V$ if*

$$\phi\in C_c(V),\ \phi(K)=\{1\},\ \phi(\mathbb{R}^n)\subseteq[0,1].$$

Also for $\phi\in C_c(\mathbb{R}^n)$, $K\prec\phi$ if

$$\phi(\mathbb{R}^n)\subseteq[0,1] \text{ and } \phi(K)=1.$$

and $\phi\prec V$ if

$$\phi(\mathbb{R}^n)\subseteq[0,1] \text{ and } \operatorname{spt}(\phi)\subseteq V.$$

Theorem A.5.13 *(Partition of unity) Let K be a compact subset of $\mathbb{R}^n$ and suppose*

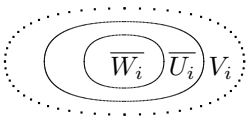
$$K \subseteq V = \cup_{i=1}^n V_i,\ V_i \text{ open.}$$

Then there exist $\psi_i \prec V_i$ with

$$\sum_{i=1}^n \psi_i(\mathbf{x}) = 1$$

for all $\mathbf{x} \in K$.

Proof: Let $K_1 = K \backslash \cup_{i=2}^n V_i$. Thus K_1 is compact because it is the intersection of a closed set with a compact set and $K_1 \subseteq V_1$. Let $K_1 \subseteq W_1 \subseteq \overline{W}_1 \subseteq V_1$ with $\overline{W}_1$compact. To obtain W_1, use Theorem A.5.10 to get f such that $K_1 \prec f \prec V_1$ and let $W_1 \equiv \{\mathbf{x} : f(\mathbf{x}) \neq 0\}$. Thus $W_1, V_2, \cdots V_n$ covers K and $\overline{W}_1 \subseteq V_1$. Let $K_2 = K \setminus (\cup_{i=3}^n V_i \cup W_1)$. Then K_2 is compact and $K_2 \subseteq V_2$. Let $K_2 \subseteq W_2 \subseteq \overline{W}_2 \subseteq V_2$ $\overline{W}_2$ compact. Continue this way finally obtaining $W_1, \cdots, W_n$, $K \subseteq W_1 \cup \cdots \cup W_n$, and $\overline{W}_i \subseteq V_i$ $\overline{W}_i$ compact. Now let $\overline{W}_i \subseteq U_i \subseteq \overline{U}_i \subseteq V_i$, $\overline{U}_i$ compact.



By Theorem A.5.10, there exist functions, ϕ_i, γ such that $\overline{U}_i \prec \phi_i \prec V_i$, $\cup_{i=1}^n \overline{W}_i \prec \gamma \prec \cup_{i=1}^n U_i$. Define

$$\psi_i(\mathbf{x}) = \begin{cases} \gamma(\mathbf{x})\phi_i(\mathbf{x}) / \sum_{j=1}^n \phi_j(\mathbf{x}) \text{ if } \sum_{j=1}^n \phi_j(\mathbf{x}) \neq 0, \\ 0 \text{ if } \sum_{j=1}^n \phi_j(\mathbf{x}) = 0. \end{cases}$$

If $\mathbf{x}$ is such that $\sum_{j=1}^n \phi_j(\mathbf{x}) = 0$, then $\mathbf{x} \notin \cup_{i=1}^n \overline{U}_i$. Consequently $\gamma(\mathbf{y}) = 0$ for all $\mathbf{y}$ near $\mathbf{x}$ and so $\psi_i(\mathbf{y}) = 0$ for all $\mathbf{y}$ near $\mathbf{x}$. Hence ψ_i is continuous at such $\mathbf{x}$. If $\sum_{j=1}^n \phi_j(\mathbf{x}) \neq 0$, this situation persists near $\mathbf{x}$ and so ψ_i is continuous at such points. Therefore ψ_i is continuous. If $\mathbf{x} \in K$, then $\gamma(\mathbf{x}) = 1$ and so $\sum_{j=1}^n \psi_j(\mathbf{x}) = 1$. Clearly $0 \leq \psi_i(\mathbf{x}) \leq 1$ and $\text{spt}(\psi_j) \subseteq V_j$. This proves the theorem.

The next lemma contains the main ideas.

Lemma A.5.14 *Let U be a bounded open set with ∂U having content 0. Also let $\mathbf{h} \in C^1(\overline{U}; \mathbb{R}^n)$ be one to one on U with $D\mathbf{h}(\mathbf{x})^{-1}$ exists for all $\mathbf{x} \in U$. Let $f \in C(\overline{U})$ be nonnegative. Then*

$$\int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f(\mathbf{z}) \, dV_n = \int \mathcal{X}_U(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| \, dV_n$$

Proof: Let $\varepsilon > 0$ be given. Then by Theorem A.5.7,

$$\mathbf{x} \to \mathcal{X}_U(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})|$$

is Riemann integrable. Therefore, there exists a grid, $\mathcal{G}$ such that, letting

$$g(\mathbf{x}) = \mathcal{X}_U(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})|,$$

$$\mathcal{L}_{\mathcal{G}}(g) + \varepsilon > \mathcal{U}_{\mathcal{G}}(g).$$

Let K denote the union of the boxes, Q of $\mathcal{G}$ for which $m_Q(g) > 0$. Thus K is a compact subset of U and it is only the terms from these boxes which contribute anything nonzero to the lower sum. By Theorem 22.2.2 on Page 394 which is stated above, it follows that for $\mathbf{p} \in K$, there exists an open set contained in U which contains $\mathbf{p}$,$O_{\mathbf{p}}$ such that for $\mathbf{x} \in O_{\mathbf{p}} - \mathbf{p}$,

$$\mathbf{h}(\mathbf{x}+\mathbf{p}) - \mathbf{h}(\mathbf{p}) = \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{G}_n \circ \cdots \circ \mathbf{G}_1(\mathbf{x})$$

where the $\mathbf{G}_i$ are primitive functions and the $\mathbf{F}_j$ are flips. Finitely many of these open sets, $\{O_j\}_{j=1}^q$ cover K. Let the distinguished point for O_j be denoted by $\mathbf{p}_j$. Now refine $\mathcal{G}$ if necessary such that the diameter of every cell of the new $\mathcal{G}$ which intersects U is smaller than a Lebesgue number for this open cover. Denote by $\mathcal{G}'$ those boxes of $\mathcal{G}$ whose union equals the set, K. Thus every box of $\mathcal{G}'$ is contained in one of these O_j. By Theorem A.5.13 there exists a partition of unity, $\{\psi_j\}$ on $\mathbf{h}(K)$ such that $\psi_j \prec \mathbf{h}(O_j)$. Then

$$\begin{aligned} \mathcal{L}_{\mathcal{G}}(g) &\leq \sum_{Q\in\mathcal{G}'} \int \mathcal{X}_Q(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| \, dx \\ &= \sum_{Q\in\mathcal{G}'} \sum_{j=1}^{q} \int \mathcal{X}_Q(\mathbf{x}) (\psi_j f)(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| \, dx. \end{aligned} \tag{1.27}$$

Consider the term $\int \mathcal{X}_Q(\mathbf{x}) (\psi_j f)(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| \, dx$. By Lemma A.5.8 and Fubini's theorem this equals

$$\begin{aligned} &\int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} \mathcal{X}_{Q-\mathbf{p}_j}(\mathbf{x}) (\psi_j f)(\mathbf{h}(\mathbf{p}_i) + \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{G}_n \circ \cdots \circ \mathbf{G}_1(\mathbf{x})) \cdot \\ &|D\mathbf{F}(\mathbf{G}_n \circ \cdots \circ \mathbf{G}_1(\mathbf{x}))| \, |D\mathbf{G}_n(\mathbf{G}_{n-1} \circ \cdots \circ \mathbf{G}_1(\mathbf{x}))| \, |D\mathbf{G}_{n-1}(\mathbf{G}_{n-2} \circ \cdots \circ \mathbf{G}_1(\mathbf{x}))| \cdot \\ &\cdots |D\mathbf{G}_2(\mathbf{G}_1(\mathbf{x}))| \, |D\mathbf{G}_1(\mathbf{x})| \, dx_1 dV_{n-1}. \end{aligned} \tag{1.28}$$

Here dV_{n-1} is with respect to the variables, $x_2, \cdots, x_n$. Also $\mathbf{F}$ denotes $\mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1}$. Now

$$\mathbf{G}_1(\mathbf{x}) = (\alpha(\mathbf{x}), x_2, \cdots, x_n)^T$$

and is one to one. Therefore, fixing $x_2, \cdots, x_n$, $x_1 \to \alpha(\mathbf{x})$ is one to one. Also, $D\mathbf{G}_1(\mathbf{x}) = \frac{\partial \alpha}{\partial x_1}(\mathbf{x})$. Fixing $x_2, \cdots, x_n$, change the variable,

$$y_1 = \alpha(x_1, x_2, \cdots, x_n).$$

Thus

$$\mathbf{x} = (x_1, x_2, \cdots, x_n)^T = \mathbf{G}_1^{-1}(y_1, x_2, \cdots, x_n) \equiv \mathbf{G}_1^{-1}(\mathbf{x}')$$

Then in 1.28 you can use Corollary A.5.3 to write 1.28 as

$$\begin{aligned} &\int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} \mathcal{X}_{Q-\mathbf{p}_j}(\mathbf{G}_1^{-1}(\mathbf{x}')) (\psi_j f)(\mathbf{h}(\mathbf{p}_i) + \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{G}_n \circ \cdots \circ \mathbf{G}_1(\mathbf{G}_1^{-1}(\mathbf{x}'))) \cdot \\ &|D\mathbf{F}(\mathbf{G}_n \circ \cdots \circ \mathbf{G}_1(\mathbf{G}_1^{-1}(\mathbf{x}')))| \, |D\mathbf{G}_n(\mathbf{G}_{n-1} \circ \cdots \circ \mathbf{G}_1(\mathbf{G}_1^{-1}(\mathbf{x}')))| \cdot \\ &|D\mathbf{G}_{n-1}(\mathbf{G}_{n-2} \circ \cdots \circ \mathbf{G}_1(\mathbf{G}_1^{-1}(\mathbf{x}')))| \cdots |D\mathbf{G}_2(\mathbf{G}_1(\mathbf{G}_1^{-1}(\mathbf{x}')))| \, dy_1 dV_{n-1} \end{aligned} \tag{1.29}$$

which reduces to

$$\begin{aligned} &\int_{\mathbb{R}^n} \mathcal{X}_{Q-\mathbf{p}_j}(\mathbf{G}_1^{-1}(\mathbf{x}')) (\psi_j f)(\mathbf{h}(\mathbf{p}_i) + \mathbf{F}_1 \circ \cdots \circ \mathbf{F}_{n-1} \circ \mathbf{G}_n \circ \cdots \circ \mathbf{G}_2(\mathbf{x}')) \cdot \\ &|D\mathbf{F}(\mathbf{G}_n \circ \cdots \circ \mathbf{G}_2(\mathbf{x}'))| \, |D\mathbf{G}_n(\mathbf{G}_{n-1} \circ \cdots \circ \mathbf{G}_2(\mathbf{x}'))| \, |D\mathbf{G}_{n-1}(\mathbf{G}_{n-2} \circ \cdots \circ \mathbf{G}_2(\mathbf{x}'))| \cdot \\ &\cdots |D\mathbf{G}_2(\mathbf{x}')| \, dV_n. \end{aligned} \tag{1.30}$$

Now use Fubini's theorem again to make the inside integral taken with respect to x_2. Exactly the same process yields

$$\int_{\mathbb{R}^{n-1}}\int_{\mathbb{R}}\mathcal{X}_{Q-\mathbf{p}_j}\left(\mathbf{G}_1^{-1}\circ\mathbf{G}_2^{-1}(\mathbf{x}'')\right)\left(\psi_j f\right)(\mathbf{h}(\mathbf{p}_i)+\mathbf{F}_1\circ\cdots\circ\mathbf{F}_{n-1}\circ\mathbf{G}_n\circ\cdots\circ\mathbf{G}_3(\mathbf{x}''))\cdot$$
$$|D\mathbf{F}(\mathbf{G}_n\circ\cdots\circ\mathbf{G}_3(\mathbf{x}''))|\,|D\mathbf{G}_n(\mathbf{G}_{n-1}\circ\cdots\circ\mathbf{G}_3(\mathbf{x}''))|\,|D\mathbf{G}_{n-1}(\mathbf{G}_{n-2}\circ\cdots\circ\mathbf{G}_3(\mathbf{x}''))|\cdot$$
$$\cdots dy_2 dV_{n-1}. \tag{1.31}$$

Now $\mathbf{F}$ is just a composition of flips and so $|D\mathbf{F}(\mathbf{G}_n\circ\cdots\circ\mathbf{G}_3(\mathbf{x}''))|=1$ and so this term can be replaced with 1. Continuing this process, eventually yields an expression of the form

$$\int_{\mathbb{R}^n}\mathcal{X}_{Q-\mathbf{p}_j}\left(\mathbf{G}_1^{-1}\circ\cdots\circ\mathbf{G}_{n-2}^{-1}\circ\mathbf{G}_{n-1}^{-1}\circ\mathbf{G}_n^{-1}\circ\mathbf{F}^{-1}(\mathbf{y})\right)\left(\psi_j f\right)(\mathbf{h}(\mathbf{p}_i)+\mathbf{y})\,dV_n. \tag{1.32}$$

Denoting by $\mathbf{G}^{-1}$ the expression, $\mathbf{G}_1^{-1}\circ\cdots\circ\mathbf{G}_{n-2}^{-1}\circ\mathbf{G}_{n-1}^{-1}\circ\mathbf{G}_n^{-1}$,

$$\mathcal{X}_{Q-\mathbf{p}_j}\left(\mathbf{G}_1^{-1}\circ\cdots\circ\mathbf{G}_{n-2}^{-1}\circ\mathbf{G}_{n-1}^{-1}\circ\mathbf{G}_n^{-1}\circ\mathbf{F}^{-1}(\mathbf{y})\right)=1$$

exactly when $\mathbf{G}^{-1}\circ\mathbf{F}^{-1}(\mathbf{y})\in Q-\mathbf{p}_j$. Now recall that $\mathbf{h}(\mathbf{p}_j+\mathbf{x})-\mathbf{h}(\mathbf{p}_j)=\mathbf{F}\circ\mathbf{G}(\mathbf{x})$ and so the above holds exactly when

$$\begin{aligned}\mathbf{y} &= \mathbf{h}\left(\mathbf{p}_j+\mathbf{G}^{-1}\circ\mathbf{F}^{-1}(\mathbf{y})\right)-\mathbf{h}(\mathbf{p}_j)\in\mathbf{h}(\mathbf{p}_j+Q-\mathbf{p}_j)-\mathbf{h}(\mathbf{p}_j)\\ &= \mathbf{h}(Q)-\mathbf{h}(\mathbf{p}_j).\end{aligned}$$

Thus 1.32 reduces to

$$\int_{\mathbb{R}^n}\mathcal{X}_{\mathbf{h}(Q)-\mathbf{h}(\mathbf{p}_j)}(\mathbf{y})\left(\psi_j f\right)(\mathbf{h}(\mathbf{p}_i)+\mathbf{y})\,dV_n=\int_{\mathbb{R}^n}\mathcal{X}_{\mathbf{h}(Q)}(\mathbf{z})\left(\psi_j f\right)(\mathbf{z})\,dV_n.$$

It follows from 1.27

$$\begin{aligned}\mathcal{U}_{\mathcal{G}}(g)-\varepsilon &\leq \mathcal{L}_{\mathcal{G}}(g)\leq\sum_{Q\in\mathcal{G}'}\int\mathcal{X}_Q(\mathbf{x})\,f(\mathbf{h}(\mathbf{x}))\,|\det D\mathbf{h}(\mathbf{x})|\,dx\\ &= \sum_{Q\in\mathcal{G}'}\sum_{j=1}^{q}\int\mathcal{X}_Q(\mathbf{x})\left(\psi_j f\right)(\mathbf{h}(\mathbf{x}))\,|\det D\mathbf{h}(\mathbf{x})|\,dx\\ &= \sum_{Q\in\mathcal{G}'}\sum_{j=1}^{q}\int_{\mathbb{R}^n}\mathcal{X}_{\mathbf{h}(Q)}(\mathbf{z})\left(\psi_j f\right)(\mathbf{z})\,dV_n\\ &= \sum_{Q\in\mathcal{G}'}\int_{\mathbb{R}^n}\mathcal{X}_{\mathbf{h}(Q)}(\mathbf{z})\,f(\mathbf{z})\,dV_n\leq\int\mathcal{X}_{\mathbf{h}(U)}(\mathbf{z})\,f(\mathbf{z})\,dV_n\end{aligned}$$

which implies the inequality,

$$\int\mathcal{X}_U(\mathbf{x})\,f(\mathbf{h}(\mathbf{x}))\,|\det D\mathbf{h}(\mathbf{x})|\,dV_n\leq\int\mathcal{X}_{\mathbf{h}(U)}(\mathbf{z})\,f(\mathbf{z})\,dV_n$$

But now you can use the same information just derived to obtain equality. $\mathbf{x}=\mathbf{h}^{-1}(\mathbf{z})$ and so from what was just done,

$$\int\mathcal{X}_U(\mathbf{x})\,f(\mathbf{h}(\mathbf{x}))\,|\det D\mathbf{h}(\mathbf{x})|\,dV_n$$

$$\begin{aligned} &= \int \mathcal{X}_{\mathbf{h}^{-1}(\mathbf{h}(U))}(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| \, dV_n \\ &\geq \int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f(\mathbf{z}) \left|\det D\mathbf{h}\left(\mathbf{h}^{-1}(\mathbf{z})\right)\right| \left|\det D\mathbf{h}^{-1}(\mathbf{z})\right| dV_n \\ &= \int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f(\mathbf{z}) \, dV_n \end{aligned}$$

from the chain rule. In fact,

$$I = D\mathbf{h}\left(\mathbf{h}^{-1}(\mathbf{z})\right) D\mathbf{h}^{-1}(\mathbf{z})$$

and so

$$1 = \left|\det D\mathbf{h}\left(\mathbf{h}^{-1}(\mathbf{z})\right)\right| \left|\det D\mathbf{h}^{-1}(\mathbf{z})\right|.$$

This proves the lemma.

The change of variables theorem follows.

Theorem A.5.15 *Let U be a bounded open set with ∂U having content 0. Also let $\mathbf{h} \in C^1\left(\overline{U}; \mathbb{R}^n\right)$ be one to one on U with $D\mathbf{h}(\mathbf{x})^{-1}$ exists for all $\mathbf{x} \in U$. Let $f \in C\left(\overline{U}\right)$. Then*

$$\int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f(\mathbf{z}) \, dz = \int \mathcal{X}_U(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| \, dx$$

Proof: You note that the formula holds for $f^+ \equiv \frac{|f|+f}{2}$ and $f^- \equiv \frac{|f|-f}{2}$. Now $f = f^+ - f^-$ and so

$$\int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f(\mathbf{z}) \, dz$$

$$\begin{aligned} &= \int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f^+(\mathbf{z}) \, dz - \int \mathcal{X}_{\mathbf{h}(U)}(\mathbf{z}) f^-(\mathbf{z}) \, dz \\ &= \int \mathcal{X}_U(\mathbf{x}) f^+(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| \, dx - \int \mathcal{X}_U(\mathbf{x}) f^-(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| \, dx \\ &= \int \mathcal{X}_U(\mathbf{x}) f(\mathbf{h}(\mathbf{x})) |\det D\mathbf{h}(\mathbf{x})| \, dx. \end{aligned}$$

A.6 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. However, most of the difficulties are artifacts. They shouldn't even be considered! 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 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. An introduction to it is in [21]. Another very good source is [11]. 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 [17], [18], [22], and [19]. There is also a still more general integral called the generalized Riemann integral. A recent book on this subject is [5]. It is far easier to define than the Lebesgue integral but the convergence theorems are much harder to prove. An introduction is also in [17].

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] **Fitzpatrick P. M.,** *Advanced Calculus a course in Mathematical Analysis,* PWS Publishing Company 1996.

[11] **Fleming W**., *Functions of Several Variables,* Springer Verlag 1976.

[12] **Greenberg, M**. Advanced Engineering Mathematics, Second edition, Prentice Hall, 1998

[13] **Gurtin M.** *An introduction to continuum mechanics,* Academic press 1981.

[14] **Hardy G.,** *A Course Of Pure Mathematics, Tenth edition,* Cambridge University Press 1992.

[15] **Horn R. and Johnson C.** *matrix Analysis,* Cambridge University Press, 1985.

[16] **Karlin S. and Taylor H.** *A First Course in Stochastic Processes,* Academic Press, 1975.

[17] **Kuttler K. L.**, *Basic Analysis*, Rinton

[18] **Kuttler K.L.,** *Modern Analysis* CRC Press 1998.

[19] **Lang S.** *Real and Functional analysis* third edition Springer Verlag 1993. Press, 2001.

[20] **Nobel B. and Daniel J.** *Applied Linear Algebra,* Prentice Hall, 1977.

[21] **Rudin, W.,** *Principles of mathematical analysis,* McGraw Hill third edition 1976

[22] **Rudin W.,** *Real and Complex Analysis,* third edition, McGraw-Hill, 1987.

[23] **Salas S. and Hille E.,** *Calculus One and Several Variables,* Wiley 1990.

[24] **Sears and Zemansky,** *University Physics, Third edition, Addison Wesley 1963.*

[25] **Tierney John,** *Calculus and Analytic Geometry,* fourth edition, Allyn and Bacon, Boston, 1969.

Index