An Introduction To Linear Algebra

Kenneth Kuttler

January 6, 2007

Contents

The Real And Complex Numbers

1.1 The Number Line And Algebra Of The Real Numbers

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

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

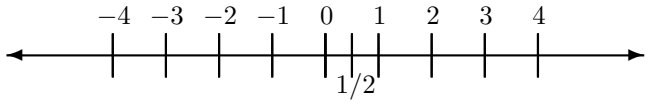
the natural numbers,

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

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

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

are each subsets of $\mathbb{R}$ as indicated in the following picture.



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

Axiom 1.1.1 $x + y = y + x$, *(commutative law for addition)*

Axiom 1.1.2 $x + 0 = x$, *(additive identity).*

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

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

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

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

Axiom 1.1.7 $1x=x$, (*multiplicative identity).*

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

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

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

1.2 The Complex Numbers

Just as a real number should be considered as a point on the line, a complex number is considered a point in the plane which can be identified in the usual way using the Cartesian coordinates of the point. Thus (a,b) identifies a point whose x coordinate is a and whose y coordinate is b. In dealing with complex numbers, such a point is written as $a+ib$ and multiplication and addition are defined in the most obvious way subject to the convention that $i^2=-1$. Thus,

$$(a+ib)+(c+id)=(a+c)+i\,(b+d)$$

and

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

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

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

You should prove the following theorem.

Theorem 1.2.1 *The complex numbers with multiplication and addition defined as above form a field satisfying all the field axioms listed on Page 7.*

The field of complex numbers is denoted as $\mathbb{C}$. An important construction regarding complex numbers is the complex conjugate denoted by a horizontal line above the number. It is defined as follows.

$$\overline{a+ib} \equiv a-ib.$$

What it does is reflect a given complex number across the x axis. Algebraically, the following formula is easy to obtain.

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

Definition 1.2.2 *Define the absolute value of a complex number as follows.*

$$|a+ib| \equiv \sqrt{a^2+b^2}.$$

Thus, denoting by z the complex number, $z = a+ib$,

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

With this definition, it is important to note the following. Be sure to verify this. It is not too hard but you need to do it.

Remark 1.2.3 *: Let $z = a+ib$ and $w = c+id$. Then $|z-w| = \sqrt{(a-c)^2+(b-d)^2}$. Thus the distance between the point in the plane determined by the ordered pair, (a,b) and the ordered pair (c,d) equals $|z-w|$ where z and w are as just described.*

For example, consider the distance between $(2,5)$ and $(1,8)$. From the distance formula this distance equals $\sqrt{(2-1)^2+(5-8)^2} = \sqrt{10}$. On the other hand, letting $z = 2+i5$ and $w = 1+i8$, $z-w = 1-i3$ and so $(z-w)\left(\overline{z-w}\right) = (1-i3)(1+i3) = 10$ so $|z-w| = \sqrt{10}$, the same thing obtained with the distance formula.

Complex numbers, are often written in the so called polar form which is described next. Suppose $x+iy$ is a complex number. Then

$$x+iy = \sqrt{x^2+y^2}\left(\frac{x}{\sqrt{x^2+y^2}} + i\frac{y}{\sqrt{x^2+y^2}}\right).$$

Now note that

$$\left(\frac{x}{\sqrt{x^2+y^2}}\right)^2 + \left(\frac{y}{\sqrt{x^2+y^2}}\right)^2 = 1$$

and so

$$\left(\frac{x}{\sqrt{x^2+y^2}}, \frac{y}{\sqrt{x^2+y^2}}\right)$$

is a point on the unit circle. Therefore, there exists a unique angle, $\theta \in [0, 2\pi)$ such that

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

The polar form of the complex number is then

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

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

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

Theorem 1.2.4 *Let $r > 0$ be given. Then if n is a positive integer,*

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

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

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

which by induction equals

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

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

Corollary 1.2.5 *Let z be a non zero complex number. Then there are always exactly k k^{th} roots of z in $\mathbb{C}$.*

Proof: Let $z = x + iy$ and let $z = |z|(\cos t + i\sin t)$ be the polar form of the complex number. By De Moivre's theorem, a complex number,

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

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

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

This requires $r^k = |z|$ and so $r = |z|^{1/k}$ and also both $\cos(k\alpha) = \cos t$ and $\sin(k\alpha) = \sin t$. This can only happen if

$$k\alpha = t + 2l\pi$$

for l an integer. Thus

$$\alpha = \frac{t + 2l\pi}{k}, l \in \mathbb{Z}$$

and so the k^{th} roots of z are of the form

$$|z|^{1/k}\left(\cos\left(\frac{t+2l\pi}{k}\right) + i\sin\left(\frac{t+2l\pi}{k}\right)\right),\ l \in \mathbb{Z}.$$

Since the cosine and sine are periodic of period 2π, there are exactly k distinct numbers which result from this formula.

Example 1.2.6 *Find the three cube roots of i.*

First note that $i = 1\left(\cos\left(\frac{\pi}{2}\right) + i\sin\left(\frac{\pi}{2}\right)\right)$. Using the formula in the proof of the above corollary, the cube roots of i are

$$1\left(\cos\left(\frac{(\pi/2) + 2l\pi}{3}\right) + i\sin\left(\frac{(\pi/2) + 2l\pi}{3}\right)\right)$$

where $l = 0, 1, 2$. Therefore, the roots are

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

and

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

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

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

Example 1.2.7 *Factor the polynomial* $x^3 - 27$.

First find the cube roots of 27. By the above proceedure using De Moivre's theorem, these cube roots are $3, 3\left(\frac{-1}{2} + i\frac{\sqrt{3}}{2}\right)$, and $3\left(\frac{-1}{2} - i\frac{\sqrt{3}}{2}\right)$. Therefore, $x^3 + 27 =$

$$(x-3)\left(x - 3\left(\frac{-1}{2} + i\frac{\sqrt{3}}{2}\right)\right)\left(x - 3\left(\frac{-1}{2} - i\frac{\sqrt{3}}{2}\right)\right).$$

Note also $\left(x - 3\left(\frac{-1}{2} + i\frac{\sqrt{3}}{2}\right)\right)\left(x - 3\left(\frac{-1}{2} - i\frac{\sqrt{3}}{2}\right)\right) = x^2 + 3x + 9$ and so

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

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

The real and complex numbers both are fields satisfying the axioms on Page 7 and it is usually one of these two fields which is used in linear algebra. The numbers are often called scalars. However, it turns out that all algebraic notions work for any field and there are many others. For this reason, I will often refer to the field of scalars as $\mathbb{F}$ although $\mathbb{F}$ will usually be either the real or complex numbers. If there is any doubt, assume it is the field of complex numbers which is meant.

1.3 Exercises

1. Let $z = 5 + i9$. Find z^{-1}.
2. Let $z = 2 + i7$ and let $w = 3 - i8$. Find $zw, z + w, z^2$, and w/z.
3. Give the complete solution to $x^4 + 16 = 0$.
4. Graph the complex cube roots of 8 in the complex plane. Do the same for the four fourth roots of 16.
5. If z is a complex number, show there exists ω a complex number with $|\omega| = 1$ and $\omega z = |z|$.
6. De Moivre's theorem says $[r(\cos t + i\sin t)]^n = r^n(\cos nt + i\sin nt)$ for n a positive integer. Does this formula continue to hold for all integers, n, even negative integers? Explain.
7. You already know formulas for $\cos(x+y)$ and $\sin(x+y)$ and these were used to prove De Moivre's theorem. Now using De Moivre's theorem, derive a formula for $\sin(5x)$ and one for $\cos(5x)$. **Hint:** Use the binomial theorem.
8. If z and w are two complex numbers and the polar form of z involves the angle θ while the polar form of w involves the angle ϕ, show that in the polar form for zw the angle involved is $\theta + \phi$. Also, show that in the polar form of a complex number, z, $r = |z|$.
9. Factor $x^3 + 8$ as a product of linear factors.
10. Write $x^3 + 27$ in the form $(x+3)\left(x^2 + ax + b\right)$ where $x^2 + ax + b$ cannot be factored any more using only real numbers.
11. Completely factor $x^4 + 16$ as a product of linear factors.

12. Factor $x^4 + 16$ as the product of two quadratic polynomials each of which cannot be factored further without using complex numbers.

13. If z, w are complex numbers prove $\overline{zw} = \overline{z}\,\overline{w}$ and then show by induction that $\overline{z_1 \cdots z_m} = \overline{z_1} \cdots \overline{z_m}$. Also verify that $\overline{\sum_{k=1}^m z_k} = \sum_{k=1}^m \overline{z_k}$. In words this says the conjugate of a product equals the product of the conjugates and the conjugate of a sum equals the sum of the conjugates.

14. Suppose $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ where all the a_k are real numbers. Suppose also that $p(z) = 0$ for some $z \in \mathbb{C}$. Show it follows that $p(\overline{z}) = 0$ also.

15. I claim that $1 = -1$. Here is why.

$$-1 = i^2 = \sqrt{-1}\sqrt{-1} = \sqrt{(-1)^2} = \sqrt{1} = 1.$$

This is clearly a remarkable result but is there something wrong with it? If so, what is wrong?

16. De Moivre's theorem is really a grand thing. I plan to use it now for rational exponents, not just integers.

$$1 = 1^{(1/4)} = (\cos 2\pi + i \sin 2\pi)^{1/4} = \cos(\pi/2) + i \sin(\pi/2) = i.$$

Therefore, squaring both sides it follows $1 = -1$ as in the previous problem. What does this tell you about De Moivre's theorem? Is there a profound difference between raising numbers to integer powers and raising numbers to non integer powers?

Systems Of Equations

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

$$x + y = 7 \text{ and } 2x - y = 8. \tag{2.1}$$

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

$$x + y = 7, \overbrace{2x - y + (-2)(x + y) = 8 + (-2)(7)}^{-3y=-6}. \tag{2.2}$$

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

Why exactly does the replacement of one equation with a multiple of another added to it not change the solution set? The two equations of 2.1 are of the form

$$E_1 = f_1, E_2 = f_2 \tag{2.3}$$

where E_1 and E_2 are expressions involving the variables. The claim is that if a is a number, then 2.3 has the same solution set as

$$E_1 = f_1,\ E_2 + aE_1 = f_2 + af_1. \tag{2.4}$$

Why is this?

If (x, y) solves 2.3 then it solves the first equation in 2.4. Also, it satisfies $aE_1 = af_1$ and so, since it also solves $E_2 = f_2$ it must solve the second equation in 2.4. If (x, y) solves 2.4 then it solves the first equation of 2.3. Also $aE_1 = af_1$ and it is given that the second equation of 2.4 is verified. Therefore, $E_2 = f_2$ and it follows (x, y) is a solution of the second equation in 2.3. This shows the solutions to 2.3 and 2.4 are exactly the same which means they have the same solution set. Of course the same reasoning applies with no change if there are many more variables than two and many more equations than two. It is still the case that when one equation is replaced with a multiple of another one added to itself, the solution set of the whole system does not change.

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

Example 2.0.1 *Find the solutions to the system,*

$$\begin{array}{c} x+3y+6z=25 \\ 2x+7y+14z=58 \\ 2y+5z=19 \end{array} \tag{2.5}$$

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

$$\begin{array}{c} x+3y+6z=25 \\ y+2z=8 \\ 2y+5z=19 \end{array} \tag{2.6}$$

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

$$\begin{array}{c} x+3y+6z=25 \\ y+2z=8 \\ z=3 \end{array} \tag{2.7}$$

At this point, you can tell what the solution is. This system has the same solution as the original system and in the above, $z=3$. Then using this in the second equation, it follows $y+6=8$ and so $y=2$. Now using this in the top equation yields $x+6+18=25$ and so $x=1$.

This process is not really much different from what you have always done in solving a single equation. For example, suppose you wanted to solve $2x+5=3x-6$. You did the same thing to both sides of the equation thus preserving the solution set until you obtained an equation which was simple enough to give the answer. In this case, you would add $-2x$ to both sides and then add 6 to both sides. This yields $x=11$.

In 2.7 you could have continued as follows. Add (-2) times the bottom equation to the middle and then add (-6) times the bottom to the top. This yields

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

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

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

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

It is foolish to write the variables every time you do these operations. It is easier to write the system 2.5 as the following "augmented matrix"

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

It has exactly the same information as the original system but here it is understood there is an x column, $\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}$, a y column, $\begin{pmatrix} 3 \\ 7 \\ 2 \end{pmatrix}$ and a z column, $\begin{pmatrix} 6 \\ 14 \\ 5 \end{pmatrix}$. The rows correspond

to the equations in the system. Thus the top row in the augmented matrix corresponds to the equation,

$$x + 3y + 6z = 25.$$

Now when you replace an equation with a multiple of another equation added to itself, you are just taking a row of this augmented matrix and replacing it with a multiple of another row added to it. Thus the first step in solving 2.5 would be to take (-2) times the first row of the augmented matrix above and add it to the second row,

$$\begin{pmatrix} 1 & 3 & 6 & 25 \\ 0 & 1 & 2 & 8 \\ 0 & 2 & 5 & 19 \end{pmatrix}.$$

Note how this corresponds to 2.6. Next take (-2) times the second row and add to the third,

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

which is the same as 2.7. You get the idea I hope. Write the system as an augmented matrix and follow the proceedure of either switching rows, multiplying a row by a non zero number, or replacing a row by a multiple of another row added to it. Each of these operations leaves the solution set unchanged. These operations are called row operations.

Definition 2.0.2 *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.*

Example 2.0.3 *Give the complete solution to the system of equations,* $5x+10y-7z=-2$, $2x+4y-3z=-1$, *and* $3x+6y+5z=9$.

The augmented matrix for this system is

$$\begin{pmatrix} 2 & 4 & -3 & -1 \\ 5 & 10 & -7 & -2 \\ 3 & 6 & 5 & 9 \end{pmatrix}$$

Multiply the second row by 2, the first row by 5, and then take (-1) times the first row and add to the second. Then multiply the first row by $1/5$. This yields

$$\begin{pmatrix} 2 & 4 & -3 & -1 \\ 0 & 0 & 1 & 1 \\ 3 & 6 & 5 & 9 \end{pmatrix}$$

Now, combining some row operations, take (-3) times the first row and add this to 2 times the last row and replace the last row with this. This yields.

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

Putting in the variables, the last two rows say $z = 1$ and $z = 21$. This is impossible so the last system of equations determined by the above augmented matrix has no solution.

However, it has the same solution set as the first system of equations. This shows there is no solution to the three given equations. When this happens, the system is called inconsistent.

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

Example 2.0.4 *Give the complete solution to the system of equations,* $3x - y - 5z = 9$, $y - 10z = 0$, *and* $-2x + y = -6$.

The augmented matrix of this system is

$$\begin{pmatrix} 3 & -1 & -5 & 9 \\ 0 & 1 & -10 & 0 \\ -2 & 1 & 0 & -6 \end{pmatrix}$$

Replace the last row with 2 times the top row added to 3 times the bottom row. This gives

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

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

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

Take the middle row and add to the top and then divide the top row which results by 3.

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

This says $y = 10z$ and $x = 3 + 5z$. Apparently z can equal any number. Therefore, the solution set of this system is $x = 3 + 5t, y = 10t$, and $z = t$ where t is completely arbitrary. The system has an infinite set of solutions and this is a good description of the solutions. This is what it is all about, finding the solutions to the system.

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

Definition 2.0.5 *A system of linear equations is a list of equations,*

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

where a_{ij} *are numbers,* f_j *is a number, and it is desired to find* $(x_1, \cdots, x_n)$ *solving each of the equations listed.*

As illustrated above, such a system of linear equations may have a unique solution, no solution, or infinitely many solutions. It turns out these are the only three cases which can occur for linear systems. Furthermore, you do exactly the same things to solve any linear system. You write the augmented matrix and do row operations until you get a simpler system in which it is possible to see the solution. All is based on the observation that the row operations do not change the solution set. You can have more equations than variables, fewer equations than variables, etc. It doesn't matter. You always set up the augmented matrix and go to work on it. These things are all the same.

Example 2.0.6 *Give the complete solution to the system of equations,* $-41x + 15y = 168$, $109x - 40y = -447$, $-3x + y = 12$, *and* $2x + z = -1$.

The augmented matrix is

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

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

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

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

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

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

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

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

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

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

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

2.1 Exercises

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

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

3. Consider the system $-5x + 2y - z = 0$ and $-5x - 2y - z = 0$. Both equations equal zero and so $-5x + 2y - z = -5x - 2y - z$ which is equivalent to $y = 0$. Thus x and z can equal anything. But when $x = 1$, $z = -4$, and $y = 0$ are plugged in to the equations, it doesn't work. Why?

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

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

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

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

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

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

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

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

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

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

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

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

16. Four times the weight of Gaston is 150 pounds more than the weight of Ichabod. Four times the weight of Ichabod is 660 pounds less than seventeen times the weight of Gaston. Four times the weight of Gaston plus the weight of Siegfried equals 290 pounds. Brunhilde would balance all three of the others. Find the weights of the four girls.

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

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

$\mathbb{F}^n$

The notation, $\mathbb{C}^n$ refers to the collection of ordered lists of n complex numbers. Since every real number is also a complex number, this simply generalizes the usual notion of $\mathbb{R}^n$, the collection of all ordered lists of n real numbers. In order to avoid worrying about whether it is real or complex numbers which are being referred to, the symbol $\mathbb{F}$ will be used. If it is not clear, always pick $\mathbb{C}$.

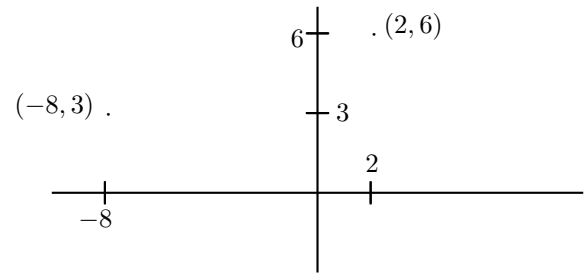
Definition 3.0.1 *Define* $\mathbb{F}^n \equiv \{(x_1, \cdots, x_n) : x_j \in \mathbb{F} \text{ for } j = 1, \cdots, n\}$. $(x_1, \cdots, x_n) = (y_1, \cdots, y_n)$ *if and only if for all* $j = 1, \cdots, n$, $x_j = y_j$. *When* $(x_1, \cdots, x_n) \in \mathbb{F}^n$, *it is conventional to denote* $(x_1, \cdots, x_n)$ *by the single bold face letter,* $\mathbf{x}$. *The numbers,* x_j *are called the coordinates. The set*

$$\{(0, \cdots, 0, t, 0, \cdots, 0) : t \in \mathbb{F}\}$$

for t *in the* i^{th} *slot is called the* i^{th} *coordinate axis. The point* $\mathbf{0} \equiv (0, \cdots, 0)$ *is called the origin.*

Thus $(1, 2, 4i) \in \mathbb{F}^3$ and $(2, 1, 4i) \in \mathbb{F}^3$ but $(1, 2, 4i) \neq (2, 1, 4i)$ because, even though the same numbers are involved, they don't match up. In particular, the first entries are not equal.

The geometric significance of $\mathbb{R}^n$ for $n \leq 3$ has been encountered already in calculus or in precalculus. Here is a short review. First consider the case when $n = 1$. Then from the definition, $\mathbb{R}^1 = \mathbb{R}$. Recall that $\mathbb{R}$ is identified with the points of a line. Look at the number line again. Observe that this amounts to identifying a point on this line with a real number. In other words a real number determines where you are on this line. Now suppose $n = 2$ and consider two lines which intersect each other at right angles as shown in the following picture.



Notice how you can identify a point shown in the plane with the ordered pair, $(2,6)$. You go to the right a distance of 2 and then up a distance of 6. Similarly, you can identify another point in the plane with the ordered pair $(-8,3)$. Go to the left a distance of 8 and then up a distance of 3. The reason you go to the left is that there is a $-$ sign on the eight. From this reasoning, every ordered pair determines a unique point in the plane. Conversely, taking a point in the plane, you could draw two lines through the point, one vertical and the other horizontal and determine unique points, x_1 on the horizontal line in the above picture and x_2 on the vertical line in the above picture, such that the point of interest is identified with the ordered pair, (x_1, x_2). In short, points in the plane can be identified with ordered pairs similar to the way that points on the real line are identified with real numbers. Now suppose $n = 3$. As just explained, the first two coordinates determine a point in a plane. Letting the third component determine how far up or down you go, depending on whether this number is positive or negative, this determines a point in space. Thus, $(1, 4, -5)$ would mean to determine the point in the plane that goes with $(1, 4)$ and then to go below this plane a distance of 5 to obtain a unique point in space. You see that the ordered triples correspond to points in space just as the ordered pairs correspond to points in a plane and single real numbers correspond to points on a line.

You can't stop here and say that you are only interested in $n \leq 3$. What if you were interested in the motion of two objects? You would need three coordinates to describe where the first object is and you would need another three coordinates to describe where the other object is located. Therefore, you would need to be considering $\mathbb{R}^6$. If the two objects moved around, you would need a time coordinate as well. As another example, consider a hot object which is cooling and suppose you want the temperature of this object. How many coordinates would be needed? You would need one for the temperature, three for the position of the point in the object and one more for the time. Thus you would need to be considering $\mathbb{R}^5$. Many other examples can be given. Sometimes n is very large. This is often the case in applications to business when they are trying to maximize profit subject to constraints. It also occurs in numerical analysis when people try to solve hard problems on a computer.

There are other ways to identify points in space with three numbers but the one presented is the most basic. In this case, the coordinates are known as Cartesian coordinates after Descartes[1] who invented this idea in the first half of the seventeenth century. I will often not bother to draw a distinction between the point in n dimensional space and its Cartesian coordinates.

The geometric significance of $\mathbb{C}^n$ for $n > 1$ is not available because each copy of $\mathbb{C}$ corresponds to the plane or $\mathbb{R}^2$.

3.1 Algebra in $\mathbb{F}^n$

There are two algebraic operations done with elements of $\mathbb{F}^n$. One is addition and the other is multiplication by numbers, called scalars. In the case of $\mathbb{C}^n$ the scalars are complex numbers while in the case of $\mathbb{R}^n$ the only allowed scalars are real numbers. Thus, the scalars always come from $\mathbb{F}$ in either case.

Definition 3.1.1 *If $\mathbf{x} \in \mathbb{F}^n$ and $a \in \mathbb{F}$, also called a scalar, then $a\mathbf{x} \in \mathbb{F}^n$ is defined by*

$$a\mathbf{x} = a(x_1, \cdots, x_n) \equiv (ax_1, \cdots, ax_n). \tag{3.1}$$

[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.

This is known as scalar multiplication. If $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$ *then* $\mathbf{x} + \mathbf{y} \in \mathbb{F}^n$ *and is defined by*

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1, \cdots, x_n) + (y_1, \cdots, y_n) \\ &\equiv (x_1 + y_1, \cdots, x_n + y_n) \end{aligned} \tag{3.2}$$

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

Theorem 3.1.2 *For* $\mathbf{v}, \mathbf{w} \in \mathbb{F}^n$ *and* α, β *scalars, (real numbers), the following hold.*

$$\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}, \tag{3.3}$$

the commutative law of addition,

$$(\mathbf{v} + \mathbf{w}) + \mathbf{z} = \mathbf{v} + (\mathbf{w} + \mathbf{z}), \tag{3.4}$$

the associative law for addition,

$$\mathbf{v} + \mathbf{0} = \mathbf{v}, \tag{3.5}$$

the existence of an additive identity,

$$\mathbf{v} + (-\mathbf{v}) = \mathbf{0}, \tag{3.6}$$

the existence of an additive inverse, Also

$$\alpha(\mathbf{v} + \mathbf{w}) = \alpha\mathbf{v} + \alpha\mathbf{w}, \tag{3.7}$$

$$(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}, \tag{3.8}$$

$$\alpha(\beta\mathbf{v}) = \alpha\beta(\mathbf{v}), \tag{3.9}$$

$$1\mathbf{v} = \mathbf{v}. \tag{3.10}$$

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

You should verify these properties all hold. For example, consider 3.7

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

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

3.2 Exercises

1. Verify all the properties 3.3-3.10.
2. Compute $5(1, 2 + 3i, 3, -2) + 6(2 - i, 1, -2, 7)$.
3. Draw a picture of the points in $\mathbb{R}^2$ which are determined by the following ordered pairs.
   (a) $(1, 2)$

(b) $(-2,-2)$
(c) $(-2,3)$
(d) $(2,-5)$

4. Does it make sense to write $(1,2)+(2,3,1)$? Explain.

5. Draw a picture of the points in $\mathbb{R}^3$ which are determined by the following ordered triples.

(a) $(1,2,0)$
(b) $(-2,-2,1)$
(c) $(-2,3,-2)$

3.3 Distance in $\mathbb{R}^n$

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

Definition 3.3.1 *Let* $\mathbf{x}=(x_1,\cdots,x_n)$ *and* $\mathbf{y}=(y_1,\cdots,y_n)$ *be two points in* $\mathbb{R}^n$. *Then* $|\mathbf{x}-\mathbf{y}|$ *to indicates the distance between these points and is defined as*

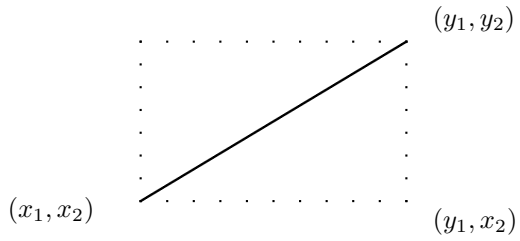
$$\text{distance between } \mathbf{x} \text{ and } \mathbf{y} \equiv |\mathbf{x}-\mathbf{y}| \equiv \left(\sum_{k=1}^{n} |x_k-y_k|^2\right)^{1/2}.$$

This is called the distance formula. The symbol, $B(\mathbf{a},r)$ *is defined by*

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

This is called an open ball of radius r *centered at* $\mathbf{a}$. *It gives all the points in* $\mathbb{R}^n$ *which are closer to* $\mathbf{a}$ *than* r.

First of all note this is a generalization of the notion of distance in $\mathbb{R}$. There the distance between two points, x and y was given by the absolute value of their difference. Thus $|x-y|$ is equal to the distance between these two points on $\mathbb{R}$. Now $|x-y|=\left((x-y)^2\right)^{1/2}$ where the square root is always the positive square root. Thus it is the same formula as the above definition except there is only one term in the sum. Geometrically, this is the right way to define distance which is seen from the Pythagorean theorem. Consider the following picture in the case that $n=2$.



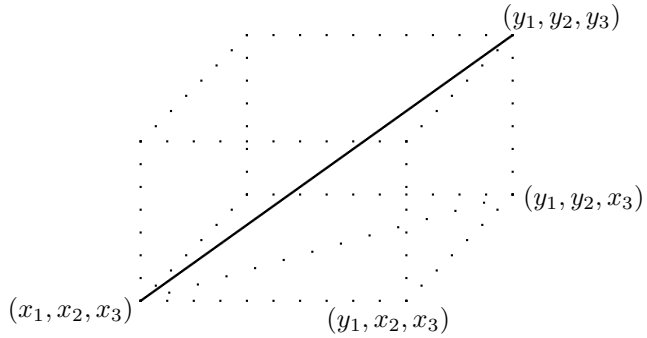
There are two points in the plane whose Cartesian coordinates are (x_1,x_2) and (y_1,y_2) respectively. Then the solid line joining these two points is the hypotenuse of a right triangle

which is half of the rectangle shown in dotted lines. What is its length? Note the lengths of the sides of this triangle are $|y_1 - x_1|$ and $|y_2 - x_2|$. Therefore, the Pythagorean theorem implies the length of the hypotenuse equals

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

which is just the formula for the distance given above.

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



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

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

while the length of the line joining (y_1, y_2, x_3) to (y_1, y_2, y_3) is just $|y_3 - x_3|$. Therefore, by the Pythagorean theorem again, the length of the line joining the points (x_1, x_2, x_3) and (y_1, y_2, y_3) equals

$$\begin{aligned}&\left\{\left[\left((y_1 - x_1)^2 + (y_2 - x_2)^2\right)^{1/2}\right]^2 + (y_3 - x_3)^2\right\}^{1/2}\\ &= \left((y_1 - x_1)^2 + (y_2 - x_2)^2 + (y_3 - x_3)^2\right)^{1/2},\end{aligned}$$

which is again just the distance formula above.

This completes the argument that the above definition is reasonable. Of course you cannot continue drawing pictures in ever higher dimensions but there is not problem with the formula for distance in any number of dimensions. Here is an example.

Example 3.3.2 *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 3.3.3 *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

$$2x+2y+2z = -9. \tag{3.11}$$

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 3.11 holds.

3.4 Distance in $\mathbb{F}^n$

How is distance between two points in $\mathbb{F}^n$ defined? It is done in exactly the same way as $\mathbb{R}^n$.

Definition 3.4.1 *Let* $\mathbf{x} = (x_1, \cdots, x_n)$ *and* $\mathbf{y} = (y_1, \cdots, y_n)$ *be two points in* $\mathbb{F}^n$. *Then writing* $|\mathbf{x}-\mathbf{y}|$ *to indicates the distance between these points,*

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

This is called the distance formula. Here the scalars, x_k *and* y_k *are complex numbers and* $|x_k - y_k|$ *refers to the complex absolute value defined earlier. Thus for* $z = x+iy \in \mathbb{C}$,

$$|z| \equiv \sqrt{x^2+y^2} = (z\overline{z})^{1/2}.$$

Note $|\mathbf{x}| = |\mathbf{x}-\mathbf{0}|$ *and gives the distance from* $\mathbf{x}$ *to* $\mathbf{0}$. *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 gives all the points in* $\mathbb{F}^n$ *which are closer to* $\mathbf{a}$ *than* r.

The following lemma is called the Cauchy Schwarz inequality. First here is a simple lemma which makes the proof easier.

Lemma 3.4.2 *If $z \in \mathbb{C}$ there exists $\theta \in \mathbb{C}$ such that $\theta z = |z|$ and $|\theta| = 1$.*

Proof: Let $\theta = 1$ if $z = 0$ and otherwise, let $\theta = \frac{\overline{z}}{|z|}$.

Lemma 3.4.3 *Let $\mathbf{x} = (x_1, \cdots, x_n)$ and $\mathbf{y} = (y_1, \cdots, y_n)$ be two points in $\mathbb{F}^n$. Then*

$$\left| \sum_{i=1}^{n} x_i \overline{y}_i \right| \leq |\mathbf{x}| \, |\mathbf{y}| . \tag{3.12}$$

Proof: Let $\theta \in \mathbb{C}$ such that $|\theta| = 1$ and

$$\theta \sum_{i=1}^{n} x_i \overline{y}_i = \left| \sum_{i=1}^{n} x_i \overline{y}_i \right|$$

Thus

$$\theta \sum_{i=1}^{n} x_i \overline{y}_i = \sum_{i=1}^{n} x_i \overline{\left(\overline{\theta} y_i\right)} = \left| \sum_{i=1}^{n} x_i \overline{y}_i \right| .$$

Consider $p(t) \equiv \sum_{i=1}^{n} \left(x_i + t\overline{\theta} y_i\right) \left(\overline{x_i + t\overline{\theta} y_i}\right)$ where $t \in \mathbb{R}$.

$$\begin{aligned} 0 &\leq p(t) = \sum_{i=1}^{n} |x_i|^2 + 2t \operatorname{Re} \left(\theta \sum_{i=1}^{n} x_i \overline{y}_i \right) + t^2 \sum_{i=1}^{n} |y_i|^2 \\ &= |\mathbf{x}|^2 + 2t \left| \sum_{i=1}^{n} x_i \overline{y}_i \right| + t^2 |\mathbf{y}|^2 \end{aligned}$$

If $|\mathbf{y}| = 0$ then 3.12 is obviously true because both sides equal zero. Therefore, assume $|\mathbf{y}| \neq 0$ and then $p(t)$ is a polynomial of degree two whose graph opens up. Therefore, it either has no zeroes, two zeros or one repeated zero. If it has two zeros, the above inequality must be violated because in this case the graph must dip below the x axis. Therefore, it either has no zeros or exactly one. From the quadratic formula this happens exactly when

$$4 \left| \sum_{i=1}^{n} x_i \overline{y}_i \right|^2 - 4 |\mathbf{x}|^2 |\mathbf{y}|^2 \leq 0$$

and so

$$\left| \sum_{i=1}^{n} x_i \overline{y}_i \right| \leq |\mathbf{x}| \, |\mathbf{y}|$$

as claimed. This proves the inequality.

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

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

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

The third fundamental property of distance is known as the triangle inequality. Recall that in any triangle the sum of the lengths of two sides is always at least as large as the third side.

Corollary 3.4.4 *Let* $\mathbf{x}, \mathbf{y}$ *be points of* $\mathbb{F}^n$. *Then*

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

Proof: Using the above lemma,

$$\begin{aligned} |\mathbf{x}+\mathbf{y}|^2 &\equiv \sum_{i=1}^{n} (x_i + y_i)\left(\overline{x_i + y_i}\right) \\ &= \sum_{i=1}^{n} |x_i|^2 + 2\,\mathrm{Re}\sum_{i=1}^{n} x_i\overline{y}_i + \sum_{i=1}^{n} |y_i|^2 \\ &\leq |\mathbf{x}|^2 + 2\left|\sum_{i=1}^{n} x_i\overline{y}_i\right| + |\mathbf{y}|^2 \\ &\leq |\mathbf{x}|^2 + 2\,|\mathbf{x}|\,|\mathbf{y}| + |\mathbf{y}|^2 \\ &= (|\mathbf{x}| + |\mathbf{y}|)^2 \end{aligned}$$

and so upon taking square roots of both sides,

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

and this proves the corollary.

3.5 Exercises

1. You are given two points in $\mathbb{R}^3, (4,5,-4)$ and $(2,3,0)$. Show the distance from the point, $(3,4,-2)$ to the first of these points is the same as the distance from this point to the second of the original pair of points. Note that $3 = \frac{4+2}{2}, 4 = \frac{5+3}{2}$. Obtain a theorem which will be valid for general pairs of points, (x,y,z) and (x_1,y_1,z_1) and prove your theorem using the distance formula.

2. A sphere is the set of all points which are at a given distance from a single given point. Find an equation for the sphere which is the set of all points that are at a distance of 4 from the point $(1,2,3)$ in $\mathbb{R}^3$.

3. A sphere centered at the point $(x_0,y_0,z_0) \in \mathbb{R}^3$ having radius r consists of all points, (x,y,z) whose distance to (x_0,y_0,z_0) equals r. Write an equation for this sphere in $\mathbb{R}^3$.

4. Suppose the distance between (x,y) and (x',y') were defined to equal the larger of the two numbers $|x-x'|$ and $|y-y'|$. Draw a picture of the sphere centered at the point, $(0,0)$ if this notion of distance is used.

5. Repeat the same problem except this time let the distance between the two points be $|x-x'| + |y-y'|$.

6. If (x_1,y_1,z_1) and (x_2,y_2,z_2) are two points such that $|(x_i,y_i,z_i)| = 1$ for $i = 1,2$, show that in terms of the usual distance, $\left|\left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}, \frac{z_1+z_2}{2}\right)\right| < 1$. What would happen if you used the way of measuring distance given in Problem 4 ($|(x,y,z)|$ = maximum of $|z|, |x|, |y|$.)?

7. Give a simple description using the distance formula of the set of points which are at an equal distance between the two points (x_1,y_1,z_1) and (x_2,y_2,z_2).

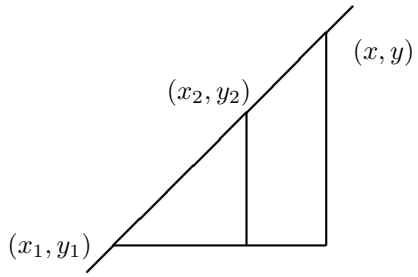
8. Show that $\mathbb{C}^n$ is essentially equal to $\mathbb{R}^{2n}$ and that the two notions of distance give exactly the same thing.

3.6 Lines in $\mathbb{R}^n$

To begin with consider the case $n = 1, 2$. In the case where $n = 1$, the only line is just $\mathbb{R}^1 = \mathbb{R}$. Therefore, if x_1 and x_2 are two different points in $\mathbb{R}$, consider

$$x = x_1 + t(x_2 - x_1)$$

where $t \in \mathbb{R}$ and the totality of all such points will give $\mathbb{R}$. You see that you can always solve the above equation for t, showing that every point on $\mathbb{R}$ is of this form. Now consider the plane. Does a similar formula hold? Let (x_1, y_1) and (x_2, y_2) be two different points in $\mathbb{R}^2$ which are contained in a line, l. Suppose that $x_1 \neq x_2$. Then if (x, y) is an arbitrary point on l,



Now by similar triangles,

$$m \equiv \frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1}$$

and so the point slope form of the line, l, is given as

$$y - y_1 = m(x - x_1).$$

If t is defined by

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

you obtain this equation along with

$$\begin{aligned} y &= y_1 + mt(x_2 - x_1) \\ &= y_1 + t(y_2 - y_1). \end{aligned}$$

Therefore,

$$(x, y) = (x_1, y_1) + t(x_2 - x_1, y_2 - y_1).$$

If $x_1 = x_2$, then in place of the point slope form above, $x = x_1$. Since the two given points are different, $y_1 \neq y_2$ and so you still obtain the above formula for the line. Because of this, the following is the definition of a line in $\mathbb{R}^n$.

Definition 3.6.1 *A line in $\mathbb{R}^n$ containing the two different points, $\mathbf{x}^1$ and $\mathbf{x}^2$ is the collection of points of the form*

$$\mathbf{x} = \mathbf{x}^1 + t\left(\mathbf{x}^2 - \mathbf{x}^1\right)$$

where $t \in \mathbb{R}$. This is known as a parametric equation and the variable t is called the parameter.

Often t denotes time in applications to Physics. Note this definition agrees with the usual notion of a line in two dimensions and so this is consistent with earlier concepts.

Lemma 3.6.2 *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.*

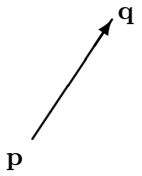
Proof: Let $\mathbf{x}^1 = \mathbf{b}$ and let $\mathbf{x}^2 - \mathbf{x}^1 = \mathbf{a}$ so that $\mathbf{x}^2 \neq \mathbf{x}^1$. Then $t\mathbf{a} + \mathbf{b} = \mathbf{x}^1 + t\left(\mathbf{x}^2 - \mathbf{x}^1\right)$ and so $\mathbf{x} = t\mathbf{a} + \mathbf{b}$ is a line containing the two different points, $\mathbf{x}^1$ and $\mathbf{x}^2$. This proves the lemma.

Definition 3.6.3 *Let* $\mathbf{p}$ *and* $\mathbf{q}$ *be two points in* $\mathbb{R}^n$, $\mathbf{p} \neq \mathbf{q}$. *The directed line segment from* $\mathbf{p}$ *to* $\mathbf{q}$, *denoted by* $\overrightarrow{\mathbf{pq}}$, *is defined to be the collection of points,*

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

with the direction corresponding to increasing t.

Think of $\overrightarrow{\mathbf{pq}}$ as an arrow whose point is on $\mathbf{q}$ and whose base is at $\mathbf{p}$ as shown in the following picture.



This line segment is a part of a line from the above Definition.

Example 3.6.4 *Find a parametric equation for the line through the points* $(1, 2, 0)$ *and* $(2, -4, 6)$.

Use the definition of a line given above to write

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

The reason for the word, "a", rather than the word, "the" is there are infinitely many different parametric equations for the same line. To see this replace t with $3s$. Then you obtain a parametric equation for the same line because the same set of points are obtained. The difference is they are obtained from different values of the parameter. What happens is this: The line is a set of points but the parametric description gives more information than that. It tells us how the set of points are obtained. Obviously, there are many ways to trace out a given set of points and each of these ways corresponds to a different parametric equation for the line.

3.7 Exercises

1. Suppose you are given two points, $(-a, 0)$ and $(a, 0)$ in $\mathbb{R}^2$ and a number, $r > 2a$. The set of points described by

   $$\left\{(x, y) \in \mathbb{R}^2 : |(x, y) - (-a, 0)| + |(x, y) - (a, 0)| = r\right\}$$

   is known as an ellipse. The two given points are known as the focus points of the ellipse. Simplify this to the form $\left(\frac{x-A}{\alpha}\right)^2 + \left(\frac{y}{\beta}\right)^2 = 1$. This is a nice exercise in messy algebra.

2. Let (x_1, y_1) and (x_2, y_2) be two points in $\mathbb{R}^2$. Give a simple description using the distance formula of the perpendicular bisector of the line segment joining these two points. Thus you want all points, (x, y) such that $|(x, y) - (x_1, y_1)| = |(x, y) - (x_2, y_2)|$.

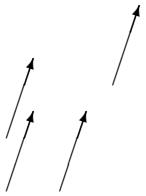
3. Find a parametric equation for the line through the points $(2, 3, 4, 5)$ and $(-2, 3, 0, 1)$.

4. Let $(x, y) = (2\cos(t), 2\sin(t))$ where $t \in [0, 2\pi]$. Describe the set of points encountered as t changes.

5. Let $(x, y, z) = (2\cos(t), 2\sin(t), t)$ where $t \in \mathbb{R}$. Describe the set of points encountered as t changes.

3.8 Physical Vectors In $\mathbb{R}^n$

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. Vectors are used to model this. 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 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 3.6.3 that $\overrightarrow{\mathbf{pq}}$ consists of the points of the form

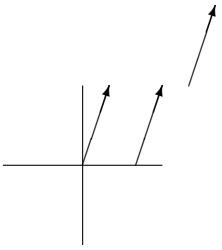
$$\mathbf{p} + t(\mathbf{q} - \mathbf{p})$$

where $t \in [0, 1]$. Subtract $\mathbf{p}$ from all these points to obtain the directed line segment consisting of the points

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

The point in $\mathbb{R}^n$, $\mathbf{q} - \mathbf{p}$, will represent the vector.

Geometrically, the arrow, $\overrightarrow{\mathbf{pq}}$, was slid so it points in the same direction and the base is at the origin, $\mathbf{0}$. For example, see the following picture.



In this way vectors can be identified with elements of $\mathbb{R}^n$.

The magnitude of a vector determined by a directed line segment $\overrightarrow{\mathbf{pq}}$ is just the distance between the point $\mathbf{p}$ and the point $\mathbf{q}$. By the distance formula this equals

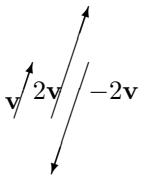
$$\left(\sum_{k=1}^{n}\left(q_k-p_k\right)^2\right)^{1/2}=|\mathbf{p}-\mathbf{q}|$$

and for $\mathbf{v}$ any vector in $\mathbb{R}^n$ the magnitude of $\mathbf{v}$ equals $\left(\sum_{k=1}^{n} v_k^2\right)^{1/2}=|\mathbf{v}|$.

What is the geometric significance of scalar multiplication? If $\mathbf{a}$ represents the vector, $\mathbf{v}$ in the sense that when it is slid to place its tail at the origin, the element of $\mathbb{R}^n$ at its point is $\mathbf{a}$, what is $r\mathbf{v}$?

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

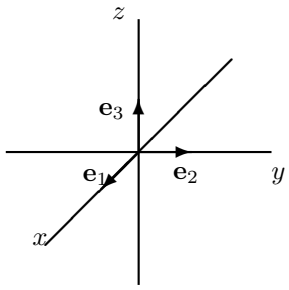
Thus the magnitude of $r\mathbf{v}$ equals $|r|$ times the magnitude of $\mathbf{v}$. If r is positive, then the vector represented by $r\mathbf{v}$ has the same direction as the vector, $\mathbf{v}$ because multiplying by the scalar, r, only has the effect of scaling all the distances. Thus the unit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

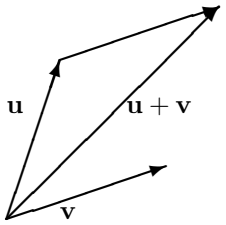
$$\begin{aligned} \mathbf{a} + \mathbf{b} &= (a_1 + b_1, \cdots, a_n + b_n). \\ &= \sum_{i=1}^{n} (a_i + b_i)\,\mathbf{e}_i. \end{aligned}$$

Thus the addition of vectors according to the rules of addition in $\mathbb{R}^n$, 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.

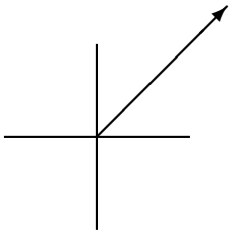
Example 3.8.1 *There are three ropes attached to a car and three people pull on these ropes. The first exerts a force of* $2\mathbf{i}+3\mathbf{j}-2\mathbf{k}$ *Newtons, the second exerts a force of* $3\mathbf{i}+5\mathbf{j}+\mathbf{k}$ *Newtons and the third exerts a force of* $5\mathbf{i}-\mathbf{j}+2\mathbf{k}$. *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.

The Newton is a unit of force like pounds.

Example 3.8.2 *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 $100/\sqrt{2}\mathbf{i} + 100/\sqrt{2}\mathbf{j}$.

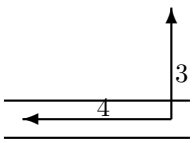
Example 3.8.3 *An airplane is traveling at* $100\mathbf{i}+\mathbf{j}+\mathbf{k}$ *kilometers per hour and at a certain instant of time its position is* $(1,2,1)$. *Here imagine a Cartesian coordinate system in which the third component is altitude and the first and second components are measured on 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 3.8.4 *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

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

3.9 Exercises

1. The wind blows from West to East at a speed of 50 kilometers per hour and an airplane is heading North West at a speed of 300 Kilometers per hour. What is the velocity of the airplane relative to the ground? What is the component of this velocity in the direction North?

2. In the situation of Problem 1 how many degrees to the West of North should the airplane head in order to fly exactly North. What will be the speed of the airplane?

3. In the situation of 2 suppose the airplane uses 34 gallons of fuel every hour at that air speed and that it needs to fly North a distance of 600 miles. Will the airplane have enough fuel to arrive at its destination given that it has 63 gallons of fuel?

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

5. A certain river is one half mile wide with a current flowing at 2 miles per hour from East to West. A man can swim at 3 miles per hour in still water. In what direction should he swim in order to travel directly across the river? What would the answer to this problem be if the river flowed at 3 miles per hour and the man could swim only at the rate of 2 miles per hour?

6. 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.

3.10 The Inner Product In $\mathbb{F}^n$

There are two ways of multiplying elements of $\mathbb{F}^n$ 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 3.10.1 *Let* $\mathbf{a},\mathbf{b}\in\mathbb{F}^n$ *define* $\mathbf{a}\cdot\mathbf{b}$ *as*

$$\mathbf{a}\cdot\mathbf{b}\equiv\sum_{k=1}^{n}a_k\overline{b}_k.$$

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 or in other words, points in $\mathbb{F}^n$.

Proposition 3.10.2 *The dot product satisfies the following properties.*

$$\mathbf{a}\cdot\mathbf{b}=\overline{\mathbf{b}\cdot\mathbf{a}} \tag{3.13}$$

$$\mathbf{a}\cdot\mathbf{a}\geq 0 \text{ and equals zero if and only if } \mathbf{a}=\mathbf{0} \tag{3.14}$$

$$(\alpha\mathbf{a}+\beta\mathbf{b})\cdot\mathbf{c}=\alpha(\mathbf{a}\cdot\mathbf{c})+\beta(\mathbf{b}\cdot\mathbf{c}) \tag{3.15}$$

$$\mathbf{c}\cdot(\alpha\mathbf{a}+\beta\mathbf{b})=\overline{\alpha}(\mathbf{c}\cdot\mathbf{a})+\overline{\beta}(\mathbf{c}\cdot\mathbf{b}) \tag{3.16}$$

$$|\mathbf{a}|^2=\mathbf{a}\cdot\mathbf{a} \tag{3.17}$$

You should verify these properties. Also be sure you understand that 3.16 follows from the first three and is therefore redundant. It is listed here for the sake of convenience.

Example 3.10.3 *Find* $(1,2,0,-1)\cdot(0,i,2,3)$.

This equals $0+2(-i)+0+-3=-3-2i$

The Cauchy Schwarz inequality takes the following form in terms of the inner product. I will prove it all over again, using only the above axioms for the dot product.

Theorem 3.10.4 *The dot product satisfies the inequality*

$$|\mathbf{a}\cdot\mathbf{b}|\leq|\mathbf{a}||\mathbf{b}|. \tag{3.18}$$

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

Proof: First define $\theta\in\mathbb{C}$ such that

$$\overline{\theta}(\mathbf{a}\cdot\mathbf{b})=|\mathbf{a}\cdot\mathbf{b}|,|\theta|=1,$$

and define a function of $t\in\mathbb{R}$

$$f(t)=(\mathbf{a}+t\theta\mathbf{b})\cdot(\mathbf{a}+t\theta\mathbf{b}).$$

Then by 3.14, $f(t)\geq 0$ for all $t\in\mathbb{R}$. Also from 3.15,3.16,3.13, and 3.17

$$\begin{aligned}f(t)&=\mathbf{a}\cdot(\mathbf{a}+t\theta\mathbf{b})+t\theta\mathbf{b}\cdot(\mathbf{a}+t\theta\mathbf{b})\\&=\mathbf{a}\cdot\mathbf{a}+t\overline{\theta}(\mathbf{a}\cdot\mathbf{b})+t\theta(\mathbf{b}\cdot\mathbf{a})+t^2|\theta|^2\mathbf{b}\cdot\mathbf{b}\\&=|\mathbf{a}|^2+2t\,\mathrm{Re}\,\overline{\theta}(\mathbf{a}\cdot\mathbf{b})+|\mathbf{b}|^2t^2\\&=|\mathbf{a}|^2+2t|\mathbf{a}\cdot\mathbf{b}|+|\mathbf{b}|^2t^2\end{aligned}$$

Now if $|\mathbf{b}|^2=0$ it must be the case that $\mathbf{a}\cdot\mathbf{b}=0$ because otherwise, you could pick large negative values of t and violate $f(t)\geq 0$. Therefore, in this case, the Cauchy Schwarz inequality holds. In the case that $|\mathbf{b}|\neq 0$, $y=f(t)$ is a polynomial which opens up and therefore, if it is always nonnegative, the quadratic formula requires that

$$\overbrace{4|\mathbf{a}\cdot\mathbf{b}|^2-4|\mathbf{a}|^2|\mathbf{b}|^2}^{\text{The discriminant}}\leq 0$$

since otherwise the function, $f(t)$ would have two real zeros and would necessarily have a graph which dips below the t axis. This proves 3.18.

It is clear from the axioms of the inner product that equality holds in 3.18 whenever one of the vectors is a scalar multiple of the other. It only remains to verify this is the only way equality can occur. If either vector equals zero, then equality is obtained in 3.18 so it can be assumed both vectors are non zero. Then if equality is achieved, it follows $f(t)$ has exactly one real zero because the discriminant vanishes. Therefore, for some value of $t,\mathbf{a}+t\theta\mathbf{b}=\mathbf{0}$ showing that $\mathbf{a}$ is a multiple of $\mathbf{b}$. This proves the theorem.

You should note that the entire argument was based only on the properties of the dot product listed in 3.13 - 3.17. This means that whenever something satisfies these properties, the Cauchy Schwartz inequality holds. There are many other instances of these properties besides vectors in $\mathbb{F}^n$.

The Cauchy Schwartz inequality allows a proof of the triangle inequality for distances in $\mathbb{F}^n$ in much the same way as the triangle inequality for the absolute value.

Theorem 3.10.5 *(Triangle inequality) For* $\mathbf{a}, \mathbf{b} \in \mathbb{F}^n$

$$|\mathbf{a}+\mathbf{b}| \leq |\mathbf{a}| + |\mathbf{b}| \tag{3.19}$$

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

$$||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a}-\mathbf{b}| \tag{3.20}$$

Proof: By properties of the dot product and the Cauchy Schwartz inequality,

$$\begin{aligned}
|\mathbf{a}+\mathbf{b}|^2 &= (\mathbf{a}+\mathbf{b})\cdot(\mathbf{a}+\mathbf{b}) \\
&= (\mathbf{a}\cdot\mathbf{a}) + (\mathbf{a}\cdot\mathbf{b}) + (\mathbf{b}\cdot\mathbf{a}) + (\mathbf{b}\cdot\mathbf{b}) \\
&= |\mathbf{a}|^2 + 2\,\mathrm{Re}\,(\mathbf{a}\cdot\mathbf{b}) + |\mathbf{b}|^2 \\
&\leq |\mathbf{a}|^2 + 2\,|\mathbf{a}\cdot\mathbf{b}| + |\mathbf{b}|^2 \\
&\leq |\mathbf{a}|^2 + 2\,|\mathbf{a}|\,|\mathbf{b}| + |\mathbf{b}|^2 \\
&= (|\mathbf{a}| + |\mathbf{b}|)^2 .
\end{aligned}$$

Taking square roots of both sides you obtain 3.19.

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 3.10.4 implies one of the vectors must be a multiple of the other. Say $\mathbf{b} = \alpha\mathbf{a}$. Also, to get equality in the first inequality, $(\mathbf{a}\cdot\mathbf{b})$ must be a nonnegative real number. Thus

$$0 \leq (\mathbf{a}\cdot\mathbf{b}) = (\mathbf{a}\cdot\alpha\mathbf{a}) = \overline{\alpha}\,|\mathbf{a}|^2 .$$

Therefore, α must be a real number which is nonnegative.

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,

$$|\mathbf{a}| - |\mathbf{b}| \leq |\mathbf{a}-\mathbf{b}| \tag{3.21}$$

Similarly,

$$|\mathbf{b}| - |\mathbf{a}| \leq |\mathbf{b}-\mathbf{a}| = |\mathbf{a}-\mathbf{b}| . \tag{3.22}$$

It follows from 3.21 and 3.22 that 3.20 holds. This is because $||\mathbf{a}| - |\mathbf{b}||$ equals the left side of either 3.21 or 3.22 and either way, $||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a}-\mathbf{b}|$. This proves the theorem.

3.11 Exercises

1. Find $(1,2,3,4)\cdot(2,0,1,3)$.

2. Show the angle between two vectors, $\mathbf{x}$ and $\mathbf{y}$ is a right angle if and only if $\mathbf{x} \cdot \mathbf{y} = 0$. **Hint:** Argue this happens when $|\mathbf{x} - \mathbf{y}|$ is the length of the hypotenuse of a right triangle having $|\mathbf{x}|$ and $|\mathbf{y}|$ as its legs. Now apply the pythagorean theorem and the observation that $|\mathbf{x} - \mathbf{y}|$ as defined above is the length of the segment joining $\mathbf{x}$ and $\mathbf{y}$.

3. Use the result of Problem 2 to consider the equation of a plane in $\mathbb{R}^3$. A plane in $\mathbb{R}^3$ through the point $\mathbf{a}$ is the set of points, $\mathbf{x}$ such that $\mathbf{x} - \mathbf{a}$ and a given normal vector, $\mathbf{n}$ form a right angle. Show that $\mathbf{n} \cdot (\mathbf{x} - \mathbf{a}) = 0$ is the equation of a plane. Now find the equation of the plane perpendicular to $\mathbf{n} = (1, 2, 3)$ which contains the point $(2, 0, 1)$. Give your answer in the form $ax + by + cz = d$ where a, b, c, and d are suitable constants.

4. Show that $(\mathbf{a} \cdot \mathbf{b}) = \frac{1}{4}\left[|\mathbf{a} + \mathbf{b}|^2 - |\mathbf{a} - \mathbf{b}|^2\right]$.

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

Applications In The Case $\mathbb{F} = \mathbb{R}$

4.1 Work And The Angle Between Vectors

4.1.1 Work And Projections

Our first application will be to the concept of work. The physical concept of work does not in any way correspond to the notion of work employed in ordinary conversation. For example, if you were to slide a 150 pound weight off a table which is three feet high and shuffle along the floor for 50 yards, sweating profusely and exerting all your strength to keep the weight from falling on your feet, keeping the height always three feet and then deposit this weight on another three foot high table, the physical concept of work would indicate that the force exerted by your arms did no work during this project even though the muscles in your hands and arms would likely be very tired. The reason for such an unusual definition is that even though your arms exerted considerable force on the weight, enough to keep it from falling, the direction of motion was at right angles to the force they exerted. The only part of a force which does work in the sense of physics is the component of the force in the direction of motion.

Theorem 4.1.1 *Let* $\mathbf{F}$ *and* $\mathbf{D}$ *be nonzero vectors. Then there exist unique vectors* $\mathbf{F}_{||}$ *and* $\mathbf{F}_{\perp}$ *such that*

$$\mathbf{F} = \mathbf{F}_{||} + \mathbf{F}_{\perp} \tag{4.1}$$

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

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

and $\mathbf{F}_{\perp} \cdot \mathbf{D} = 0$.

Proof: Suppose 4.1 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.

The following defines the concept of work.

Definition 4.1.2 *Let* $\mathbf{F}$ *be a force and let* $\mathbf{p}$ *and* $\mathbf{q}$ *be points in* $\mathbb{R}^n$. *Then the work,* W, *done by* $\mathbf{F}$ *on an object which moves from point* $\mathbf{p}$ *to point* $\mathbf{q}$ *is defined as*

$$\begin{aligned}W &\equiv \mathrm{proj}_{\mathbf{p}-\mathbf{q}}(\mathbf{F}) \cdot \frac{\mathbf{p}-\mathbf{q}}{|\mathbf{p}-\mathbf{q}|}\,|\mathbf{p}-\mathbf{q}| \\ &= \mathrm{proj}_{\mathbf{p}-\mathbf{q}}(\mathbf{F}) \cdot (\mathbf{p}-\mathbf{q}) \\ &= \mathbf{F}\cdot(\mathbf{p}-\mathbf{q}),\end{aligned}$$

the last equality holding because $\mathbf{F}_\perp \cdot (\mathbf{p}-\mathbf{q}) = 0$ *and* $\mathbf{F}_\perp + \mathrm{proj}_{\mathbf{p}-\mathbf{q}}(\mathbf{F}) = \mathbf{F}$.

Example 4.1.3 *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* $(-9,-3,4)$ *where distances are measured in meters.*

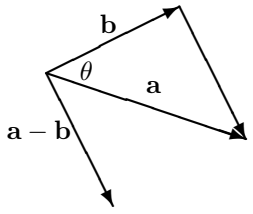
According to the definition, this work is

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

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.

4.1.2 The Angle Between Two Vectors

Given two vectors, $\mathbf{a}$ and $\mathbf{b}$, the included angle is the angle between these two vectors which is less than or equal to 180 degrees. The dot product can be used to determine the included angle between two vectors. To see how to do this, consider the following picture.



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,

$$\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}|\,|\mathbf{b}|\cos\theta. \tag{4.2}$$

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 4.1.4 *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 4.2 the cosine of the included angle equals

$$\cos\theta = \frac{9}{\sqrt{26}\sqrt{6}} = .720\,58$$

Now the cosine is known, the angle can be determines by solving the equation, $\cos\theta = .720\,58$. This will involve using a calculator or a table of trigonometric functions. The answer is $\theta = .766\,16$ radians or in terms of degrees, $\theta = .766\,16\times\frac{360}{2\pi} = 43.898°$. Recall how this last computation is done. Set up a proportion, $\frac{x}{.766\,16} = \frac{360}{2\pi}$ because $360°$ corresponds to 2π radians. However, in calculus, you should get used to thinking in terms of radians and not degrees. This is because all the important calculus formulas are defined in terms of radians.

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 $\mathbf{b}_\perp$?

$$\begin{aligned}|\mathbf{b}_\perp|^2 &= (\mathbf{b}-\mathrm{proj}_{\mathbf{a}}(\mathbf{b}))\cdot(\mathbf{b}-\mathrm{proj}_{\mathbf{a}}(\mathbf{b}))\\ &= \left(\mathbf{b}-\frac{\mathbf{b}\cdot\mathbf{a}}{|\mathbf{a}|^2}\mathbf{a}\right)\cdot\left(\mathbf{b}-\frac{\mathbf{b}\cdot\mathbf{a}}{|\mathbf{a}|^2}\mathbf{a}\right)\\ &= |\mathbf{b}|^2-2\frac{(\mathbf{b}\cdot\mathbf{a})^2}{|\mathbf{a}|^2}+\left(\frac{\mathbf{b}\cdot\mathbf{a}}{|\mathbf{a}|^2}\right)^2|\mathbf{a}|^2\\ &= |\mathbf{b}|^2\left(1-\frac{(\mathbf{b}\cdot\mathbf{a})^2}{|\mathbf{a}|^2\,|\mathbf{b}|^2}\right)\\ &= |\mathbf{b}|^2\left(1-\cos^2\theta\right) = |\mathbf{b}|^2\sin^2(\theta)\end{aligned}$$

where θ is the included angle between $\mathbf{a}$ and $\mathbf{b}$ which is less than π radians. Therefore, taking square roots,

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

4.2 Exercises

1. Use formula 4.2 to verify the Cauchy Schwartz inequality and to show that equality occurs if and only if one of the vectors is a scalar multiple of the other.
2. Find the angle between the vectors $3\mathbf{i}-\mathbf{j}-\mathbf{k}$ and $\mathbf{i}+4\mathbf{j}+2\mathbf{k}$.
3. Find the angle between the vectors $\mathbf{i}-2\mathbf{j}+\mathbf{k}$ and $\mathbf{i}+2\mathbf{j}-7\mathbf{k}$.

4. If $\mathbf{F}$ is a force and $\mathbf{D}$ is a vector, show $\text{proj}_{\mathbf{D}}(\mathbf{F}) = (|\mathbf{F}|\cos\theta)\,\mathbf{u}$ where $\mathbf{u}$ is the unit vector in the direction of $\mathbf{D}$, $\mathbf{u} = \mathbf{D}/|\mathbf{D}|$ and θ is the included angle between the two vectors, $\mathbf{F}$ and $\mathbf{D}$.

5. Show that the work done by a force $\mathbf{F}$ in moving an object along the line from $\mathbf{p}$ to $\mathbf{q}$ equals $|\mathbf{F}|\cos\theta\,|\mathbf{p}-\mathbf{q}|$. What is the geometric significance of the work being negative? What is the physical significance?

6. A boy drags a sled for 100 feet along the ground by pulling on a rope which is 20 degrees from the horizontal with a force of 10 pounds. How much work does this force do?

7. An object moves 10 meters in the direction of $\mathbf{j}$. There are two forces acting on this object, $\mathbf{F}_1 = \mathbf{i}+\mathbf{j}+2\mathbf{k}$, and $\mathbf{F}_2 = -5\mathbf{i}+2\mathbf{j}-6\mathbf{k}$. Find the total work done on the object by the two forces. **Hint:** You can take the work done by the resultant of the two forces or you can add the work done by each force.

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

9. In the discussion of the reflecting mirror which directs all rays to a particular point, $(0,p)$. Show that for any choice of positive C this point is the focus of the parabola and the directrix is $y = p - \frac{1}{C}$.

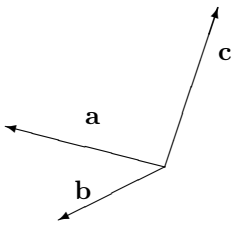
10. Suppose you wanted to make a solar powered oven to cook food. Are there reasons for using a mirror which is not parabolic? Also describe how you would design a good flash light with a beam which does not spread out too quickly.

4.3 The Cross Product

The cross product is the other way of multiplying two vectors in $\mathbb{R}^3$. It is very different from the dot product in many ways. First the geometric meaning is discussed and then a description in terms of coordinates is given. Both descriptions of the cross product are important. The geometric description is essential in order to understand the applications to physics and geometry while the coordinate description is the only way to practically compute the cross product.

Definition 4.3.1 *Three vectors,* $\mathbf{a}, \mathbf{b}, \mathbf{c}$ *form a right handed system if when you extend the fingers of your right hand along the vector,* $\mathbf{a}$ *and close them in the direction of* $\mathbf{b}$*, the thumb points roughly in the direction of* $\mathbf{c}$.

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



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

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

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

Definition 4.3.2 *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two vectors in* $\mathbb{R}^n$. *Then* $\mathbf{a} \times \mathbf{b}$ *is defined by the following two rules.*

1. $|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}|\,|\mathbf{b}| \sin\theta$ *where* θ *is the included angle.*
2. $\mathbf{a} \times \mathbf{b} \cdot \mathbf{a} = 0$, $\mathbf{a} \times \mathbf{b} \cdot \mathbf{b} = 0$, *and* $\mathbf{a}, \mathbf{b}, \mathbf{a} \times \mathbf{b}$ *forms a right hand system.*

The cross product satisfies the following properties.

$$\mathbf{a} \times \mathbf{b} = -(\mathbf{b} \times \mathbf{a}) \ , \ \mathbf{a} \times \mathbf{a} = \mathbf{0}, \tag{4.3}$$

For α a scalar,

$$(\alpha \mathbf{a}) \times \mathbf{b} = \alpha(\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (\alpha \mathbf{b}), \tag{4.4}$$

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

$$\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}, \tag{4.5}$$

$$(\mathbf{b} + \mathbf{c}) \times \mathbf{a} = \mathbf{b} \times \mathbf{a} + \mathbf{c} \times \mathbf{a}. \tag{4.6}$$

Formula 4.3 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 4.4 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 4.4. 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 4.3,

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

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

Now from the definition of the cross product,

$$\begin{array}{ll}\mathbf{i} \times \mathbf{j} = \mathbf{k} & \mathbf{j} \times \mathbf{i} = -\mathbf{k} \\ \mathbf{k} \times \mathbf{i} = \mathbf{j} & \mathbf{i} \times \mathbf{k} = -\mathbf{j} \\ \mathbf{j} \times \mathbf{k} = \mathbf{i} & \mathbf{k} \times \mathbf{j} = -\mathbf{i}\end{array}$$

With this information, the following gives the coordinate description of the cross product.

Proposition 4.3.3 *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*

$$\begin{aligned}\mathbf{a} \times \mathbf{b} = &(a_2b_3 - a_3b_2)\,\mathbf{i} + (a_3b_1 - a_1b_3)\,\mathbf{j} + \\ &+ (a_1b_2 - a_2b_1)\,\mathbf{k}.\end{aligned} \tag{4.7}$$

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

$$\begin{aligned}
&(a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}) \times (b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}) = \\
&a_1b_2\mathbf{i} \times \mathbf{j} + a_1b_3\mathbf{i} \times \mathbf{k} + a_2b_1\mathbf{j} \times \mathbf{i} + a_2b_3\mathbf{j} \times \mathbf{k} + \\
&+a_3b_1\mathbf{k} \times \mathbf{i} + a_3b_2\mathbf{k} \times \mathbf{j} \\
&= a_1b_2\mathbf{k} - a_1b_3\mathbf{j} - a_2b_1\mathbf{k} + a_2b_3\mathbf{i} + a_3b_1\mathbf{j} - a_3b_2\mathbf{i} \\
&= (a_2b_3 - a_3b_2)\,\mathbf{i} + (a_3b_1 - a_1b_3)\,\mathbf{j} + (a_1b_2 - a_2b_1)\,\mathbf{k} \qquad (4.8)
\end{aligned}$$

This proves the proposition.

It is probably impossible for most people to remember 4.7. Fortunately, there is a somewhat easier way to remember it. This involves the notion of a determinant. A determinant is a single number assigned to a square array of numbers as follows.

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

This is the definition of the determinant of a square array of numbers having two rows and two columns. Now using this, the determinant of a square array of numbers in which there are three rows and three columns is defined as follows.

$$\begin{aligned}
\det \begin{pmatrix} a & b & c \\ d & e & f \\ h & i & j \end{pmatrix} &\equiv (-1)^{1+1} a \begin{vmatrix} e & f \\ i & j \end{vmatrix} \\
&+ (-1)^{1+2} b \begin{vmatrix} d & f \\ h & j \end{vmatrix} + (-1)^{1+3} c \begin{vmatrix} d & e \\ h & i \end{vmatrix}.
\end{aligned}$$

Take the first entry in the top row, a, multiply by (-1) raised to the $1+1$ since a is in the first row and the first column, and then multiply by the determinant obtained by crossing out the row and the column in which a appears. Then add to this a similar number obtained from the next element in the first row, b. This time multiply by $(-1)^{1+2}$ because b is in the second column and the first row. When this is done do the same for c, the last element in the first row using a similar process. Using the definition of a determinant for square arrays of numbers having two columns and two rows, this equals

$$a\,(ej - if) + b\,(fh - dj) + c\,(di - eh)\,,$$

an expression which, like the one for the cross product will be impossible to remember, although the process through which it is obtained is not too bad. It turns out these two impossible to remember expressions are linked through the process of finding a determinant which was just described. The easy way to remember the description of the cross product in terms of coordinates, is to write

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} \qquad (4.9)$$

and then follow the same process which was just described for calculating determinants above. This yields

$$(a_2b_3 - a_3b_2)\,\mathbf{i} - (a_1b_3 - a_3b_1)\,\mathbf{j} + (a_1b_2 - a_2b_1)\,\mathbf{k} \qquad (4.10)$$

which is the same as 4.8. Later in the book a complete discussion of determinants is given but this will suffice for now.

Example 4.3.4 *Find* $(\mathbf{i}-\mathbf{j}+2\mathbf{k})\times(3\mathbf{i}-2\mathbf{j}+\mathbf{k})$.

Use 4.9 to compute this.

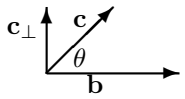
$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & -1 & 2 \\ 3 & -2 & 1 \end{vmatrix} = \begin{vmatrix} -1 & 2 \\ -2 & 1 \end{vmatrix}\mathbf{i} - \begin{vmatrix} 1 & 2 \\ 3 & 1 \end{vmatrix}\mathbf{j} + \begin{vmatrix} 1 & -1 \\ 3 & -2 \end{vmatrix}$$
$$= 3\mathbf{i}+5\mathbf{j}+\mathbf{k}.$$

4.3.1 The Distributive Law For The Cross Product

This section gives a proof for 4.5, a fairly difficult topic. It is included here for the interested student. If you are satisfied with taking the distributive law on faith, it is not necessary to read this section. The proof given here is quite clever and follows the one given in [4]. Another approach, based on volumes of parallelepipeds is found in [12] and is discussed a little later.

Lemma 4.3.5 *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,

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

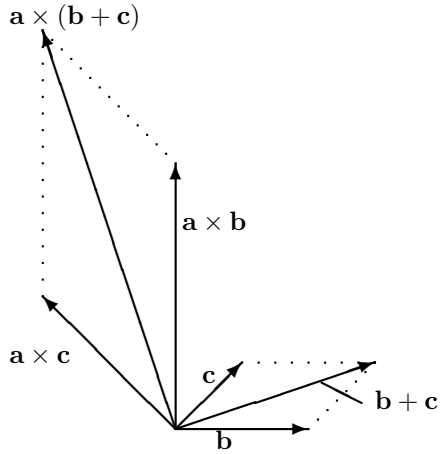
Therefore, $\mathbf{b}\times\mathbf{c}$ and $\mathbf{b}\times\mathbf{c}_\perp$ also have the same magnitude and so they are the same vector.

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

Theorem 4.3.6 *Let* $\mathbf{a},\mathbf{b}$, *and* $\mathbf{c}$ *be vectors in* $\mathbb{R}^3$. *Then*

$$\mathbf{a}\times(\mathbf{b}+\mathbf{c})=\mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c} \tag{4.11}$$

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 4.11.

Now suppose it is not necessarily the case that $\mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{c} = 0$. Then write $\mathbf{b} = \mathbf{b}_{||} + \mathbf{b}_{\perp}$ where $\mathbf{b}_{\perp} \cdot \mathbf{a} = 0$. Similarly $\mathbf{c} = \mathbf{c}_{||} + \mathbf{c}_{\perp}$. By the above lemma and what was just shown,

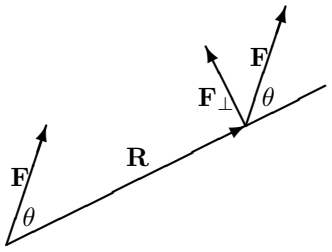
$$\begin{aligned} \mathbf{a} \times (\mathbf{b} + \mathbf{c}) &= \mathbf{a} \times (\mathbf{b} + \mathbf{c})_{\perp} \\ &= \mathbf{a} \times (\mathbf{b}_{\perp} + \mathbf{c}_{\perp}) \\ &= \mathbf{a} \times \mathbf{b}_{\perp} + \mathbf{a} \times \mathbf{c}_{\perp} \\ &= \mathbf{a} \times \mathbf{b} +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

This proves the theorem.

The result of Problem 8 of the exercises 4.2 is used to go from the first to the second line.

4.3.2 Torque

Imagine you are using a wrench to loosen a nut. The idea is to turn the nut by applying a force to the end of the wrench. If you push or pull the wrench directly toward or away from the nut, it should be obvious from experience that no progress will be made in turning the nut. The important thing is the component of force perpendicular to the wrench. It is this component of force which will cause the nut to turn. For example see the following picture.



In the picture a force, $\mathbf{F}$ is applied at the end of a wrench represented by the position vector, $\mathbf{R}$ and the angle between these two is θ. Then the tendency to turn will be $|\mathbf{R}|\,|\mathbf{F}_\perp| = |\mathbf{R}|\,|\mathbf{F}|\sin\theta$, which you recognize as the magnitude of the cross product of $\mathbf{R}$ and $\mathbf{F}$. If there were just one force acting at one point whose position vector is $\mathbf{R}$, perhaps this would be sufficient, but what if there are numerous forces acting at many different points with neither the position vectors nor the force vectors in the same plane; what then? To keep track of this sort of thing, define for each $\mathbf{R}$ and $\mathbf{F}$, the Torque vector,

$$\tau \equiv \mathbf{R} \times \mathbf{F}.$$

That way, if there are several forces acting at several points, the total torque can be obtained by simply adding up the torques associated with the different forces and positions.

Example 4.3.7 *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 4.3.8 *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.

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 4.3.9 *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,

$$\left(\sum_{k=1}^{p} \mathbf{R}_k g m_k - \mathbf{R}_0 \sum_{k=1}^{p} g m_k \right) \times \mathbf{u} = \mathbf{0} \tag{4.12}$$

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

$$\mathbf{R}_0 = \frac{\sum_{k=1}^{p} \mathbf{R}_k m_k}{\sum_{k=1}^{p} m_k}. \tag{4.13}$$

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 4.3.10 *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 4.13

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

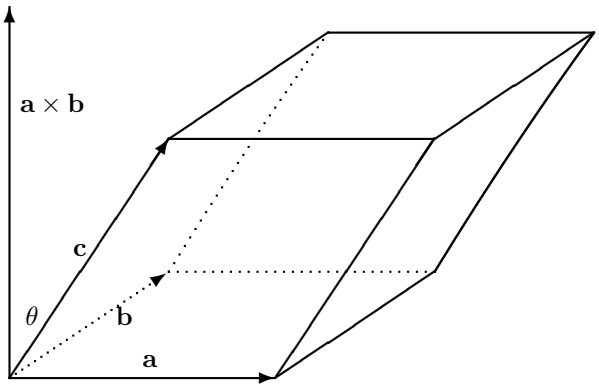
4.3.3 The Box Product

Definition 4.3.11 *A parallelepiped determined by the three vectors,* $\mathbf{a},\mathbf{b}$, *and* $\mathbf{c}$ *consists of*

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

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

The following is a picture of such a thing.



You notice the area of the base of the parallelepiped, the parallelogram determined by the vectors, $\mathbf{a}$ and $\mathbf{b}$ 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 equals

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

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 4.3.12 *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.

Here is another proof of the distributive law for the cross product. From the above picture $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$ because both of these give either the volume of a parallelepiped determined by the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ or -1 times the volume of this parallelepiped. Now to prove the distributive law, let $\mathbf{x}$ be a vector. From the above observation,

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

Therefore,

$$\mathbf{x} \cdot [\mathbf{a} \times (\mathbf{b} + \mathbf{c}) - (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c})] = 0$$

for all $\mathbf{x}$. In particular, this holds for $\mathbf{x} = \mathbf{a} \times (\mathbf{b} + \mathbf{c}) - (\mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c})$ showing that $\mathbf{a} \times (\mathbf{b} + \mathbf{c}) = \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c}$ and this proves the distributive law for the cross product another way.

4.4 Exercises

1. Show that if $\mathbf{a} \times \mathbf{u} = \mathbf{0}$ for all unit vectors, $\mathbf{u}$, then $\mathbf{a} = \mathbf{0}$.

2. If you only assume 4.12 holds for $\mathbf{u} = \mathbf{i}, \mathbf{j}, \mathbf{k}$, show that this implies 4.12 holds for all unit vectors, $\mathbf{u}$.

3. Let $m_1 = 5, m_2 = 1$, and $m_3 = 4$ where the masses are in kilograms and the distance is in meters. Suppose m_1 is located at $2\mathbf{i} - 3\mathbf{j} + \mathbf{k}$, m_2 is located at $\mathbf{i} - 3\mathbf{j} + 6\mathbf{k}$ and m_3 is located at $2\mathbf{i} + \mathbf{j} + 3\mathbf{k}$. Find the center of mass of these three masses.

4. Let $m_1 = 2, m_2 = 3$, and $m_3 = 1$ where the masses are in kilograms and the distance is in meters. Suppose m_1 is located at $2\mathbf{i} - \mathbf{j} + \mathbf{k}$, m_2 is located at $\mathbf{i} - 2\mathbf{j} + \mathbf{k}$ and m_3 is located at $4\mathbf{i} + \mathbf{j} + 3\mathbf{k}$. Find the center of mass of these three masses.

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

6. 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?

7. What does it mean geometrically if the box product of three vectors gives zero?

8. Suppose $\mathbf{a} = (a_1, a_2, a_3), \mathbf{b} = (b_1, b_2, b_3)$, and $\mathbf{c} = (c_1, c_2, c_3)$. Show the box product, $[\mathbf{a}, \mathbf{b}, \mathbf{c}]$ equals the determinant

$$\begin{vmatrix} c_1 & c_2 & c_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}.$$

9. It is desired to find an equation of a plane containing the two vectors, **a** and **b**. Using Problem 7, show an equation for this plane is

$$\begin{vmatrix} x & y & z \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = 0$$

That is, the set of all (x, y, z) such that the above expression equals zero.

10. 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.

11. Verify directly that the coordinate description of the cross product, $\mathbf{a} \times \mathbf{b}$ has the property that it is perpendicular to both **a** and **b**. Then show by direct computation that this coordinate description satisfies

$$\begin{aligned} |\mathbf{a} \times \mathbf{b}|^2 &= |\mathbf{a}|^2 |\mathbf{b}|^2 - (\mathbf{a} \cdot \mathbf{b})^2 \\ &= |\mathbf{a}|^2 |\mathbf{b}|^2 \left(1 - \cos^2(\theta)\right) \end{aligned}$$

where θ is the angle included between the two vectors. Explain why $|\mathbf{a} \times \mathbf{b}|$ has the correct magnitude. All that is missing is the material about the right hand rule. Verify directly from the coordinate description of the cross product that the right thing happens with regards to the vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$. Next verify that the distributive law holds for the coordinate description of the cross product. This gives another way to approach the cross product. First define it in terms of coordinates and then get the geometric properties from this.

4.5 Vector Identities And Notation

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 4.5.1 *The symbol, δ_{ij}, called the Kroneker delta symbol is defined as follows.*

$$\delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}.$$

With the Kroneker symbol, i and j can equal any integer in $\{1, 2, \cdots, n\}$ for any $n \in \mathbb{N}$.

Definition 4.5.2 *For i, j, and k integers in the set, $\{1, 2, 3\}$, ε_{ijk} is defined as follows.*

$$\varepsilon_{ijk} \equiv \begin{cases} 1 \text{ if } (i,j,k) = (1,2,3), (2,3,1), \text{ or } (3,1,2) \\ -1 \text{ if } (i,j,k) = (2,1,3), (1,3,2), \text{ or } (3,2,1) \\ 0 \text{ if there are any repeated integers} \end{cases}.$$

The subscripts ijk and ij in the above are called indices. A single one is called an index. This symbol, ε_{ijk} is also called the permutation symbol.

The way to think of ε_{ijk} is that $\varepsilon_{123} = 1$ and if you switch any two of the numbers in the list i, j, k, it changes the sign. Thus $\varepsilon_{ijk} = -\varepsilon_{jik}$ and $\varepsilon_{ijk} = -\varepsilon_{kji}$ etc. You should check that this rule reduces to the above definition. For example, it immediately implies that if there is a repeated index, the answer is zero. This follows because $\varepsilon_{iij} = -\varepsilon_{iij}$ and so $\varepsilon_{iij} = 0$.

It is useful to use the Einstein summation convention when dealing with these symbols. Simply stated, the convention is that you sum over the repeated index. Thus $a_i b_i$ 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 4.5.3 *The following holds.*

$$\varepsilon_{ijk}\varepsilon_{irs} = (\delta_{jr}\delta_{ks} - \delta_{kr}\delta_{js}).$$

Proof: If $\{j,k\} \neq \{r,s\}$ then every term in the sum on the left must have either ε_{ijk} or ε_{irs} contains a repeated index. Therefore, the left side equals zero. The right side also equals zero in this case. To see this, note that if the two sets are not equal, then there is one of the indices in one of the sets which is not in the other set. For example, it could be that j is not equal to either r or s. Then the right side equals zero.

Therefore, it can be assumed $\{j,k\} = \{r,s\}$. If $i = r$ and $j = s$ for $s \neq r$, then there is exactly one term in the sum on the left and it equals 1. The right also reduces to 1 in this case. If $i = s$ and $j = r$, there is exactly one term in the sum on the left which is nonzero and it must equal -1. The right side also reduces to -1 in this case. If there is a repeated index in $\{j,k\}$, then every term in the sum on the left equals zero. The right also reduces to zero in this case because then $j = k = r = s$ and so the right side becomes $(1)(1) - (-1)(-1) = 0$.

Proposition 4.5.4 *Let $\mathbf{u}, \mathbf{v}$ be vectors in $\mathbb{R}^n$ where the Cartesian coordinates of $\mathbf{u}$ are $(u_1, \cdots, u_n)$ and the Cartesian coordinates of $\mathbf{v}$ are $(v_1, \cdots, v_n)$. Then $\mathbf{u} \cdot \mathbf{v} = u_i v_i$. If $\mathbf{u}, \mathbf{v}$ are vectors in $\mathbb{R}^3$, then*

$$(\mathbf{u} \times \mathbf{v})_i = \varepsilon_{ijk} u_j v_k.$$

Also, $\delta_{ik} a_k = a_i$.

Proof: The first claim is obvious from the definition of the dot product. The second is verified by simply checking it works. For example,

$$\mathbf{u} \times \mathbf{v} \equiv \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}$$

and so

$$(\mathbf{u} \times \mathbf{v})_1 = (u_2 v_3 - u_3 v_2).$$

From the above formula in the proposition,

$$\varepsilon_{1jk} u_j v_k \equiv u_2 v_3 - u_3 v_2,$$

the same thing. The cases for $(\mathbf{u} \times \mathbf{v})_2$ and $(\mathbf{u} \times \mathbf{v})_3$ are verified similarly. The last claim follows directly from the definition.

With this notation, you can easily discover vector identities and simplify expressions which involve the cross product.

Example 4.5.5 *Discover a formula which simplifies* $(\mathbf{u}\times\mathbf{v})\times\mathbf{w}$.

From the above reduction formula,

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

Since this holds for all i, it follows that

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

This is good notation and it will be used in the rest of the book whenever convenient. Actually, this notation is a special case of something more elaborate in which the level of the indices is also important, but there is time for this more general notion later. You will see it in advanced books on mechanics in physics and engineering. It also occurs in the subject of differential geometry.

4.6 Exercises

1. Discover a vector identity for $\mathbf{u}\times\left(\mathbf{v}\times\mathbf{w}\right)$.
2. Discover a vector identity for $\left(\mathbf{u}\times\mathbf{v}\right)\cdot\left(\mathbf{z}\times\mathbf{w}\right)$.
3. Discover a vector identity for $\left(\mathbf{u}\times\mathbf{v}\right)\times\left(\mathbf{z}\times\mathbf{w}\right)$ in terms of box products.
4. Simplify $\left(\mathbf{u}\times\mathbf{v}\right)\cdot\left(\mathbf{v}\times\mathbf{w}\right)\times\left(\mathbf{w}\times\mathbf{z}\right)$.
5. Simplify $\left|\mathbf{u}\times\mathbf{v}\right|^2+\left(\mathbf{u}\times\mathbf{v}\right)^2-\left|\mathbf{u}\right|^2\left|\mathbf{v}\right|^2$.
6. Prove that $\varepsilon_{ijk}\varepsilon_{ijr}=2\delta_{kr}$.

Matrices And Linear Transformations

5.1 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 this book 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.

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

This 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 $\begin{pmatrix} 1 \\ 5 \\ 6 \end{pmatrix}$. The convention in dealing with matrices is to always list the rows first and then the columns. Also, you can remember the columns are like columns in a Greek temple. They stand up right while the rows just lay there like rows made by a tractor in a plowed field. Elements of the matrix are identified according to position in the matrix. For example, 8 is in position $2,3$ because it is in the second row and the third column. You might remember that you always list the rows before the columns by using the phrase **Row**man **C**atholic. The symbol, (a_{ij}) refers to a matrix in which the i denotes the row and the j denotes the column. 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. They can sometimes be added, multiplied by a scalar and sometimes multiplied. To illustrate scalar multiplication, consider the following example.

$$3\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 2 & 8 & 7 \\ 6 & -9 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 6 & 9 & 12 \\ 15 & 6 & 24 & 21 \\ 18 & -27 & 3 & 6 \end{pmatrix}.$$

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 which are the same size can be added. When this is done, the result is the

matrix which is obtained by adding 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 5.1.1 *Let $A = (a_{ij})$ and $B = (b_{ij})$ be two $m \times n$ matrices. Then $A + B = C$ where*

$$C = (c_{ij})$$

for $c_{ij} = a_{ij} + b_{ij}$. Also if x is a scalar,

$$xA = (c_{ij})$$

where $c_{ij} = xa_{ij}$. The number A_{ij} will typically refer to the ij^{th} entry of the matrix, A. The zero matrix, denoted by 0 will be the matrix consisting of all zeros.

Do not be upset by the use of the subscripts, ij. The expression $c_{ij} = a_{ij} + b_{ij}$ is just saying that you add corresponding entries to get the result of summing two matrices as discussed above.

Note there are 2×3 zero matrices, 3×4 zero matrices, etc. In fact for every size there is a zero matrix.

With this definition, the following properties are all obvious but you should verify all of these properties are valid for A, B, and C, $m \times n$ matrices and 0 an $m \times n$ zero matrix,

$$A + B = B + A, \tag{5.1}$$

the commutative law of addition,

$$(A + B) + C = A + (B + C), \tag{5.2}$$

the associative law for addition,

$$A + 0 = A, \tag{5.3}$$

the existence of an additive identity,

$$A + (-A) = 0, \tag{5.4}$$

the existence of an additive inverse. Also, for α, β scalars, the following also hold.

$$\alpha (A + B) = \alpha A + \alpha B, \tag{5.5}$$

$$(\alpha + \beta) A = \alpha A + \beta A, \tag{5.6}$$

$$\alpha (\beta A) = \alpha\beta (A), \tag{5.7}$$

$$1A = A. \tag{5.8}$$

The above properties, 5.1 - 5.8 are known as the vector space axioms and the fact that the $m \times n$ matrices satisfy these axioms is what is meant by saying this set of matrices forms a vector space. You may need to study these later.

Definition 5.1.2 *Matrices which are $n \times 1$ or $1 \times n$ are especially called vectors and are often denoted by a bold letter. Thus*

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

is a $n \times 1$ matrix also called a column vector while a $1 \times n$ matrix of the form $(x_1 \cdots x_n)$ is referred to as a row vector.

All the above is fine, but the real reason for considering matrices is that they can be multiplied. This is where things quit being banal.

First consider the problem of multiplying an $m \times n$ matrix by an $n \times 1$ column vector. Consider the following example

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

The way I like to remember this is as follows. Slide the vector, placing it on top the two rows as shown

$$\begin{pmatrix} {\scriptstyle 7} & {\scriptstyle 8} & {\scriptstyle 9} \\ 1 & 2 & 3 \\ {\scriptstyle 7} & {\scriptstyle 8} & {\scriptstyle 9} \\ 4 & 5 & 6 \end{pmatrix},$$

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. These numbers are listed in the same order giving, in this case, a 2×1 matrix. Thus

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 7 \times 1 + 8 \times 2 + 9 \times 3 \\ 7 \times 4 + 8 \times 5 + 9 \times 6 \end{pmatrix} = \begin{pmatrix} 50 \\ 122 \end{pmatrix}.$$

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. Motivated by this example, here is the definition of how to multiply an $m \times n$ matrix by an $n \times 1$ matrix. (vector)

Definition 5.1.3 *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

$$(A\mathbf{v})_i = \sum_{j=1}^{n} A_{ij} v_j.$$

Thus

$$A\mathbf{v} = \begin{pmatrix} \sum_{j=1}^{n} A_{1j} v_j \\ \vdots \\ \sum_{j=1}^{n} A_{mj} v_j \end{pmatrix}. \tag{5.9}$$

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 5.9 and the observation that the j^{th} column of A is

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

so 5.9 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 5.1.4 *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 entry 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

$$\begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 2 & 1 & -2 \\ 2 & 1 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 8 \\ 2 \\ 5 \end{pmatrix}.$$

With this done, 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.

$$(m \times \overbrace{\quad n)\,(n}^{\text{these must match}} \times p \quad) = m \times p$$

If the two middle numbers don't match, you can't multiply the matrices!

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. Then an $m \times p$ matrix, AB is defined as follows:

$$AB \equiv (A\mathbf{b}_1, \cdots, A\mathbf{b}_p) \tag{5.10}$$

where $A\mathbf{b}_k$ is an $m \times 1$ matrix. Hence AB as just defined is an $m \times p$ matrix. For example,

Example 5.1.5 *Multiply the following.*

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

The first thing you need to check before doing anything else is whether it is possible to do the multiplication. The first matrix is a 2×3 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

$$\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 1 \\ -2 & 1 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 9 & 3 \\ -2 & 7 & 3 \end{pmatrix}.$$

Here is another example.

Example 5.1.6 *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!

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

$$\sum_{k=1}^{n} A_{ik}B_{kj}. \tag{5.11}$$

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 **??**.

This motivates the definition for matrix multiplication which identifies the ij^{th} entries of the product.

Definition 5.1.7 *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*

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}. \tag{5.12}$$

Two matrices, A and B are said to be conformable in a particular order if they can be multiplied in that order. Thus if A is an $r \times s$ matrix and B is a $s \times p$ then A and B are conformable in the order, AB.

Example 5.1.8 *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, it must be possible to do this and 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 5.1.9 *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.

Example 5.1.10 *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. 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 5.1.11 *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

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

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 they be equal then?

Example 5.1.12 *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

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

and you see these are not equal. Therefore, you cannot conclude that $AB = BA$ for matrix multiplication. However, there are some properties which do hold.

Proposition 5.1.13 *If all multiplications and additions make sense, the following hold for matrices,* A, B, C *and* a, b *scalars.*

$$A(aB+bC) = a(AB) + b(AC) \tag{5.13}$$

$$(B+C)A = BA + CA \tag{5.14}$$

$$A(BC) = (AB)C \tag{5.15}$$

Proof: Using the repeated index summation convention and the above definition of matrix multiplication,

$$\begin{aligned} (A(aB+bC))_{ij} &= \sum_k A_{ik}(aB+bC)_{kj} \\ &= \sum_k A_{ik}(aB_{kj} + bC_{kj}) \\ &= a\sum_k A_{ik}B_{kj} + b\sum_k A_{ik}C_{kj} \\ &= a(AB)_{ij} + b(AC)_{ij} \\ &= (a(AB) + b(AC))_{ij} \end{aligned}$$

showing that $A(B+C) = AB + AC$ as claimed. Formula 5.14 is entirely similar.

Consider 5.15, the associative law of multiplication. Before reading this, review the definition of matrix multiplication in terms of entries of the matrices.

$$\begin{aligned}(A(BC))_{ij} &= \sum_k A_{ik}(BC)_{kj} \\ &= \sum_k A_{ik}\sum_l B_{kl}C_{lj} \\ &= \sum_l (AB)_{il}C_{lj} \\ &= ((AB)C)_{ij}.\end{aligned}$$

This proves 5.15.

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.

$$\begin{pmatrix} 1 & 1+2i \\ 3 & 1 \\ 2 & 6 \end{pmatrix}^T = \begin{pmatrix} 1 & 3 & 2 \\ 1+2i & 1 & 6 \end{pmatrix}$$

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. This motivates the following definition of the transpose of a matrix.

Definition 5.1.14 *Let A be an $m\times n$ matrix. Then A^T denotes the $n\times m$ matrix which is defined as follows.*

$$\left(A^T\right)_{ij} = A_{ji}$$

The transpose of a matrix has the following important property.

Lemma 5.1.15 *Let A be an $m\times n$ matrix and let B be a $n\times p$ matrix. Then*

$$(AB)^T = B^T A^T \tag{5.16}$$

and if α and β are scalars,

$$(\alpha A + \beta B)^T = \alpha A^T + \beta B^T \tag{5.17}$$

Proof: From the definition,

$$\begin{aligned}\left((AB)^T\right)_{ij} &= (AB)_{ji} \\ &= \sum_k A_{jk}B_{ki} \\ &= \sum_k \left(B^T\right)_{ik}\left(A^T\right)_{kj} \\ &= \left(B^T A^T\right)_{ij}\end{aligned}$$

5.17 is left as an exercise and this proves the lemma.

Definition 5.1.16 *An $n\times n$ matrix, A is said to be symmetric if $A = A^T$. It is said to be skew symmetric if $A^T = -A$.*

Example 5.1.17 *Let*

$$A = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 5 & -3 \\ 3 & -3 & 7 \end{pmatrix}.$$

Then A is symmetric.

Example 5.1.18 *Let*

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

Then A is skew symmetric.

There is a special matrix called I and defined by

$$I_{ij} = \delta_{ij}$$

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 5.1.19 *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 5.1.20 *An $n \times n$ matrix, A has an inverse, A^{-1} if and only if $AA^{-1} = A^{-1}A = I$ where $I = (\delta_{ij})$ for*

$$\delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

Such a matrix is called invertible.

5.1.1 Finding The Inverse Of A Matrix

A little later a formula is given for the inverse of a matrix. However, it is not a good way to find the inverse for a matrix. There is a much easier way and it is this which is presented here. It is also important to note that not all matrices have inverses.

Example 5.1.21 *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

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = A^{-1}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) = A^{-1}\left(A\begin{pmatrix} -1 \\ 1 \end{pmatrix}\right) =$$

$$= (A^{-1}A)\begin{pmatrix} -1 \\ 1 \end{pmatrix} = I\begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

a contradiction. Thus the answer is that A does not have an inverse.

Example 5.1.22 *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.

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

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix} \tag{5.18}$$

for the first system and

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \end{pmatrix} \tag{5.19}$$

for the second. Lets solve the first system. Take (-1) times the first row and add to the second to get

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

Now take (-1) times the second row and add to the first to get

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

Putting in the variables, this says $x = 2$ and $y = -1$.

Now solve the second system, 5.19 to find z and w. Take (-1) times the first row and add to the second to get

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

Now take (-1) times the second row and add to the first to get

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

Putting in the variables, this says $z = -1$ and $w = 1$. Therefore, the inverse is

$$\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}.$$

Didn't the above seem rather repetitive? Note that exactly the same row operations were used in both systems. In each case, the end result was something of the form $(I|\mathbf{v})$ 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 $\begin{pmatrix} z \\ w \end{pmatrix}$.

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 5.1.23 *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 do row operations until you obtain an $n \times 2n$ matrix of the form

$$(I|B) \tag{5.20}$$

if possible. When this has been done, $B = A^{-1}$. The matrix, A has no inverse exactly when it is impossible to do row operations and end up with one like 5.20.

Example 5.1.24 *Let* $A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}$. *Find* A^{-1}.

Form the augmented matrix,

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

Now do row operations untill the $n \times n$ matrix on the left becomes the identity matrix. This yields after some computations,

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

and so the inverse of A is the matrix on the right,

$$\begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & -1 & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix}.$$

Checking the answer is easy. Just multiply the matrices and see if it works.

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & -1 & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Always check your answer because if you are like some of us, you will usually have made a mistake.

Example 5.1.25 *Let* $A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 3 & 1 & -1 \end{pmatrix}$. *Find* A^{-1}.

Set up the augmented matrix, $(A|I)$

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

Next take (-1) times the first row and add to the second followed by (-3) times the first row added to the last. This yields

$$\begin{pmatrix} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -2 & 0 & -1 & 1 & 0 \\ 0 & -5 & -7 & -3 & 0 & 1 \end{pmatrix}.$$

Then take 5 times the second row and add to -2 times the last row.

$$\begin{pmatrix} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{pmatrix}$$

Next take the last row and add to (-7) times the top row. This yields

$$\begin{pmatrix} -7 & -14 & 0 & -6 & 5 & -2 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{pmatrix}.$$

Now take $(-7/5)$ times the second row and add to the top.

$$\begin{pmatrix} -7 & 0 & 0 & 1 & -2 & -2 \\ 0 & -10 & 0 & -5 & 5 & 0 \\ 0 & 0 & 14 & 1 & 5 & -2 \end{pmatrix}.$$

Finally divide the top row by -7, the second row by -10 and the bottom row by 14 which yields

$$\begin{pmatrix} 1 & 0 & 0 & -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ 0 & 1 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{pmatrix}.$$

Therefore, the inverse is

$$\begin{pmatrix} -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{pmatrix}$$

Example 5.1.26 *Let* $A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 4 \end{pmatrix}$. *Find* A^{-1}.

Write the augmented matrix, $(A|I)$

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

and proceed to do row operations attempting to obtain $\left(I|A^{-1}\right)$. Take (-1) times the top row and add to the second. Then take (-2) times the top row and add to the bottom.

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

Next add (-1) times the second row to the bottom row.

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

At this point, you can see there will be no inverse because you have obtained a row of zeros in the left half of the augmented matrix, $(A|I)$. Thus there will be no way to obtain I on the left. In other words, the three systems of equations you must solve to find the inverse have no solution. In particular, there is no solution for the first column of A^{-1} which must solve

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

because a sequence of row operations leads to the impossible equation, $0x + 0y + 0z = -1$.

5.2 Exercises

1. In 5.1 - 5.8 describe $-A$ and 0.
2. Let A be an $n \times n$ matrix. Show A equals the sum of a symmetric and a skew symmetric matrix.
3. Show every skew symmetric matrix has all zeros down the main diagonal. The main diagonal consists of every entry of the matrix which is of the form a_{ii}. It runs from the upper left down to the lower right.
4. Using only the properties 5.1 - 5.8 show $-A$ is unique.
5. Using only the properties 5.1 - 5.8 show 0 is unique.
6. Using only the properties 5.1 - 5.8 show $0A = 0$. Here the 0 on the left is the scalar 0 and the 0 on the right is the zero for $m \times n$ matrices.
7. Using only the properties 5.1 - 5.8 and previous problems show $(-1)\, A = -A$.
8. Prove 5.17.

9. Prove that $I_m A = A$ where A is an $m \times n$ matrix.

10. Let A and be a real $m \times n$ matrix and let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Show $(A\mathbf{x}, \mathbf{y})_{\mathbb{R}^m} = \left(\mathbf{x}, A^T\mathbf{y}\right)_{\mathbb{R}^n}$ where $(\cdot,\cdot)_{\mathbb{R}^k}$ denotes the dot product in $\mathbb{R}^k$.

11. Use the result of Problem 10 to verify directly that $(AB)^T = B^T A^T$ without making any reference to subscripts.

12. Let $\mathbf{x} = (-1,-1,1)$ and $\mathbf{y} = (0,1,2)$. Find $\mathbf{x}^T\mathbf{y}$ and $\mathbf{x}\mathbf{y}^T$ if possible.

13. Give an example of matrices, A, B, C such that $B \neq C$, $A \neq 0$, and yet $AB = AC$.

14. Let $A = \begin{pmatrix} 1 & 1 \\ -2 & -1 \\ 1 & 2 \end{pmatrix}$, $B = \begin{pmatrix} 1 & -1 & -2 \\ 2 & 1 & -2 \end{pmatrix}$, and $C = \begin{pmatrix} 1 & 1 & -3 \\ -1 & 2 & 0 \\ -3 & -1 & 0 \end{pmatrix}$. Find if possible.

    (a) AB
    (b) BA
    (c) AC
    (d) CA
    (e) CB
    (f) BC

15. 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 $B = A^{-1}$.

16. Show $(AB)^{-1} = B^{-1}A^{-1}$.

17. Show that if A is an invertible $n \times n$ matrix, then so is A^T and $\left(A^T\right)^{-1} = \left(A^{-1}\right)^T$.

18. Show that if A is an $n \times n$ invertible matrix and $\mathbf{x}$ is a $n \times 1$ matrix such that $A\mathbf{x} = \mathbf{b}$ for $\mathbf{b}$ an $n \times 1$ matrix, then $\mathbf{x} = A^{-1}\mathbf{b}$.

19. Give an example of a matrix, A such that $A^2 = I$ and yet $A \neq I$ and $A \neq -I$.

20. Give an example of matrices, A, B such that neither A nor B equals zero and yet $AB = 0$.

21. Write $\begin{pmatrix} x_1 - x_2 + 2x_3 \\ 2x_3 + x_1 \\ 3x_3 \\ 3x_4 + 3x_2 + x_1 \end{pmatrix}$ in the form $A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$ where A is an appropriate matrix.

22. Give another example other than the one given in this section of two square matrices, A and B such that $AB \neq BA$.

23. Suppose A and B are square matrices of the same size. Which of the following are correct?

    (a) $(A-B)^2 = A^2 - 2AB + B^2$
    (b) $(AB)^2 = A^2B^2$
    (c) $(A+B)^2 = A^2 + 2AB + B^2$

(d) $(A+B)^2 = A^2 + AB + BA + B^2$

(e) $A^2B^2 = A(AB)B$

(f) $(A+B)^3 = A^3 + 3A^2B + 3AB^2 + B^3$

(g) $(A+B)(A-B) = A^2 - B^2$

(h) None of the above. They are all wrong.

(i) All of the above. They are all right.

24. Let $A = \begin{pmatrix} -1 & -1 \\ 3 & 3 \end{pmatrix}$. Find all 2×2 matrices, B such that $AB = 0$.

25. Prove that if A^{-1} exists and $A\mathbf{x} = \mathbf{0}$ then $\mathbf{x} = \mathbf{0}$.

26. Let
$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 1 & 0 & 2 \end{pmatrix}.$$
Find A^{-1} if possible. If A^{-1} does not exist, determine why.

27. Let
$$A = \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}.$$
Find A^{-1} if possible. If A^{-1} does not exist, determine why.

28. Let
$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 4 & 5 & 10 \end{pmatrix}.$$
Find A^{-1} if possible. If A^{-1} does not exist, determine why.

29. Let
$$A = \begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 1 & 2 & 0 \\ 2 & 1 & -3 & 2 \\ 1 & 2 & 1 & 2 \end{pmatrix}$$
Find A^{-1} if possible. If A^{-1} does not exist, determine why.

5.3 Linear Transformations

By 5.13, if A is an $m \times n$ matrix, then for $\mathbf{v}, \mathbf{u}$ vectors in $\mathbb{F}^n$ and a, b scalars,

$$A\left(\overbrace{a\mathbf{u} + b\mathbf{v}}^{\in \mathbb{F}^n}\right) = aA\mathbf{u} + bA\mathbf{v} \in \mathbb{F}^m \tag{5.21}$$

Definition 5.3.1 *A function, $A : \mathbb{F}^n \to \mathbb{F}^m$ is called a linear transformation if for all $\mathbf{u}, \mathbf{v} \in \mathbb{F}^n$ and a, b scalars, 5.21 holds.*

From 5.21, matrix multiplication defines a linear transformation as just defined. It turns out this is the only type of linear transformation available. Thus if A is a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$, there is always a matrix which produces A. Before showing this, here is a simple definition.

Definition 5.3.2 *A vector, $\mathbf{e}_i \in \mathbb{F}^n$ is defined as follows:*

$$\mathbf{e}_i \equiv \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix},$$

where the 1 is in the i^{th} position and there are zeros everywhere else. Thus

$$\mathbf{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)^T.$$

Of course the $\mathbf{e}_i$ for a particular value of i in $\mathbb{F}^n$ would be different than the $\mathbf{e}_i$ for that same value of i in $\mathbb{F}^m$ for $m \neq n$. One of them is longer than the other. However, which one is meant will be determined by the context in which they occur.

These vectors have a significant property.

Lemma 5.3.3 *Let $\mathbf{v} \in \mathbb{F}^n$. Thus $\mathbf{v}$ is a list of numbers arranged vertically, $v_1, \cdots, v_n$. Then*

$$\mathbf{e}_i^T \mathbf{v} = v_i. \tag{5.22}$$

Also, if A is an $m \times n$ matrix, then letting $\mathbf{e}_i \in \mathbb{F}^m$ and $\mathbf{e}_j \in \mathbb{F}^n$,

$$\mathbf{e}_i^T A \mathbf{e}_j = A_{ij} \tag{5.23}$$

Proof: First note that $\mathbf{e}_i^T$ is a $1 \times n$ matrix and $\mathbf{v}$ is an $n \times 1$ matrix so the above multiplication in 5.22 makes perfect sense. It equals

$$(0, \cdots, 1, \cdots 0) \begin{pmatrix} v_1 \\ \vdots \\ v_i \\ \vdots \\ v_n \end{pmatrix} = v_i$$

as claimed.

Consider 5.23. From the definition of matrix multiplication using the repeated index summation convention, and noting that $(\mathbf{e}_j)_k = \delta_{kj}$

$$\mathbf{e}_i^T A \mathbf{e}_j = \mathbf{e}_i^T \begin{pmatrix} A_{1k} (\mathbf{e}_j)_k \\ \vdots \\ A_{ik} (\mathbf{e}_j)_k \\ \vdots \\ A_{mk} (\mathbf{e}_j)_k \end{pmatrix} = \mathbf{e}_i^T \begin{pmatrix} A_{1j} \\ \vdots \\ A_{ij} \\ \vdots \\ A_{mj} \end{pmatrix} = A_{ij}$$

by the first part of the lemma. This proves the lemma.

Theorem 5.3.4 *Let $L : \mathbb{F}^n \to \mathbb{F}^m$ be a linear transformation. Then there exists a unique $m \times n$ matrix, A such that*

$$A\mathbf{x} = L\mathbf{x}$$

for all $\mathbf{x} \in \mathbb{F}^n$. The ik^{th} entry of this matrix is given by

$$\mathbf{e}_i^T L\mathbf{e}_k \tag{5.24}$$

Proof: By the lemma,

$$(L\mathbf{x})_i = \mathbf{e}_i^T L\mathbf{x} = \mathbf{e}_i^T x_k L\mathbf{e}_k = \left(\mathbf{e}_i^T L\mathbf{e}_k\right) x_k.$$

Let $A_{ik} = \mathbf{e}_i^T L\mathbf{e}_k$, to prove the existence part of the theorem.

To verify uniqueness, suppose $B\mathbf{x} = A\mathbf{x} = L\mathbf{x}$ for all $\mathbf{x} \in \mathbb{F}^n$. Then in particular, this is true for $\mathbf{x} = \mathbf{e}_j$ and then multiply on the left by $\mathbf{e}_i^T$ to obtain

$$B_{ij} = \mathbf{e}_i^T B\mathbf{e}_j = \mathbf{e}_i^T A\mathbf{e}_j = A_{ij}$$

showing $A = B$. This proves uniqueness.

Corollary 5.3.5 *A linear transformation, $L : \mathbb{F}^n \to \mathbb{F}^m$ is completely determined by the vectors $\{L\mathbf{e}_1, \cdots, L\mathbf{e}_n\}$.*

Proof: This follows immediately from the above theorem. The unique matrix determining the linear transformation which is given in 5.24 depends only on these vectors.

This theorem shows that any linear transformation defined on $\mathbb{F}^n$ can always be considered as a matrix. Therefore, the terms "linear transformation" and "matrix" will be used interchangeably. For example, to say a matrix is one to one, means the linear transformation determined by the matrix is one to one.

Example 5.3.6 *Find the linear transformation, $L : \mathbb{R}^2 \to \mathbb{R}^2$ which has the property that $L\mathbf{e}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $L\mathbf{e}_2 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$. From the above theorem and corollary, this linear transformation is that determined by matrix multiplication by the matrix*

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

5.4 Subspaces And Spans

Definition 5.4.1 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ be vectors in $\mathbb{F}^n$. A linear combination is any expression of the form*

$$\sum_{i=1}^{p} c_i \mathbf{x}_i$$

where the c_i are scalars. The set of all linear combinations of these vectors is called $\text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$. If $V \subseteq \mathbb{F}^n$, then V is called a subspace if whenever α, β are scalars and $\mathbf{u}$ and $\mathbf{v}$ are vectors of V, it follows $\alpha\mathbf{u} + \beta\mathbf{v} \in V$. That is, it is "closed under the algebraic operations of vector addition and scalar multiplication". A linear combination of vectors is said to be trivial if all the scalars in the linear combination equal zero. A set of vectors is said to be linearly independent if the only linear combination of these vectors which equals the zero vector is the trivial linear combination. Thus $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ is called linearly independent if whenever

$$\sum_{k=1}^{p} c_k \mathbf{x}_k = \mathbf{0}$$

it follows that all the scalars, c_k equal zero. A set of vectors, $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$, is called linearly dependent if it is not linearly independent. Thus the set of vectors is linearly dependent if there exist scalars, $c_i, i = 1, \cdots, n$, not all zero such that $\sum_{k=1}^{p} c_k \mathbf{x}_k = \mathbf{0}$.

Lemma 5.4.2 *A set of vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ is linearly independent if and only if none of the vectors can be obtained as a linear combination of the others.*

Proof: Suppose first that $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ is linearly independent. If $\mathbf{x}_k = \sum_{j \neq k} c_j \mathbf{x}_j$, then

$$\mathbf{0} = 1\mathbf{x}_k + \sum_{j \neq k} (-c_j)\, \mathbf{x}_j,$$

a nontrivial linear combination, contrary to assumption. This shows that if the set is linearly independent, then none of the vectors is a linear combination of the others.

Now suppose no vector is a linear combination of the others. Is $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ linearly independent? If it is not there exist scalars, c_i, not all zero such that

$$\sum_{i=1}^{p} c_i \mathbf{x}_i = \mathbf{0}.$$

Say $c_k \neq 0$. Then you can solve for $\mathbf{x}_k$ as

$$\mathbf{x}_k = \sum_{j \neq k} (-c_j)\, / c_k \mathbf{x}_j$$

contrary to assumption. This proves the lemma.

The following is called the exchange theorem.

Theorem 5.4.3 *(Exchange Theorem) Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ be a linearly independent set of vectors such that each $\mathbf{x}_i$ is in span$(\mathbf{y}_1, \cdots, \mathbf{y}_s)$. Then $r \leq s$.*

Proof: Define span$\{\mathbf{y}_1, \cdots, \mathbf{y}_s\} \equiv V$, it follows there exist scalars, $c_1, \cdots, c_s$ such that

$$\mathbf{x}_1 = \sum_{i=1}^{s} c_i \mathbf{y}_i. \tag{5.25}$$

Not all of these scalars can equal zero because if this were the case, it would follow that $\mathbf{x}_1 = 0$ and so $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ would not be linearly independent. Indeed, if $\mathbf{x}_1 = \mathbf{0}$, $1\mathbf{x}_1 + \sum_{i=2}^{r} 0\mathbf{x}_i = \mathbf{x}_1 = \mathbf{0}$ and so there would exist a nontrivial linear combination of the vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ which equals zero.

Say $c_k \neq 0$. Then solve (5.25) for $\mathbf{y}_k$ and obtain

$$\mathbf{y}_k \in \text{span}\left(\mathbf{x}_1, \overbrace{\mathbf{y}_1, \cdots, \mathbf{y}_{k-1}, \mathbf{y}_{k+1}, \cdots, \mathbf{y}_s}^{\text{s-1 vectors here}}\right).$$

Define $\{\mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$ by

$$\{\mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\} \equiv \{\mathbf{y}_1, \cdots, \mathbf{y}_{k-1}, \mathbf{y}_{k+1}, \cdots, \mathbf{y}_s\}$$

Therefore, span $\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\} = V$ because if $\mathbf{v} \in V$, there exist constants $c_1, \cdots, c_s$ such that

$$\mathbf{v} = \sum_{i=1}^{s-1} c_i \mathbf{z}_i + c_s \mathbf{y}_k.$$

Now replace the $\mathbf{y}_k$ in the above with a linear combination of the vectors, $\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$ to obtain $\mathbf{v} \in \text{span}\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$. The vector $\mathbf{y}_k$, in the list $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$, has now been replaced with the vector $\mathbf{x}_1$ and the resulting modified list of vectors has the same span as the original list of vectors, $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$.

Now suppose that $r > s$ and that $\text{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{z}_1, \cdots, \mathbf{z}_p\} = V$ where the vectors, $\mathbf{z}_1, \cdots, \mathbf{z}_p$ are each taken from the set, $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$ and $l + p = s$. This has now been done for $l = 1$ above. Then since $r > s$, it follows that $l \leq s < r$ and so $l + 1 \leq r$. Therefore, $\mathbf{x}_{l+1}$ is a vector not in the list, $\{\mathbf{x}_1, \cdots, \mathbf{x}_l\}$ and since $\text{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{z}_1, \cdots, \mathbf{z}_p\} = V$, there exist scalars, c_i and d_j such that

$$\mathbf{x}_{l+1} = \sum_{i=1}^{l} c_i \mathbf{x}_i + \sum_{j=1}^{p} d_j \mathbf{z}_j. \tag{5.26}$$

Now not all the d_j can equal zero because if this were so, it would follow that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ would be a linearly dependent set because one of the vectors would equal a linear combination of the others. Therefore, (5.26) can be solved for one of the $\mathbf{z}_i$, say $\mathbf{z}_k$, in terms of $\mathbf{x}_{l+1}$ and the other $\mathbf{z}_i$ and just as in the above argument, replace that $\mathbf{z}_i$ with $\mathbf{x}_{l+1}$ to obtain

$$\text{span}\left\{\mathbf{x}_1, \cdots \mathbf{x}_l, \mathbf{x}_{l+1}, \overbrace{\mathbf{z}_1, \cdots \mathbf{z}_{k-1}, \mathbf{z}_{k+1}, \cdots, \mathbf{z}_p}^{\text{p-1 vectors here}}\right\} = V.$$

Continue this way, eventually obtaining

$$\text{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_s\} = V.$$

But then $\mathbf{x}_r \in \text{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_s\}$ contrary to the assumption that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is linearly independent. Therefore, $r \leq s$ as claimed.

Definition 5.4.4 *A finite set of vectors, $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is a basis for $\mathbb{F}^n$ if $\text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_r) = \mathbb{F}^n$ and $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is linearly independent.*

Corollary 5.4.5 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ and $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$ be two bases[1] of $\mathbb{F}^n$. Then $r = s = n$.*

Proof: From the exchange theorem, $r \leq s$ and $s \leq r$. Now note the vectors,

$$\mathbf{e}_i = \overbrace{(0, \cdots, 0, 1, 0 \cdots, 0)}^{1 \text{ is in the } i^{th} \text{ slot}}$$

for $i = 1, 2, \cdots, n$ are a basis for $\mathbb{F}^n$. This proves the corollary.

Lemma 5.4.6 *Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ be a set of vectors. Then $V \equiv \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_r)$ is a subspace.*

Proof: Suppose α, β are two scalars and let $\sum_{k=1}^{r} c_k \mathbf{v}_k$ and $\sum_{k=1}^{r} d_k \mathbf{v}_k$ are two elements of V. What about

$$\alpha \sum_{k=1}^{r} c_k \mathbf{v}_k + \beta \sum_{k=1}^{r} d_k \mathbf{v}_k?$$

Is it also in V?

$$\alpha \sum_{k=1}^{r} c_k \mathbf{v}_k + \beta \sum_{k=1}^{r} d_k \mathbf{v}_k = \sum_{k=1}^{r} (\alpha c_k + \beta d_k) \mathbf{v}_k \in V$$

so the answer is yes. This proves the lemma.

[1] This is the plural form of basis. We could say basiss but it would involve an inordinate amount of hissing as in "The sixth shiek's sixth sheep is sick". This is the reason that bases is used instead of basiss.

Definition 5.4.7 *A finite set of vectors,* $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ *is a basis for a subspace,* V *of* $\mathbb{F}^n$ *if* $\mathrm{span}(\mathbf{x}_1, \cdots, \mathbf{x}_r) = V$ *and* $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ *is linearly independent.*

Corollary 5.4.8 *Let* $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ *and* $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$ *be two bases for* V*. Then* $r = s$*.*

Proof: From the exchange theorem, $r \leq s$ and $s \leq r$. Therefore, this proves the corollary.

Definition 5.4.9 *Let* V *be a subspace of* $\mathbb{F}^n$*. Then* $\dim(V)$ *read as the dimension of* V *is the number of vectors in a basis.*

Of course you should wonder right now whether an arbitrary subspace even has a basis. In fact it does and this is in the next theorem. First, here is an interesting lemma.

Lemma 5.4.10 *Suppose* $\mathbf{v} \notin \mathrm{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ *and* $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ *is linearly independent. Then* $\{\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{v}\}$ *is also linearly independent.*

Proof: Suppose $\sum_{i=1}^k c_i \mathbf{u}_i + d\mathbf{v} = \mathbf{0}$. It is required to verify that each $c_i = 0$ and that $d = 0$. But if $d \neq 0$, then you can solve for $\mathbf{v}$ as a linear combination of the vectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$,

$$\mathbf{v} = -\sum_{i=1}^k \left(\frac{c_i}{d}\right) \mathbf{u}_i$$

contrary to assumption. Therefore, $d = 0$. But then $\sum_{i=1}^k c_i \mathbf{u}_i = 0$ and the linear independence of $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ implies each $c_i = 0$ also. This proves the lemma.

Theorem 5.4.11 *Let* V *be a nonzero subspace of* $\mathbb{F}^n$*. Then* V *has a basis.*

Proof: Let $\mathbf{v}_1 \in V$ where $\mathbf{v}_1 \neq 0$. If $\mathrm{span}\{\mathbf{v}_1\} = V$, stop. $\{\mathbf{v}_1\}$ is a basis for V. Otherwise, there exists $\mathbf{v}_2 \in V$ which is not in $\mathrm{span}\{\mathbf{v}_1\}$. By Lemma 5.4.10 $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a linearly independent set of vectors. If $\mathrm{span}\{\mathbf{v}_1, \mathbf{v}_2\} = V$ stop, $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a basis for V. If $\mathrm{span}\{\mathbf{v}_1, \mathbf{v}_2\} \neq V$, then there exists $\mathbf{v}_3 \notin \mathrm{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a larger linearly independent set of vectors. Continuing this way, the process must stop before $n+1$ steps because if not, it would be possible to obtain $n+1$ linearly independent vectors contrary to the exchange theorem. This proves the theorem.

In words the following corollary states that any linearly independent set of vectors can be enlarged to form a basis.

Corollary 5.4.12 *Let* V *be a subspace of* $\mathbb{F}^n$ *and let* $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ *be a linearly independent set of vectors in* V*. Then either it is a basis for* V *or there exist vectors,* $\mathbf{v}_{r+1}, \cdots, \mathbf{v}_s$ *such that* $\{\mathbf{v}_1, \cdots, \mathbf{v}_r, \mathbf{v}_{r+1}, \cdots, \mathbf{v}_s\}$ *is a basis for* V*.*

Proof: This follows immediately from the proof of Theorem 5.4.11. You do exactly the same argument except you start with $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ rather than $\{\mathbf{v}_1\}$.

It is also true that any spanning set of vectors can be restricted to obtain a basis.

Theorem 5.4.13 *Let* V *be a subspace of* $\mathbb{F}^n$ *and suppose* $\mathrm{span}(\mathbf{u}_1 \cdots, \mathbf{u}_p) = V$ *where the* $\mathbf{u}_i$ *are nonzero vectors. Then there exist vectors,* $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ *such that* $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\} \subseteq \{\mathbf{u}_1 \cdots, \mathbf{u}_p\}$ *and* $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ *is a basis for* V*.*

Proof: Let r be the smallest positive integer with the property that for some set, $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\} \subseteq \{\mathbf{u}_1 \cdots, \mathbf{u}_p\}$,

$$\mathrm{span}(\mathbf{v}_1 \cdots, \mathbf{v}_r) = V.$$

Then $r \leq p$ and it must be the case that $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ is linearly independent because if it were not so, one of the vectors, say $\mathbf{v}_k$ would be a linear combination of the others. But then you could delete this vector from $\{\mathbf{v}_1 \cdots, \mathbf{v}_r\}$ and the resulting list of $r-1$ vectors would still span V contrary to the definition of r. This proves the theorem.

5.5 An Application To Matrices

The following is a theorem of major significance.

Theorem 5.5.1 *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

$$\sum_{k=1}^{n} c_k \mathbf{e}_k = \mathbf{0}^2$$

which implies each $c_k = 0$. Therefore, $\{A\mathbf{e}_1, \cdots, A\mathbf{e}_n\}$ must be a basis for $\mathbb{F}^n$ because if not there would exist a vector, $\mathbf{y} \notin \text{span}\,(A\mathbf{e}_1, \cdots, A\mathbf{e}_n)$ and then by Lemma 5.4.10, $\{A\mathbf{e}_1, \cdots, A\mathbf{e}_n, \mathbf{y}\}$ would be an independent set of vectors having $n+1$ vectors in it, contrary to the exchange theorem. It follows that for $\mathbf{y} \in \mathbb{F}^n$ there exist constants, c_i such that

$$\mathbf{y} = \sum_{k=1}^{n} c_k A\mathbf{e}_k = A\left(\sum_{k=1}^{n} c_k \mathbf{e}_k\right)$$

showing that, since $\mathbf{y}$ was arbitrary, A is onto.

Next suppose A is onto. This means the span of the columns of A equals $\mathbb{F}^n$. If these columns are not linearly independent, then by Lemma 5.4.2 on Page 71, one of the columns is a linear combination of the others and so the span of the columns of A equals the span of the $n-1$ other columns. This violates the exchange theorem because $\{\mathbf{e}_1, \cdots, \mathbf{e}_n\}$ would be a linearly independent set of vectors contained in the span of only $n-1$ vectors. Therefore, the columns of A must be independent and this equivalent to saying that $A\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{0}$. This implies A is one to one because if $A\mathbf{x} = A\mathbf{y}$, then $A(\mathbf{x} - \mathbf{y}) = \mathbf{0}$ and so $\mathbf{x} - \mathbf{y} = \mathbf{0}$.

Now suppose $AB = I$. Why is $BA = I$? Since $AB = I$ it follows B is one to one since otherwise, there would exist, $\mathbf{x} \neq \mathbf{0}$ such that $B\mathbf{x} = \mathbf{0}$ and then $AB\mathbf{x} = A\mathbf{0} = \mathbf{0} \neq I\mathbf{x}$. Therefore, from what was just shown, B is also onto. In addition to this, A must be one to one because if $A\mathbf{y} = \mathbf{0}$, then $\mathbf{y} = B\mathbf{x}$ for some $\mathbf{x}$ and then $\mathbf{x} = AB\mathbf{x} = A\mathbf{y} = \mathbf{0}$ showing $\mathbf{y} = \mathbf{0}$. Now from what is given to be so, it follows $(AB)A = A$ and so using the associative law for matrix multiplication,

$$A(BA) - A = A(BA - I) = 0.$$

But this means $(BA - I)\mathbf{x} = \mathbf{0}$ for all $\mathbf{x}$ since otherwise, A would not be one to one. Hence $BA = I$ as claimed. This proves the theorem.

This theorem shows that if an $n \times n$ matrix, B acts like an inverse when multiplied on one side of A it follows that $B = A^{-1}$and it will act like an inverse on both sides of A.

The conclusion of this theorem pertains to square matrices only. For example, let

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \; B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & -1 \end{pmatrix} \tag{5.27}$$

Then

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

but

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

5.6 Matrices And Calculus

The study of moving coordinate systems gives a non trivial example of the usefulness of the ideas involving linear transformations and matrices. To begin with, here is the concept of the product rule extended to matrix multiplication.

Definition 5.6.1 *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 5.6.2 *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*

$$(A(t)B(t))' = A'(t)B(t) + A(t)B'(t).$$

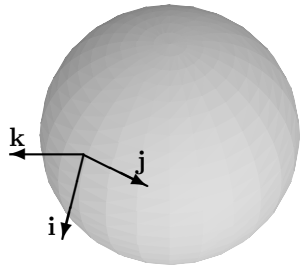
Proof: $(A(t)B(t))' = \left(C'_{ij}(t)\right)$ where $C_{ij}(t) = A_{ik}(t)B_{kj}(t)$ and the repeated index summation convention is being used. Therefore,

$$\begin{aligned} C'_{ij}(t) &= A'_{ik}(t)B_{kj}(t) + A_{ik}(t)B'_{kj}(t) \\ &= (A'(t)B(t))_{ij} + (A(t)B'(t))_{ij} \\ &= (A'(t)B(t) + A(t)B'(t))_{ij} \end{aligned}$$

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.

5.6.1 The Coriolis Acceleration

Imagine a point on the surface of the earth. Now consider unit vectors, one pointing South, one pointing East and one pointing directly away from the center of the earth.



Denote the first as $\mathbf{i}$, the second as $\mathbf{j}$ and the third as $\mathbf{k}$. If you are standing on the earth you will consider these vectors as fixed, but of course they are not. As the earth turns, they change direction and so each is in reality a function of t. Nevertheless, it is with respect to these apparently fixed vectors that you wish to understand acceleration, velocities, and displacements.

In general, let $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$ be the usual fixed vectors in space and let $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ be an orthonormal basis of vectors for each t, like the vectors described in the first paragraph. It is assumed these vectors are C^1 functions of t. Letting the positive x axis extend in the direction of $\mathbf{i}(t)$, the positive y axis extend in the direction of $\mathbf{j}(t)$, and the positive z axis extend in the direction of $\mathbf{k}(t)$, yields a moving coordinate system. Now let $\mathbf{u}$ be a vector 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.

This is exactly the situation in the case of the 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 the 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 $\mathbf{v}$ be a vector defined in the same manner as $\mathbf{u}$ and α, β, scalars,

$$Q(t)(\alpha\mathbf{u} + \beta\mathbf{v}) \equiv \left(\alpha u^1 + \beta v^1\right)\mathbf{i}(t) + \left(\alpha u^2 + \beta v^2\right)\mathbf{j}(t) + \left(\alpha u^3 + \beta v^3\right)\mathbf{k}(t)$$

$$\begin{aligned} &= \left(\alpha u^1 \mathbf{i}(t) + \alpha u^2 \mathbf{j}(t) + \alpha u^3 \mathbf{k}(t)\right) + \left(\beta v^1 \mathbf{i}(t) + \beta v^2 \mathbf{j}(t) + \beta v^3 \mathbf{k}(t)\right) \\ &= \alpha\left(u^1 \mathbf{i}(t) + u^2 \mathbf{j}(t) + u^3 \mathbf{k}(t)\right) + \beta\left(v^1 \mathbf{i}(t) + v^2 \mathbf{j}(t) + v^3 \mathbf{k}(t)\right) \\ &\equiv \alpha Q(t)\mathbf{u} + \beta Q(t)\mathbf{v} \end{aligned}$$

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,

$$|Q(t)\mathbf{u}| = \left(\sum_{i=1}^{3} \left(u^i\right)^2\right)^{1/2} = |\mathbf{u}|.$$

Lemma 5.6.3 *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 5.6.2 that

$$Q'(t)Q(t)^T + Q(t)Q'(t)^T = 0$$

and so

$$Q'(t)Q(t)^T = -\left(Q'(t)Q(t)^T\right)^T. \tag{5.28}$$

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 fixed in space orthonormal basis vectors, $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$, where these are the usual basis vectors for $\mathbb{R}^3$, it follows from 5.28 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,

$$\begin{pmatrix} u^1 \\ u^2 \\ u^3 \end{pmatrix}'(t) = \begin{pmatrix} 0 & -\omega_3(t) & \omega_2(t) \\ \omega_3(t) & 0 & -\omega_1(t) \\ -\omega_2(t) & \omega_1(t) & 0 \end{pmatrix} \begin{pmatrix} u^1 \\ u^2 \\ u^3 \end{pmatrix}(t)$$

where the u^i are the components of the vector $\mathbf{u}(t)$ in terms of the fixed vectors $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$. Therefore,

$$\mathbf{u}'(t) = \mathbf{\Omega}(t) \times \mathbf{u}(t) = Q'(t)Q(t)^T\mathbf{u}(t) \tag{5.29}$$

where

$$\mathbf{\Omega}(t) = \omega_1(t)\,\mathbf{i}^* + \omega_2(t)\,\mathbf{j}^* + \omega_3(t)\,\mathbf{k}^*.$$

because

$$\mathbf{\Omega}(t) \times \mathbf{u}(t) \equiv \begin{vmatrix} \mathbf{i}^* & \mathbf{j}^* & \mathbf{k}^* \\ w_1 & w_2 & w_3 \\ u^1 & u^2 & u^3 \end{vmatrix} \equiv$$

$$\mathbf{i}^*\left(w_2u^3 - w_3u^2\right) + \mathbf{j}^*\left(w_3u^1 - w_1^3\right) + \mathbf{k}^*\left(w_1u^2 - w_2u^1\right).$$

This proves the lemma and yields the existence part of the following theorem.

Theorem 5.6.4 *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,

$$(\mathbf{\Omega} - \mathbf{\Omega}_1) \times Q(t)\,\mathbf{u} = \mathbf{0}$$

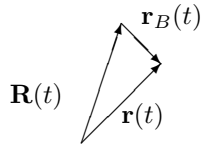
for all $\mathbf{u}$ and since $Q(t)$ is one to one and onto, this implies $(\mathbf{\Omega} - \mathbf{\Omega}_1) \times \mathbf{w} = \mathbf{0}$ for all $\mathbf{w}$ and thus $\mathbf{\Omega} - \mathbf{\Omega}_1 = \mathbf{0}$. This proves the theorem.

Now let $\mathbf{R}(t)$ be a position vector and let

$$\mathbf{r}(t) = \mathbf{R}(t) + \mathbf{r}_B(t)$$

where

$$\mathbf{r}_B(t) \equiv x(t)\,\mathbf{i}(t) + y(t)\,\mathbf{j}(t) + z(t)\,\mathbf{k}(t).$$



In the example of the earth, $\mathbf{R}(t)$ is the position vector of a point $\mathbf{p}(t)$ on the earth's surface and $\mathbf{r}_B(t)$ is the position vector of another point from $\mathbf{p}(t)$, thus regarding $\mathbf{p}(t)$ as the origin. $\mathbf{r}_B(t)$ is the position vector of a point as perceived by the observer on the earth with respect to the vectors he thinks of as fixed. Similarly, $\mathbf{v}_B(t)$ and $\mathbf{a}_B(t)$ will be the velocity and acceleration relative to $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$, and so $\mathbf{v}_B = x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k}$ and $\mathbf{a}_B = x''\mathbf{i} + y''\mathbf{j} + z''\mathbf{k}$. Then

$$\mathbf{v} \equiv \mathbf{r}' = \mathbf{R}' + x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k} + x\mathbf{i}' + y\mathbf{j}' + z\mathbf{k}'.$$

By , 5.29, if $\mathbf{e} \in \{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$, $\mathbf{e}' = \mathbf{\Omega} \times \mathbf{e}$ because the components of these vectors with respect to $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are constant. Therefore,

$$\begin{aligned} x\mathbf{i}' + y\mathbf{j}' + z\mathbf{k}' &= x\mathbf{\Omega} \times \mathbf{i} + y\mathbf{\Omega} \times \mathbf{j} + z\mathbf{\Omega} \times \mathbf{k} \\ &= \Omega(x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) \end{aligned}$$

and consequently,

$$\mathbf{v} = \mathbf{R}' + x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k} + \mathbf{\Omega} \times \mathbf{r}_B = \mathbf{R}' + x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k} + \mathbf{\Omega} \times (x\mathbf{i} + y\mathbf{j} + z\mathbf{k}).$$

Now consider the acceleration. Quantities which are relative to the moving coordinate system and quantities which are relative to a fixed coordinate system are distinguished by using the subscript, B on those relative to the moving coordinates system.

$$\begin{aligned}\mathbf{a} = \mathbf{v}' &= \mathbf{R}'' + x''\mathbf{i} + y''\mathbf{j} + z''\mathbf{k} + \overbrace{x'\mathbf{i}' + y'\mathbf{j}' + z'\mathbf{k}'}^{\boldsymbol{\Omega}\times\mathbf{v}_B} + \boldsymbol{\Omega}' \times \mathbf{r}_B \\ &\quad + \boldsymbol{\Omega}\times\left(\overbrace{x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k}}^{\mathbf{v}_B} + \overbrace{x\mathbf{i}' + y\mathbf{j}' + z\mathbf{k}'}^{\boldsymbol{\Omega}\times\mathbf{r}_B(t)}\right) \\ &= \mathbf{R}'' + \mathbf{a}_B + \boldsymbol{\Omega}' \times \mathbf{r}_B + 2\boldsymbol{\Omega} \times \mathbf{v}_B + \boldsymbol{\Omega}\times(\boldsymbol{\Omega} \times \mathbf{r}_B).\end{aligned}$$

The acceleration $\mathbf{a}_B$ is that perceived by an observer who is moving with the moving coordinate system and for whom the moving coordinate system is fixed. The term $\boldsymbol{\Omega}\times(\boldsymbol{\Omega} \times \mathbf{r}_B)$ is called the centripetal acceleration. Solving for $\mathbf{a}_B$,

$$\mathbf{a}_B = \mathbf{a} - \mathbf{R}'' - \boldsymbol{\Omega}' \times \mathbf{r}_B - 2\boldsymbol{\Omega} \times \mathbf{v}_B - \boldsymbol{\Omega}\times(\boldsymbol{\Omega} \times \mathbf{r}_B). \tag{5.30}$$

Here the term $-(\boldsymbol{\Omega}\times(\boldsymbol{\Omega} \times \mathbf{r}_B))$ is called the centrifugal acceleration, it being an acceleration felt by the observer relative to the moving coordinate system which he regards as fixed, and the term $-2\boldsymbol{\Omega} \times \mathbf{v}_B$ is called the Coriolis acceleration, an acceleration experienced by the observer as he moves relative to the moving coordinate system. The mass multiplied by the Coriolis acceleration defines the Coriolis force.

There is a ride found in some amusement parks in which the victims stand next to a circular wall covered with a carpet or some rough material. Then the whole circular room begins to revolve faster and faster. At some point, the bottom drops out and the victims are held in place by friction. The force they feel is called centrifugal force and it causes centrifugal acceleration. It is not necessary to move relative to coordinates fixed with the revolving wall in order to feel this force and it is is pretty predictable. However, if the nauseated victim moves relative to the rotating wall, he will feel the effects of the Coriolis force and this force is really strange. The difference between these forces is that the Coriolis force is caused by movement relative to the moving coordinate system and the centrifugal force is not.

5.6.2 The Coriolis Acceleration On The Rotating Earth

Now consider the earth. Let $\mathbf{i}^*, \mathbf{j}^*, \mathbf{k}^*$, be the usual basis vectors fixed in space with $\mathbf{k}^*$ pointing in the direction of the north pole from the center of the earth and let $\mathbf{i}, \mathbf{j}, \mathbf{k}$ be the unit vectors described earlier with $\mathbf{i}$ pointing South, $\mathbf{j}$ pointing East, and $\mathbf{k}$ pointing away from the center of the earth at some point of the rotating earth's surface, $\mathbf{p}$. Letting $\mathbf{R}(t)$ be the position vector of the point $\mathbf{p}$, from the center of the earth, observe the coordinates of $\mathbf{R}(t)$ are constant with respect to $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$. Also, since the earth rotates from West to East and the speed of a point on the surface of the earth relative to an observer fixed in space is $\omega|\mathbf{R}|\sin\phi$ where ω is the angular speed of the earth about an axis through the poles, it follows from the geometric definition of the cross product that

$$\mathbf{R}' = \omega\mathbf{k}^* \times \mathbf{R}$$

Therefore, the vector of Theorem 5.6.4 is $\boldsymbol{\Omega} = \omega\mathbf{k}^*$ and so

$$\mathbf{R}'' = \overbrace{\boldsymbol{\Omega}' \times \mathbf{R}}^{=\mathbf{0}} + \boldsymbol{\Omega} \times \mathbf{R}' = \boldsymbol{\Omega}\times(\boldsymbol{\Omega} \times \mathbf{R})$$

since $\mathbf{\Omega}$ does not depend on t. Formula 5.30 implies

$$\mathbf{a}_B = \mathbf{a} - \mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R}) - 2\mathbf{\Omega}\times\mathbf{v}_B - \mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{r}_B). \tag{5.31}$$

In this formula, you can totally ignore the term $\mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{r}_B)$ because it is so small whenever you are considering motion near some point on the earth's surface. To see this, note $\omega \overbrace{(24)(3600)}^{\text{seconds in a day}} = 2\pi$, and so $\omega = 7.2722\times10^{-5}$ in radians per second. If you are using seconds to measure time and feet to measure distance, this term is therefore, no larger than

$$\left(7.2722\times10^{-5}\right)^2|\mathbf{r}_B|.$$

Clearly this is not worth considering in the presence of the acceleration due to gravity which is approximately 32 feet per second squared near the surface of the earth.

If the acceleration $\mathbf{a}$, is due to gravity, then

$$\mathbf{a}_B = \mathbf{a} - \mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R}) - 2\mathbf{\Omega}\times\mathbf{v}_B =$$

$$\overbrace{-\frac{GM(\mathbf{R}+\mathbf{r}_B)}{|\mathbf{R}+\mathbf{r}_B|^3} - \mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R})}^{\equiv\mathbf{g}} - 2\mathbf{\Omega}\times\mathbf{v}_B \equiv \mathbf{g} - 2\mathbf{\Omega}\times\mathbf{v_B}.$$

Note that

$$\mathbf{\Omega}\times(\mathbf{\Omega}\times\mathbf{R}) = (\mathbf{\Omega}\cdot\mathbf{R})\mathbf{\Omega} - |\mathbf{\Omega}|^2\mathbf{R}$$

and so $\mathbf{g}$, the acceleration relative to the moving coordinate system on the earth is not directed exactly toward the center of the earth except at the poles and at the equator, although the components of acceleration which are in other directions are very small when compared with the acceleration due to the force of gravity and are often neglected. Therefore, if the only force acting on an object is due to gravity, the following formula describes the acceleration relative to a coordinate system moving with the earth's surface.

$$\mathbf{a}_B = \mathbf{g} - 2(\mathbf{\Omega}\times\mathbf{v}_B)$$

While the vector, $\mathbf{\Omega}$ is quite small, if the relative velocity, $\mathbf{v}_B$ is large, the Coriolis acceleration could be significant. This is described in terms of the vectors $\mathbf{i}(t), \mathbf{j}(t), \mathbf{k}(t)$ next.

Letting (ρ,θ,ϕ) be the usual spherical coordinates of the point $\mathbf{p}(t)$ on the surface taken with respect to $\mathbf{i}^*,\mathbf{j}^*,\mathbf{k}^*$ the usual way with ϕ the polar angle, it follows the $\mathbf{i}^*,\mathbf{j}^*,\mathbf{k}^*$ coordinates of this point are

$$\begin{pmatrix} \rho\sin(\phi)\cos(\theta) \\ \rho\sin(\phi)\sin(\theta) \\ \rho\cos(\phi) \end{pmatrix}.$$

It follows,

$$\mathbf{i} = \cos(\phi)\cos(\theta)\,\mathbf{i}^* + \cos(\phi)\sin(\theta)\,\mathbf{j}^* - \sin(\phi)\,\mathbf{k}^*$$
$$\mathbf{j} = -\sin(\theta)\,\mathbf{i}^* + \cos(\theta)\,\mathbf{j}^* + 0\mathbf{k}^*$$

and

$$\mathbf{k} = \sin(\phi)\cos(\theta)\,\mathbf{i}^* + \sin(\phi)\sin(\theta)\,\mathbf{j}^* + \cos(\phi)\,\mathbf{k}^*.$$

It is necessary to obtain $\mathbf{k}^*$ in terms of the vectors, $\mathbf{i},\mathbf{j},\mathbf{k}$. Thus the following equation needs to be solved for a,b,c to find $\mathbf{k}^* = a\mathbf{i}+b\mathbf{j}+c\mathbf{k}$

$$\overbrace{\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}}^{\mathbf{k}^*} = \begin{pmatrix} \cos(\phi)\cos(\theta) & -\sin(\theta) & \sin(\phi)\cos(\theta) \\ \cos(\phi)\sin(\theta) & \cos(\theta) & \sin(\phi)\sin(\theta) \\ -\sin(\phi) & 0 & \cos(\phi) \end{pmatrix}\begin{pmatrix} a \\ b \\ c \end{pmatrix} \tag{5.32}$$

The first column is **i**, the second is **j** and the third is **k** in the above matrix. The solution is $a = -\sin(\phi), b = 0$, and $c = \cos(\phi)$.

Now the Coriolis acceleration on the earth equals

$$2(\mathbf{\Omega} \times \mathbf{v}_B) = 2\omega \left(\overbrace{-\sin(\phi)\,\mathbf{i} + 0\mathbf{j} + \cos(\phi)\,\mathbf{k}}^{\mathbf{k}^*} \right) \times (x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k}).$$

This equals

$$2\omega \left[(-y'\cos\phi)\,\mathbf{i} + (x'\cos\phi + z'\sin\phi)\,\mathbf{j} - (y'\sin\phi)\,\mathbf{k} \right]. \tag{5.33}$$

Remember ϕ is fixed and pertains to the fixed point, $\mathbf{p}(t)$ on the earth's surface. Therefore, if the acceleration, $\mathbf{a}$ is due to gravity,

$$\mathbf{a}_B = \mathbf{g} - 2\omega \left[(-y'\cos\phi)\,\mathbf{i} + (x'\cos\phi + z'\sin\phi)\,\mathbf{j} - (y'\sin\phi)\,\mathbf{k} \right]$$

where $\mathbf{g} = -\frac{GM(\mathbf{R}+\mathbf{r}_B)}{|\mathbf{R}+\mathbf{r}_B|^3} - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{R})$ as explained above. The term $\mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{R})$ is pretty small and so it will be neglected. However, the Coriolis force will not be neglected.

Example 5.6.5 *Suppose a rock is dropped from a tall building. Where will it stike?*

Assume $\mathbf{a} = -g\mathbf{k}$ and the $\mathbf{j}$ component of $\mathbf{a}_B$ is approximately

$$-2\omega (x'\cos\phi + z'\sin\phi).$$

The dominant term in this expression is clearly the second one because x' will be small. Also, the $\mathbf{i}$ and $\mathbf{k}$ contributions will be very small. Therefore, the following equation is descriptive of the situation.

$$\mathbf{a}_B = -g\mathbf{k} - 2z'\omega \sin\phi \mathbf{j}.$$

$z' = -gt$ approximately. Therefore, considering the $\mathbf{j}$ component, this is

$$2gt\omega \sin\phi.$$

Two integrations give $\left(\omega g t^3/3\right)\sin\phi$ for the $\mathbf{j}$ component of the relative displacement at time t.

This shows the rock does not fall directly towards the center of the earth as expected but slightly to the east.

Example 5.6.6 *In 1851 Foucault set a pendulum vibrating and observed the earth rotate out from under it. It was a very long pendulum with a heavy weight at the end so that it would vibrate for a long time without stopping*[3]*. This is what allowed him to observe the earth rotate out from under it. Clearly such a pendulum will take 24 hours for the plane of vibration to appear to make one complete revolution at the north pole. It is also reasonable to expect that no such observed rotation would take place on the equator. Is it possible to predict what will take place at various latitudes?*

Using 5.33, in 5.31,

$$\mathbf{a}_B = \mathbf{a} - \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{R})$$

$$-2\omega \left[(-y'\cos\phi)\,\mathbf{i} + (x'\cos\phi + z'\sin\phi)\,\mathbf{j} - (y'\sin\phi)\,\mathbf{k} \right].$$

[3]There is such a pendulum in the Eyring building at BYU and to keep people from touching it, there is a little sign which says Warning! 1000 ohms.

Neglecting the small term, $\boldsymbol{\Omega}\times(\boldsymbol{\Omega}\times\mathbf{R})$, this becomes

$$=-g\mathbf{k}+\mathbf{T}/m-2\omega\left[\left(-y'\cos\phi\right)\mathbf{i}+\left(x'\cos\phi+z'\sin\phi\right)\mathbf{j}-\left(y'\sin\phi\right)\mathbf{k}\right]$$

where $\mathbf{T}$, the tension in the string of the pendulum, is directed towards the point at which the pendulum is supported, and m is the mass of the pendulum bob. The pendulum can be thought of as the position vector from $(0,0,l)$ to the surface of the sphere $x^2+y^2+(z-l)^2=l^2$. Therefore,

$$\mathbf{T}=-T\frac{x}{l}\mathbf{i}-T\frac{y}{l}\mathbf{j}+T\frac{l-z}{l}\mathbf{k}$$

and consequently, the differential equations of relative motion are

$$x''=-T\frac{x}{ml}+2\omega y'\cos\phi$$

$$y''=-T\frac{y}{ml}-2\omega\left(x'\cos\phi+z'\sin\phi\right)$$

and

$$z''=T\frac{l-z}{ml}-g+2\omega y'\sin\phi.$$

If the vibrations of the pendulum are small so that for practical purposes, $z''=z=0$, the last equation may be solved for T to get

$$gm-2\omega y'\sin\left(\phi\right)m=T.$$

Therefore, the first two equations become

$$x''=-\left(gm-2\omega my'\sin\phi\right)\frac{x}{ml}+2\omega y'\cos\phi$$

and

$$y''=-\left(gm-2\omega my'\sin\phi\right)\frac{y}{ml}-2\omega\left(x'\cos\phi+z'\sin\phi\right).$$

All terms of the form xy' or $y'y$ can be neglected because it is assumed x and y remain small. Also, the pendulum is assumed to be long with a heavy weight so that x' and y' are also small. With these simplifying assumptions, the equations of motion become

$$x''+g\frac{x}{l}=2\omega y'\cos\phi$$

and

$$y''+g\frac{y}{l}=-2\omega x'\cos\phi.$$

These equations are of the form

$$x''+a^2x=by',\ y''+a^2y=-bx' \tag{5.34}$$

where $a^2=\frac{g}{l}$ and $b=2\omega\cos\phi$. Then it is fairly tedious but routine to verify that for each constant, c,

$$x=c\sin\left(\frac{bt}{2}\right)\sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right),\ y=c\cos\left(\frac{bt}{2}\right)\sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right) \tag{5.35}$$

yields a solution to 5.34 along with the initial conditions,

$$x\left(0\right)=0,y\left(0\right)=0,x'\left(0\right)=0,y'\left(0\right)=\frac{c\sqrt{b^2+4a^2}}{2}. \tag{5.36}$$

It is clear from experiments with the pendulum that the earth does indeed rotate out from under it causing the plane of vibration of the pendulum to appear to rotate. The purpose of this discussion is not to establish these self evident facts but to predict how long it takes for the plane of vibration to make one revolution. Therefore, there will be some instant in time at which the pendulum will be vibrating in a plane determined by $\mathbf{k}$ and $\mathbf{j}$. (Recall $\mathbf{k}$ points away from the center of the earth and $\mathbf{j}$ points East.) At this instant in time, defined as $t = 0$, the conditions of 5.36 will hold for some value of c and so the solution to 5.34 having these initial conditions will be those of 5.35 by uniqueness of the initial value problem. Writing these solutions differently,

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = c \begin{pmatrix} \sin\left(\frac{bt}{2}\right) \\ \cos\left(\frac{bt}{2}\right) \end{pmatrix} \sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right)$$

This is very interesting! The vector, $c\begin{pmatrix} \sin\left(\frac{bt}{2}\right) \\ \cos\left(\frac{bt}{2}\right) \end{pmatrix}$ always has magnitude equal to $|c|$ but its direction changes very slowly because b is very small. The plane of vibration is determined by this vector and the vector $\mathbf{k}$. The term $\sin\left(\frac{\sqrt{b^2+4a^2}}{2}t\right)$ changes relatively fast and takes values between -1 and 1. This is what describes the actual observed vibrations of the pendulum. Thus the plane of vibration will have made one complete revolution when $t = T$ for

$$\frac{bT}{2} \equiv 2\pi.$$

Therefore, the time it takes for the earth to turn out from under the pendulum is

$$T = \frac{4\pi}{2\omega\cos\phi} = \frac{2\pi}{\omega}\sec\phi.$$

Since ω is the angular speed of the rotating earth, it follows $\omega = \frac{2\pi}{24} = \frac{\pi}{12}$ in radians per hour. Therefore, the above formula implies

$$T = 24\sec\phi.$$

I think this is really amazing. You could actually determine latitude, not by taking readings with instuments using the North Star but by doing an experiment with a big pendulum. You would set it vibrating, observe T in hours, and then solve the above equation for ϕ. Also note the pendulum would not appear to change its plane of vibration at the equator because $\lim_{\phi\to\pi/2}\sec\phi = \infty$.

The Coriolis acceleration is also responsible for the phenomenon of the next example.

Example 5.6.7 *It is known that low pressure areas rotate counterclockwise as seen from above in the Northern hemisphere but clockwise in the Southern hemisphere. Why?*

Neglect accelerations other than the Coriolis acceleration and the following acceleration which comes from an assumption that the point $\mathbf{p}(t)$ is the location of the lowest pressure.

$$\mathbf{a} = -a(r_B)\,\mathbf{r}_B$$

where $r_B = r$ will denote the distance from the fixed point $\mathbf{p}(t)$ on the earth's surface which is also the lowest pressure point. Of course the situation could be more complicated but this will suffice to expain the above question. Then the acceleration observed by a person on the earth relative to the apparantly fixed vectors, $\mathbf{i}, \mathbf{k}, \mathbf{j}$, is

$$\mathbf{a}_B = -a(r_B)(x\mathbf{i}+y\mathbf{j}+z\mathbf{k}) - 2\omega\left[-y'\cos(\phi)\,\mathbf{i} + (x'\cos(\phi)+z'\sin(\phi))\,\mathbf{j} - (y'\sin(\phi)\,\mathbf{k})\right]$$

Therefore, one obtains some differential equations from $\mathbf{a}_B = x''\mathbf{i} + y''\mathbf{j} + z''\mathbf{k}$ by matching the components. These are

$$\begin{aligned} x'' + a(r_B)\, x &= 2\omega y' \cos\phi \\ y'' + a(r_B)\, y &= -2\omega x' \cos\phi - 2\omega z' \sin(\phi) \\ z'' + a(r_B)\, z &= 2\omega y' \sin\phi \end{aligned}$$

Now remember, the vectors, $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are fixed relative to the earth and so are constant vectors. Therefore, from the properties of the determinant and the above differential equations,

$$\left(\mathbf{r}_B' \times \mathbf{r}_B\right)' = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x' & y' & z' \\ x & y & z \end{vmatrix}' = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x'' & y'' & z'' \\ x & y & z \end{vmatrix}$$

$$= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -a(r_B)\, x + 2\omega y' \cos\phi & -a(r_B)\, y - 2\omega x' \cos\phi - 2\omega z' \sin(\phi) & -a(r_B)\, z + 2\omega y' \sin\phi \\ x & y & z \end{vmatrix}$$

Then the $\mathbf{k}^{th}$ component of this cross product equals

$$\omega \cos(\phi)\left(y^2 + x^2\right)' + 2\omega x z' \sin(\phi).$$

The first term will be negative because it is assumed $\mathbf{p}(t)$ is the location of low pressure causing $y^2 + x^2$ to be a decreasing function. If it is assumed there is not a substantial motion in the $\mathbf{k}$ direction, so that z is fairly constant and the last term can be neglected, then the $\mathbf{k}^{th}$ component of $\left(\mathbf{r}_B' \times \mathbf{r}_B\right)'$ is negative provided $\phi \in \left(0, \frac{\pi}{2}\right)$ and positive if $\phi \in \left(\frac{\pi}{2}, \pi\right)$. Beginning with a point at rest, this implies $\mathbf{r}_B' \times \mathbf{r}_B = \mathbf{0}$ initially and then the above implies its $\mathbf{k}^{th}$ component is negative in the upper hemisphere when $\phi < \pi/2$ and positive in the lower hemisphere when $\phi > \pi/2$. Using the right hand and the geometric definition of the cross product, this shows clockwise rotation in the lower hemisphere and counter clockwise rotation in the upper hemisphere.

Note also that as ϕ gets close to $\pi/2$ near the equator, the above reasoning tends to break down because $\cos(\phi)$ becomes close to zero. Therefore, the motion towards the low pressure has to be more pronounced in comparison with the motion in the $\mathbf{k}$ direction in order to draw this conclusion.

5.7 Exercises

1. Remember the Coriolis force was $2\mathbf{\Omega} \times \mathbf{v}_B$ where $\mathbf{\Omega}$ was a particular vector which came from the matrix, $Q(t)$ as described above. Show that

$$Q(t) = \begin{pmatrix} \mathbf{i}(t) \cdot \mathbf{i}(t_0) & \mathbf{j}(t) \cdot \mathbf{i}(t_0) & \mathbf{k}(t) \cdot \mathbf{i}(t_0) \\ \mathbf{i}(t) \cdot \mathbf{j}(t_0) & \mathbf{j}(t) \cdot \mathbf{j}(t_0) & \mathbf{k}(t) \cdot \mathbf{j}(t_0) \\ \mathbf{i}(t) \cdot \mathbf{k}(t_0) & \mathbf{j}(t) \cdot \mathbf{k}(t_0) & \mathbf{k}(t) \cdot \mathbf{k}(t_0) \end{pmatrix}.$$

   There will be no Coriolis force exactly when $\mathbf{\Omega} = \mathbf{0}$ which corresponds to $Q'(t) = 0$. When will $Q'(t) = 0$?

2. An illustration used in many beginning physics books is that of firing a rifle horizontally and dropping an identical bullet from the same height above the perfectly flat ground followed by an assertion that the two bullets will hit the ground at exactly the same time. Is this true on the rotating earth assuming the experiment

takes place over a large perfectly flat field so the curvature of the earth is not an issue? Explain. What other irregularities will occur? Recall the Coriolis force is $2\omega\left[(-y'\cos\phi)\,\mathbf{i}+(x'\cos\phi+z'\sin\phi)\,\mathbf{j}-(y'\sin\phi)\,\mathbf{k}\right]$ where $\mathbf{k}$ points away from the center of the earth, $\mathbf{j}$ points East, and $\mathbf{i}$ points South.

Determinants

6.1 Basic Techniques And Properties

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 6.1.1 *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 6.1.2 *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?

Example 6.1.3 *Find the determinant of*

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

Here is how it is done by "expanding along the first column".

$$(-1)^{1+1} 1 \begin{vmatrix} 3 & 2 \\ 2 & 1 \end{vmatrix} + (-1)^{2+1} 4 \begin{vmatrix} 2 & 3 \\ 2 & 1 \end{vmatrix} + (-1)^{3+1} 3 \begin{vmatrix} 2 & 3 \\ 3 & 2 \end{vmatrix} = 0.$$

What is going on here? Take the 1 in the upper left corner and cross out the row and the column containing the 1. Then take the determinant of the resulting 2×2 matrix. Now multiply this determinant by 1 and then multiply by $(-1)^{1+1}$ because this 1 is in the first row and the first column. This gives the first term in the above sum. Now go to the 4. Cross out the row and the column which contain 4 and take the determinant of the 2×2 matrix which remains. Multiply this by 4 and then by $(-1)^{2+1}$ because the 4 is in the first column and the second row. Finally consider the 3 on the bottom of the first column. Cross out the row and column containing this 3 and take the determinant of what is left. Then

multiply this by 3 and by $(-1)^{3+1}$ because this 3 is in the third row and the first column. This is the pattern used to evaluate the determinant by expansion along the first column.

You could also expand the determinant along the second row as follows.

$$(-1)^{2+1}4\begin{vmatrix}2&3\\2&1\end{vmatrix}+(-1)^{2+2}3\begin{vmatrix}1&3\\3&1\end{vmatrix}+(-1)^{2+3}2\begin{vmatrix}1&2\\3&2\end{vmatrix}=0.$$

It follows exactly the same pattern and you see it gave the same answer. You pick a row or column and corresponding to each number in that row or column, you cross out the row and column containing it, take the determinant of what is left, multiply this by the number and by $(-1)^{i+j}$ assuming the number is in the i^{th} row and the j^{th} column. Then adding these gives the value of the determinant.

What about a 4×4 matrix?

Example 6.1.4 *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 $4\times 3\times 2=24$ terms to evaluate in order to find the determinant using the method of Laplace expansion. Suppose now you have a 10×10 matrix. I hope you see that from the above pattern 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, I think 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 which follows. The above examples motivate the following incredible theorem and definition.

Definition 6.1.5 *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.*

Theorem 6.1.6 *Let A be an $n \times n$ matrix where $n \geq 2$. Then*

$$\det(A) = \sum_{j=1}^{n} a_{ij} \operatorname{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij} \operatorname{cof}(A)_{ij}. \tag{6.1}$$

The first formula consists of expanding the determinant along the i^{th} row and the second expands the determinant along the j^{th} column.

Notwithstanding the difficulties involved in using the method of Laplace expansion, certain types of matrices are very easy to deal with.

Definition 6.1.7 *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 6.1.8 *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 6.1.9 *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

$$1 \begin{vmatrix} 2 & 6 & 7 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{vmatrix}$$

and now expand this along the first column to get this equals

$$1 \times 2 \times \begin{vmatrix} 3 & 33.7 \\ 0 & -1 \end{vmatrix}$$

Next expand the last along the first column which reduces to the product of the main diagonal elements as claimed. This example also demonstrates why the above corollary is true.

There are many properties satisfied by determinants. Some of the most important are listed in the following theorem.

Theorem 6.1.10 *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". In addition to this, if A and B are $n \times n$ matrices, then*

$$\det(AB) = \det(A) \det(B),$$

and if A is an $n \times n$ matrix, then

$$\det(A) = \det\left(A^T\right).$$

This theorem implies the following corollary which gives a way to find determinants. As I pointed out above, the method of Laplace expansion will not be practical for any matrix of large size.

Corollary 6.1.11 *Let A be an $n \times n$ matrix and let B be the matrix obtained by replacing the i^{th} row (column) of A with the sum of the i^{th} row (column) added to a multiple of another row (column). Then* $\det(A) = \det(B)$. *If B is the matrix obtained from A be replacing the i^{th} row (column) of A by a times the i^{th} row (column) then $a \det(A) = \det(B)$.*

Here is an example which shows how to use this corollary to find a determinant.

Example 6.1.12 *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 the above corollary, it has the same determinant as A. Now using the corollary some more, $\det(B) = \left(\frac{-1}{3}\right) \det(C)$ 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 and then the last row was multiplied by (-3). Now replace the last row with 2 times the third added to it and then switch the third and second rows. Then $\det(C) = -\det(D)$ 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 $\det(A) = \det(B) = \left(\frac{-1}{3}\right)(-1485) = 495$.

The theorem about expanding a matrix along any row or column also provides a way to give a formula for the inverse of a matrix. Recall the definition of the inverse of a matrix in Definition 5.1.20 on Page 62.

Theorem 6.1.13 *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 6.1.6 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 6.1.10. 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,

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} \det(A)^{-1} = \delta_{rk}.$$

Now

$$\sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ik} = \sum_{i=1}^{n} a_{ir} \operatorname{cof}(A)_{ki}^{T}$$

which is the kr^{th} entry of $\operatorname{cof}(A)^T A$. Therefore,

$$\frac{\operatorname{cof}(A)^T}{\det(A)} A = I. \tag{6.2}$$

Using the other formula in Theorem 6.1.6, 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,

$$A \frac{\operatorname{cof}(A)^{T}}{\det(A)} = I, \tag{6.3}$$

and it follows from 6.2 and 6.3 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 6.1.10,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A) \det\left(A^{-1}\right)$$

so $\det(A) \neq 0$. This proves the theorem.

Theorem 6.1.13 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.

Example 6.1.14 *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 Corollary 6.1.11 on Page 90, 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

$$\frac{1}{12} \begin{pmatrix} -2 & -2 & 6 \\ 4 & -2 & 0 \\ 2 & 8 & -6 \end{pmatrix}^{T} = \begin{pmatrix} -\frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{6} & \frac{2}{3} \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{pmatrix}.$$

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 6.1.15 *Suppose*

$$A(t) = \begin{pmatrix} e^t & 0 & 0 \\ 0 & \cos t & \sin t \\ 0 & -\sin t & \cos t \end{pmatrix}$$

Find $A(t)^{-1}$.

First note $\det(A(t)) = e^t$. 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

$$\frac{1}{e^t} \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{pmatrix}^T = \begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}.$$

This formula for the inverse also implies a famous proceedure known as Cramer's rule. Cramer's rule gives a formula for the solutions, $\mathbf{x}$, to a system of equations, $A\mathbf{x} = \mathbf{y}$.

In case you are solving a system of equations, $A\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$, it follows that if A^{-1} exists,

$$\mathbf{x} = \left(A^{-1}A\right)\mathbf{x} = A^{-1}(A\mathbf{x}) = A^{-1}\mathbf{y}$$

thus solving the system. Now in the case that A^{-1} exists, there is a formula for A^{-1} given above. Using this formula,

$$x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \sum_{j=1}^{n} \frac{1}{\det(A)} \operatorname{cof}(A)_{ji}\, y_j.$$

By the formula for the expansion of a determinant along a column,

$$x_i = \frac{1}{\det(A)} \det \begin{pmatrix} * & \cdots & y_1 & \cdots & * \\ \vdots & & \vdots & & \vdots \\ * & \cdots & y_n & \cdots & * \end{pmatrix},$$

where here the i^{th} column of A is replaced with the column vector, $(y_1 \cdots, y_n)^T$, and the determinant of this modified matrix is taken and divided by $\det(A)$. This formula is known as Cramer's rule.

Procedure 6.1.16 *Suppose A is an $n \times n$ matrix and it is desired to solve the system $A\mathbf{x} = \mathbf{y}, \mathbf{y} = (y_1, \cdots, y_n)^T$ for $\mathbf{x} = (x_1, \cdots, x_n)^T$. Then Cramer's rule says*

$$x_i = \frac{\det A_i}{\det A}$$

where A_i is obtained from A by replacing the i^{th} column of A with the column $(y_1, \cdots, y_n)^T$.

The following theorem is of fundamental importance and ties together many of the ideas presented above. It is proved in the next section.

Theorem 6.1.17 *Let A be an $n \times n$ matrix. Then the following are equivalent.*

1. *A is one to one.*
2. *A is onto.*
3. $\det(A) \neq 0$.

6.2 Exercises

1. Find the determinants of the following matrices.

   (a) $\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 2 \\ 0 & 9 & 8 \end{pmatrix}$ (The answer is 31.)

   (b) $\begin{pmatrix} 4 & 3 & 2 \\ 1 & 7 & 8 \\ 3 & -9 & 3 \end{pmatrix}$ (The answer is 375.)

   (c) $\begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 3 & 2 & 3 \\ 4 & 1 & 5 & 0 \\ 1 & 2 & 1 & 2 \end{pmatrix}$, (The answer is -2.)

2. A matrix is said to be orthogonal if $A^T A = I$. Thus the inverse of an orthogonal matrix is just its transpose. What are the possible values of $\det(A)$ if A is an orthogonal matrix?

3. If A^{-1} exist, what is the relationship between $\det(A)$ and $\det\left(A^{-1}\right)$. Explain your answer.

4. Is it true that $\det(A+B) = \det(A) + \det(B)$? If this is so, explain why it is so and if it is not so, give a counter example.

5. Let A be an $r \times r$ matrix and suppose there are $r-1$ rows (columns) such that all rows (columns) are linear combinations of these $r-1$ rows (columns). Show $\det(A) = 0$.

6. Show $\det(aA) = a^n \det(A)$ where here A is an $n \times n$ matrix and a is a scalar.

7. Suppose A is an upper triangular matrix. Show that A^{-1} exists if and only if all elements of the main diagonal are non zero. Is it true that A^{-1} will also be upper triangular? Explain. Is everything the same for lower triangular matrices?

8. Let A and B be two $n \times n$ matrices. $A \sim B$ (A is similar to B) means there exists an invertible matrix, S such that $A = S^{-1}BS$. Show that if $A \sim B$, then $B \sim A$. Show also that $A \sim A$ and that if $A \sim B$ and $B \sim C$, then $A \sim C$.

9. In the context of Problem 8 show that if $A \sim B$, then $\det(A) = \det(B)$.

10. Let A be an $n \times n$ matrix and let $\mathbf{x}$ be a nonzero vector such that $A\mathbf{x} = \lambda\mathbf{x}$ for some scalar, λ. When this occurs, the vector, $\mathbf{x}$ is called an eigenvector and the scalar, λ is called an eigenvalue. It turns out that not every number is an eigenvalue. Only certain ones are. Why? **Hint:** Show that if $A\mathbf{x} = \lambda\mathbf{x}$, then $(\lambda I - A)\mathbf{x} = \mathbf{0}$. Explain why this shows that $(\lambda I - A)$ is not one to one and not onto. Now use Theorem 6.1.17 to argue $\det(\lambda I - A) = 0$. What sort of equation is this? How many solutions does it have?

11. Suppose $\det(\lambda I - A) = 0$. Show using Theorem 6.1.17 there exists $\mathbf{x} \neq \mathbf{0}$ such that $(\lambda I - A)\mathbf{x} = \mathbf{0}$.

12. Let $F(t) = \det \begin{pmatrix} a(t) & b(t) \\ c(t) & d(t) \end{pmatrix}$. Verify

$$F'(t) = \det \begin{pmatrix} a'(t) & b'(t) \\ c(t) & d(t) \end{pmatrix} + \det \begin{pmatrix} a(t) & b(t) \\ c'(t) & d'(t) \end{pmatrix}.$$

Now suppose

$$F(t) = \det \begin{pmatrix} a(t) & b(t) & c(t) \\ d(t) & e(t) & f(t) \\ g(t) & h(t) & i(t) \end{pmatrix}.$$

Use Laplace expansion and the first part to verify $F'(t) =$

$$\det \begin{pmatrix} a'(t) & b'(t) & c'(t) \\ d(t) & e(t) & f(t) \\ g(t) & h(t) & i(t) \end{pmatrix} + \det \begin{pmatrix} a(t) & b(t) & c(t) \\ d'(t) & e'(t) & f'(t) \\ g(t) & h(t) & i(t) \end{pmatrix}$$
$$+ \det \begin{pmatrix} a(t) & b(t) & c(t) \\ d(t) & e(t) & f(t) \\ g'(t) & h'(t) & i'(t) \end{pmatrix}.$$

Conjecture a general result valid for $n \times n$ matrices and explain why it will be true. Can a similar thing be done with the columns?

13. Use the formula for the inverse in terms of the cofactor matrix to find the inverse of the matrix,

$$A = \begin{pmatrix} e^t & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & e^t \cos t - e^t \sin t & e^t \cos t + e^t \sin t \end{pmatrix}.$$

14. Let A be an $r \times r$ matrix and let B be an $m \times m$ matrix such that $r + m = n$. Consider the following $n \times n$ block matrix

$$C = \begin{pmatrix} A & 0 \\ D & B \end{pmatrix}.$$

where the D is an $m \times r$ matrix, and the 0 is a $r \times m$ matrix. Letting I_k denote the $k \times k$ identity matrix, tell why

$$C = \begin{pmatrix} A & 0 \\ D & I_m \end{pmatrix} \begin{pmatrix} I_r & 0 \\ 0 & B \end{pmatrix}.$$

Now explain why $\det(C) = \det(A)\det(B)$. **Hint:** Part of this will require an explantion of why

$$\det \begin{pmatrix} A & 0 \\ D & I_m \end{pmatrix} = \det(A).$$

See Corollary 6.1.11.

6.3 The Mathematical Theory Of Determinants

It is easiest to give a different definition of the determinant which is clearly well defined and then prove the earlier one in terms of Laplace expansion. Let $(i_1, \cdots, i_n)$ be an ordered list of numbers from $\{1, \cdots, n\}$. This means the order is important so $(1,2,3)$ and $(2,1,3)$ are different. There will be some repetition between this section and the earlier section on determinants. The main purpose is to give all the missing proofs. Two books which give a good introduction to determinants are Apostol [1] and Rudin [11]. A recent book which also has a good introduction is Baker [2]

The following Lemma will be essential in the definition of the determinant.

Lemma 6.3.1 *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.*

$$\operatorname{sgn}_n(1, \cdots, n) = 1 \tag{6.4}$$

$$\operatorname{sgn}_n(i_1, \cdots, p, \cdots, q, \cdots, i_n) = -\operatorname{sgn}_n(i_1, \cdots, q, \cdots, p, \cdots, i_n) \tag{6.5}$$

In words, the second property states that if two of the numbers are switched, the value of the function is multiplied by -1*. Also, in the case where* $n > 1$ *and* $\{i_1, \cdots, i_n\} = \{1, \cdots, n\}$ *so that every number from* $\{1, \cdots, n\}$ *appears in the ordered list,* $(i_1, \cdots, i_n)$*,*

$$\operatorname{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) \equiv$$

$$(-1)^{n-\theta} \operatorname{sgn}_{n-1}(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n) \tag{6.6}$$

where $n = i_\theta$ *in the ordered list,* $(i_1, \cdots, i_n)$*.*

Proof: To begin with, it is necessary to show the existence of such a function. This is clearly true if $n = 1$. Define $\operatorname{sgn}_1(1) \equiv 1$ and observe that it works. No switching is possible. In the case where $n = 2$, it is also clearly true. Let $\operatorname{sgn}_2(1,2) = 1$ and $\operatorname{sgn}_2(2,1) = 0$ while $\operatorname{sgn}_2(2,2) = \operatorname{sgn}_2(1,1) = 0$ and verify it works. Assuming such a function exists for n, sgn_{n+1} will be defined in terms of sgn_n. If there are any repeated numbers in $(i_1, \cdots, i_{n+1})$, $\operatorname{sgn}_{n+1}(i_1, \cdots, i_{n+1}) \equiv 0$. If there are no repeats, then $n+1$ appears somewhere in the ordered list. Let θ be the position of the number $n+1$ in the list. Thus, the list is of the form $(i_1, \cdots, i_{\theta-1}, n+1, i_{\theta+1}, \cdots, i_{n+1})$. From 6.6 it must be that

$$\operatorname{sgn}_{n+1}(i_1, \cdots, i_{\theta-1}, n+1, i_{\theta+1}, \cdots, i_{n+1}) \equiv$$

$$(-1)^{n+1-\theta} \operatorname{sgn}_n(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_{n+1}).$$

It is necessary to verify this satisfies 6.4 and 6.5 with n replaced with $n+1$. The first of these is obviously true because

$$\operatorname{sgn}_{n+1}(1, \cdots, n, n+1) \equiv (-1)^{n+1-(n+1)} \operatorname{sgn}_n(1, \cdots, n) = 1.$$

If there are repeated numbers in $(i_1, \cdots, i_{n+1})$, then it is obvious 6.5 holds because both sides would equal zero from the above definition. It remains to verify 6.5 in the case where there are no numbers repeated in $(i_1, \cdots, i_{n+1})$. Consider

$$\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{p}, \cdots, \overset{s}{q}, \cdots, i_{n+1}\right),$$

where the r above the p indicates the number, p is in the r^{th} position and the s above the q indicates that the number, q is in the s^{th} position. Suppose first that $r < \theta < s$. Then

$$\operatorname{sgn}_{n+1}\left(i_1, \cdots, \overset{r}{p}, \cdots, \overset{\theta}{n+1}, \cdots, \overset{s}{q}, \cdots, i_{n+1}\right) \equiv$$

$$(-1)^{n+1-\theta}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{p},\cdots,\overset{s-1}{q},\cdots,i_{n+1}\right)$$

while

$$\operatorname{sgn}_{n+1}\left(i_1,\cdots,\overset{r}{q},\cdots,\overset{\theta}{n+1},\cdots,\overset{s}{p},\cdots,i_{n+1}\right)=$$
$$(-1)^{n+1-\theta}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,\overset{s-1}{p},\cdots,i_{n+1}\right)$$

and so, by induction, a switch of p and q introduces a minus sign in the result. Similarly, if $\theta > s$ or if $\theta < r$ it also follows that 6.5 holds. The interesting case is when $\theta = r$ or $\theta = s$. Consider the case where $\theta = r$ and note the other case is entirely similar.

$$\operatorname{sgn}_{n+1}\left(i_1,\cdots,\overset{r}{n+1},\cdots,\overset{s}{q},\cdots,i_{n+1}\right)=$$
$$(-1)^{n+1-r}\operatorname{sgn}_n\left(i_1,\cdots,\overset{s-1}{q},\cdots,i_{n+1}\right) \tag{6.7}$$

while

$$\operatorname{sgn}_{n+1}\left(i_1,\cdots,\overset{r}{q},\cdots,\overset{s}{n+1},\cdots,i_{n+1}\right)=$$
$$(-1)^{n+1-s}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right). \tag{6.8}$$

By making $s-1-r$ switches, move the q which is in the $s-1^{th}$ position in 6.7 to the r^{th} position in 6.8. By induction, each of these switches introduces a factor of -1 and so

$$\operatorname{sgn}_n\left(i_1,\cdots,\overset{s-1}{q},\cdots,i_{n+1}\right)=(-1)^{s-1-r}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right).$$

Therefore,

$$\operatorname{sgn}_{n+1}\left(i_1,\cdots,\overset{r}{n+1},\cdots,\overset{s}{q},\cdots,i_{n+1}\right)=(-1)^{n+1-r}\operatorname{sgn}_n\left(i_1,\cdots,\overset{s-1}{q},\cdots,i_{n+1}\right)$$
$$=(-1)^{n+1-r}(-1)^{s-1-r}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right)$$
$$=(-1)^{n+s}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right)=(-1)^{2s-1}(-1)^{n+1-s}\operatorname{sgn}_n\left(i_1,\cdots,\overset{r}{q},\cdots,i_{n+1}\right)$$
$$=-\operatorname{sgn}_{n+1}\left(i_1,\cdots,\overset{r}{q},\cdots,\overset{s}{n+1},\cdots,i_{n+1}\right).$$

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 6.4 and 6.5, you could start with $f(1,\cdots,n)=g(1,\cdots,n)$ and applying the same sequence of switches, eventually arrive at $f(i_1,\cdots,i_n)=g(i_1,\cdots,i_n)$. If any numbers are repeated, then 6.5 gives both functions are equal to zero for that ordered list. This proves the lemma.

In what follows sgn will often be used rather than sgn_n because the context supplies the appropriate n.

Definition 6.3.2 *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 6.3.3 *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

$$\det(A(r_1,\cdots,r_n)) = \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n} \tag{6.9}$$

and

$$A(1,\cdots,n) = A.$$

Proposition 6.3.4 *Let*

$$(r_1,\cdots,r_n)$$

be an ordered list of numbers from $\{1,\cdots,n\}$. Then

$$\operatorname{sgn}(r_1,\cdots,r_n)\det(A)$$

$$= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n} \tag{6.10}$$

$$= \det(A(r_1,\cdots,r_n)). \tag{6.11}$$

Proof: Let $(1,\cdots,n) = (1,\cdots,r,\cdots s,\cdots,n)$ so $r < s$.

$$\det(A(1,\cdots,r,\cdots,s,\cdots,n)) = \tag{6.12}$$

$$\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_r,\cdots,k_s,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_r}\cdots a_{sk_s}\cdots a_{nk_n},$$

and renaming the variables, calling k_s, k_r and k_r, k_s, this equals

$$= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_s,\cdots,k_r,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_s}\cdots a_{sk_r}\cdots a_{nk_n}$$

$$= \sum_{(k_1,\cdots,k_n)} -\operatorname{sgn}\left(k_1,\cdots,\overbrace{k_r,\cdots,k_s}^{\text{These got switched}},\cdots,k_n\right) a_{1k_1}\cdots a_{sk_r}\cdots a_{rk_s}\cdots a_{nk_n}$$

$$= -\det(A(1,\cdots,s,\cdots,r,\cdots,n)). \tag{6.13}$$

Consequently,

$$\det(A(1,\cdots,s,\cdots,r,\cdots,n)) =$$

$$-\det(A(1,\cdots,r,\cdots,s,\cdots,n)) = -\det(A)$$

Now letting $A(1,\cdots,s,\cdots,r,\cdots,n)$ play the role of A, and continuing in this way, switching pairs of numbers,

$$\det(A(r_1,\cdots,r_n)) = (-1)^p \det(A)$$

where it took p switches to obtain$(r_1,\cdots,r_n)$ from $(1,\cdots,n)$. By Lemma 6.3.1, this implies

$$\det(A(r_1,\cdots,r_n)) = (-1)^p \det(A) = \operatorname{sgn}(r_1,\cdots,r_n)\det(A)$$

and proves the proposition in the case when there are no repeated numbers in the ordered list, $(r_1,\cdots,r_n)$. However, if there is a repeat, say the r^{th} row equals the s^{th} row, then the reasoning of 6.12 -6.13 shows that $A(r_1,\cdots,r_n) = 0$ and also $\operatorname{sgn}(r_1,\cdots,r_n) = 0$ so the formula holds in this case also.

Observation 6.3.5 *There are $n!$ ordered lists of distinct numbers from $\{1,\cdots,n\}$.*

To see this, consider n slots placed in order. There are n choices for the first slot. For each of these choices, there are $n-1$ choices for the second. Thus there are $n(n-1)$ ways to fill the first two slots. Then for each of these ways there are $n-2$ choices left for the third slot. Continuing this way, there are $n!$ ordered lists of distinct numbers from $\{1,\cdots,n\}$ as stated in the observation.

With the above, it is possible to give a more symmetric description of the determinant from which it will follow that $\det(A) = \det\left(A^T\right)$.

Corollary 6.3.6 *The following formula for $\det(A)$ is valid.*

$$\det(A) = \frac{1}{n!}\cdot$$

$$\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n}. \tag{6.14}$$

And also $\det\left(A^T\right) = \det(A)$ where A^T is the transpose of A. (Recall that for $A^T = \left(a_{ij}^T\right)$, $a_{ij}^T = a_{ji}$.)

Proof: From Proposition 6.3.4, if the r_i are distinct,

$$\det(A) = \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n}.$$

Summing over all ordered lists, $(r_1,\cdots,r_n)$ where the r_i are distinct, (If the r_i are not distinct, $\operatorname{sgn}(r_1,\cdots,r_n) = 0$ and so there is no contribution to the sum.)

$$n!\det(A) =$$

$$\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n}.$$

This proves the corollary since the formula gives the same number for A as it does for A^T.

Corollary 6.3.7 *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 6.3.4 when two rows are switched, the determinant of the resulting matrix is (-1) times the determinant of the original matrix. By Corollary 6.3.6 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 6.3.8 *A vector,* $\mathbf{w}$*, is a linear combination of the vectors* $\{\mathbf{v}_1,\cdots,\mathbf{v}_r\}$ *if there exists scalars,* $c_1,\cdots c_r$ *such that* $\mathbf{w} = \sum_{k=1}^r c_k\mathbf{v}_k$*. This is the same as saying*

$$\mathbf{w} \in \operatorname{span}\{\mathbf{v}_1,\cdots,\mathbf{v}_r\}.$$

The following corollary is also of great use.

Corollary 6.3.9 *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 6.3.7 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 6.3.7

$$\det(A) = \sum_{k=1}^r c_k \det\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \mathbf{a}_k\end{pmatrix} = 0.$$

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 6.3.10 *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 6.3.11 *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 6.3.4,

$$\begin{aligned}
&\det(AB) = \\
&\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, c_{1k_1}\cdots c_{nk_n} \\
&= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n) \left(\sum_{r_1} a_{1r_1}b_{r_1k_1}\right)\cdots\left(\sum_{r_n} a_{nr_n}b_{r_nk_n}\right) \\
&= \sum_{(r_1\cdots,r_n)} \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, b_{r_1k_1}\cdots b_{r_nk_n}\left(a_{1r_1}\cdots a_{nr_n}\right) \\
&= \sum_{(r_1\cdots,r_n)} \operatorname{sgn}(r_1\cdots r_n)\, a_{1r_1}\cdots a_{nr_n}\det(B) = \det(A)\det(B).
\end{aligned}$$

This proves the theorem.

Lemma 6.3.12 *Suppose a matrix is of the form*

$$M = \begin{pmatrix} A & * \\ \mathbf{0} & a \end{pmatrix} \tag{6.15}$$

or

$$M = \begin{pmatrix} A & \mathbf{0} \\ * & a \end{pmatrix} \tag{6.16}$$

where a is a number and A is an $(n-1)\times(n-1)$ matrix and $*$ denotes either a column or a row having length $n-1$ and the $\mathbf{0}$ denotes either a column or a row of length $n-1$ consisting entirely of zeros. Then

$$\det(M) = a\det(A).$$

Proof: Denote M by (m_{ij}). Thus in the first case, $m_{nn} = a$ and $m_{ni} = 0$ if $i \neq n$ while in the second case, $m_{nn} = a$ and $m_{in} = 0$ if $i \neq n$. From the definition of the determinant,

$$\det(M) \equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}_n(k_1,\cdots,k_n)\, m_{1k_1}\cdots m_{nk_n}$$

Letting θ denote the position of n in the ordered list, $(k_1,\cdots,k_n)$ then using the earlier conventions used to prove Lemma 6.3.1, $\det(M)$ 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 6.16. Then if $k_n \neq n$, the term involving m_{nk_n} in the above expression equals zero. Therefore, the only terms which survive are those for which $\theta = n$ or in other words, those for which $k_n = n$. Therefore, the above expression reduces to

$$a \sum_{(k_1,\cdots,k_{n-1})} \operatorname{sgn}_{n-1}(k_1,\cdots k_{n-1})\, m_{1k_1}\cdots m_{(n-1)k_{n-1}} = a\det(A).$$

To get the assertion in the situation of 6.15 use Corollary 6.3.6 and 6.16 to write

$$\det(M) = \det\left(M^T\right) = \det\left(\begin{pmatrix} A^T & \mathbf{0} \\ * & a \end{pmatrix}\right) = a\det\left(A^T\right) = a\det(A).$$

This proves the lemma.

In terms of the theory of determinants, arguably the most important idea is that of Laplace expansion along a row or a column. This will follow from the above definition of a determinant.

Definition 6.3.13 *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 6.3.14 *Let A be an $n \times n$ matrix where $n \geq 2$. Then*

$$\det(A) = \sum_{j=1}^{n} a_{ij}\operatorname{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij}\operatorname{cof}(A)_{ij}. \tag{6.17}$$

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 6.3.7,

$$\det(A) = \sum_{j=1}^{n} \det(B_j)$$

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 6.3.4, 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 6.3.12,

$$\begin{aligned} \det(B_j) &= (-1)^{n-j}(-1)^{n-i}\det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij} \end{pmatrix}\right) \\ &= (-1)^{i+j}\det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij} \end{pmatrix}\right) = a_{ij}\operatorname{cof}(A)_{ij}. \end{aligned}$$

Therefore,

$$\det(A) = \sum_{j=1}^{n} a_{ij}\operatorname{cof}(A)_{ij}$$

which is the formula for expanding $\det(A)$ along the i^{th} row. Also,

$$\begin{aligned} \det(A) &= \det\left(A^T\right) = \sum_{j=1}^{n} a_{ij}^T \operatorname{cof}\left(A^T\right)_{ij} \\ &= \sum_{j=1}^{n} a_{ji}\operatorname{cof}(A)_{ji} \end{aligned}$$

which is the formula for expanding $\det(A)$ along the i^{th} column. This proves the theorem.

Note that this gives an easy way to write a formula for the inverse of an $n \times n$ matrix. Recall the definition of the inverse of a matrix in Definition 5.1.20 on Page 62.

Theorem 6.3.15 *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 6.3.14 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 6.3.7. 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 6.3.14, 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 6.3.11,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A) \det\left(A^{-1}\right)$$

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 6.3.16 *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 6.3.11 implies

$$\det B \det A = 1$$

and so $\det A \neq 0$. Therefore from Theorem 6.3.15, 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 6.3.15 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 6.3.17 *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 6.3.14.

Corollary 6.3.18 *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 6.3.19 *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 6.3.20 *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 6.3.21 *The determinant rank equals the row rank.*

Proof: From Theorem 6.3.20, 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 6.3.22 *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 6.3.21, 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 6.3.23 *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 6.3.20. 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 6.3.15 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 6.3.24 *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.

6.4 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 $A\mathbf{x} = \mathbf{0}$.

6.5 The Cayley Hamilton Theorem

Definition 6.5.1 *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 6.5.2 *Suppose for all $|\lambda|$ large enough,*

$$A_0 + A_1\lambda + \cdots + A_m\lambda^m = 0,$$

where the A_i are $n \times n$ matrices. Then each $A_i = 0$.

Proof: Multiply by λ^{-m} to obtain

$$A_0\lambda^{-m} + A_1\lambda^{-m+1} + \cdots + A_{m-1}\lambda^{-1} + A_m = 0.$$

Now let $|\lambda| \to \infty$ to obtain $A_m = 0$. With this, multiply by λ to obtain

$$A_0\lambda^{-m+1} + A_1\lambda^{-m+2} + \cdots + A_{m-1} = 0.$$

Now let $|\lambda| \to \infty$ to obtain $A_{m-1} = 0$. Continue multiplying by λ and letting $\lambda \to \infty$ to obtain that all the $A_i = 0$. This proves the lemma.

With the lemma, here is a simple corollary.

Corollary 6.5.3 *Let A_i and B_i be $n \times n$ matrices and suppose*

$$A_0 + A_1\lambda + \cdots + A_m\lambda^m = B_0 + B_1\lambda + \cdots + B_m\lambda^m$$

for all $|\lambda|$ large enough. Then $A_i = B_i$ for all i. Consequently if λ is replaced by any $n \times n$ matrix, the two sides will be equal. That is, for C any $n \times n$ matrix,

$$A_0 + A_1C + \cdots + A_mC^m = B_0 + B_1C + \cdots + B_mC^m.$$

Proof: Subtract and use the result of the lemma.

With this preparation, here is a relatively easy proof of the Cayley Hamilton theorem.

Theorem 6.5.4 *Let A be an $n \times n$ matrix and let $p(\lambda) \equiv \det(\lambda I - A)$ be the characteristic polynomial. Then $p(A) = 0$.*

Proof: Let $C(\lambda)$ equal the transpose of the cofactor matrix of $(\lambda I - A)$ for $|\lambda|$ large. (If $|\lambda|$ is large enough, then λ cannot be in the finite list of eigenvalues of A and so for such λ, $(\lambda I - A)^{-1}$ exists.) Therefore, by Theorem 6.3.15

$$C(\lambda) = p(\lambda)(\lambda I - A)^{-1}.$$

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 6.5.3 may be used. It follows the matrix coefficients corresponding to equal powers of λ are equal on both sides of this equation. Therefore, if λ is replaced with A, the two sides will be equal. Thus

$$0 = (A - A)\left(C_0 + C_1A + \cdots + C_{n-1}A^{n-1}\right) = p(A) I = p(A).$$

This proves the Cayley Hamilton theorem.

6.6 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 this is true.

Suppose A is a matrix of the form

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{r1} & \cdots & A_{rm} \end{pmatrix} \tag{6.18}$$

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 6.6.1 *Let A be an $m \times n$ block matrix as in 6.18 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

$$\begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix} \tag{6.19}$$

and A is a block matrix of the form

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{p1} & \cdots & A_{pm} \end{pmatrix} \tag{6.20}$$

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 6.6.2 *Let B be a block matrix as in 6.19 and let A be a block matrix as in 6.20 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*

$$\sum_s B_{is}A_{sj}. \tag{6.21}$$

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 6.6.1 applied twice,

$$\begin{pmatrix} \begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix} \begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{p1} & \cdots & A_{pm} \end{pmatrix} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{pmatrix}$$

$$= \begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix} \begin{pmatrix} \sum_j A_{1j}\mathbf{x}_j \\ \vdots \\ \sum_j A_{rj}\mathbf{x}_j \end{pmatrix}$$

$$= \begin{pmatrix} \sum_s \sum_j B_{1s}A_{sj}\mathbf{x}_j \\ \vdots \\ \sum_s \sum_j B_{rs}A_{sj}\mathbf{x}_j \end{pmatrix} = \begin{pmatrix} \sum_j \left(\sum_s B_{1s}A_{sj}\right)\mathbf{x}_j \\ \vdots \\ \sum_j \left(\sum_s B_{rs}A_{sj}\right)\mathbf{x}_j \end{pmatrix}$$

$$= \begin{pmatrix} \sum_s B_{1s}A_{s1} & \cdots & \sum_s B_{1s}A_{sm} \\ \vdots & \ddots & \vdots \\ \sum_s B_{rs}A_{s1} & \cdots & \sum_s B_{rs}A_{sm} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{pmatrix}$$

By Lemma 6.6.1, this shows that $(BA)\,\mathbf{x}$ equals the block matrix whose ij^{th} entry is given by 6.21 times $\mathbf{x}$. Since $\mathbf{x}$ is an arbitrary vector in $\mathbb{F}^n$, this proves the theorem.

The message of this theorem is that you can formally multiply block matrices as though the blocks were numbers. You just have to pay attention to the preservation of order.

This simple idea of block multiplication turns out to be very useful later. For now here is an interesting and significant application. In this theorem, $p_M(t)$ denotes the polynomial, $\det(tI - M)$. Thus the zeros of this polynomial are the eigenvalues of the matrix, M.

Theorem 6.6.3 *Let A be an $m \times n$ matrix and let B be an $n \times m$ matrix for $m \leq n$. Then*

$$p_{BA}(t) = t^{n-m} p_{AB}(t),$$

so the eigenvalues of BA and AB are the same including multiplicities except that BA has $n - m$ extra zero eigenvalues.

Proof: Use block multiplication to write

$$\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}$$

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}.$$

Therefore,

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$$

By Problem 11 of Page 111, it follows that $\begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$ and $\begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix}$ have the same characteristic polynomials. Therefore, noting that BA is an $n \times n$ matrix and AB is an $m \times m$ matrix,

$$t^m \det(tI - BA) = t^n \det(tI - AB)$$

and so $\det(tI - BA) = p_{BA}(t) = t^{n-m} \det(tI - AB) = t^{n-m} p_{AB}(t)$. This proves the theorem.

6.7 Exercises

1. Show that matrix multiplication is associative. That is, $(AB)\,C = A\,(BC)$.

2. Show the inverse of a matrix, if it exists, is unique. Thus if $AB = BA = I$, then $B = A^{-1}$.

3. In the proof of Theorem 6.3.15 it was claimed that $\det(I) = 1$. Here $I = (\delta_{ij})$. Prove this assertion. Also prove Corollary 6.3.18.

4. Let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ be vectors in $\mathbb{F}^n$ and let $M(\mathbf{v}_1, \cdots, \mathbf{v}_n)$ denote the matrix whose i^{th} column equals $\mathbf{v}_i$. Define

   $$d(\mathbf{v}_1, \cdots, \mathbf{v}_n) \equiv \det(M(\mathbf{v}_1, \cdots, \mathbf{v}_n)).$$

   Prove that d is linear in each variable, (multilinear), that

   $$d(\mathbf{v}_1, \cdots, \mathbf{v}_i, \cdots, \mathbf{v}_j, \cdots, \mathbf{v}_n) = -d(\mathbf{v}_1, \cdots, \mathbf{v}_j, \cdots, \mathbf{v}_i, \cdots, \mathbf{v}_n), \tag{6.22}$$

and

$$d(\mathbf{e}_1, \cdots, \mathbf{e}_n) = 1 \tag{6.23}$$

where here $\mathbf{e}_j$ is the vector in $\mathbb{F}^n$ which has a zero in every position except the j^{th} position in which it has a one.

5. Suppose $f : \mathbb{F}^n \times \cdots \times \mathbb{F}^n \to \mathbb{F}$ satisfies 6.22 and 6.23 and is linear in each variable. Show that $f = d$.

6. Show that if you replace a row (column) of an $n \times n$ matrix A with itself added to some multiple of another row (column) then the new matrix has the same determinant as the original one.

7. If $A = (a_{ij})$, show $\det(A) = \sum_{(k_1, \cdots, k_n)} \operatorname{sgn}(k_1, \cdots, k_n) a_{k_1 1} \cdots a_{k_n n}$.

8. Use the result of Problem 6 to evaluate by hand the determinant

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

9. Find the inverse if it exists of the matrix,

$$\begin{pmatrix} e^t & \cos t & \sin t \\ e^t & -\sin t & \cos t \\ e^t & -\cos t & -\sin t \end{pmatrix}.$$

10. Let $Ly = y^{(n)} + a_{n-1}(x) y^{(n-1)} + \cdots + a_1(x) y' + a_0(x) y$ where the a_i are given continuous functions defined on a closed interval, (a, b) and y is some function which has n derivatives so it makes sense to write Ly. Suppose $Ly_k = 0$ for $k = 1, 2, \cdots, n$. The Wronskian of these functions, y_i is defined as

$$W(y_1, \cdots, y_n)(x) \equiv \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n-1)}(x) & \cdots & y_n^{(n-1)}(x) \end{pmatrix}$$

Show that for $W(x) = W(y_1, \cdots, y_n)(x)$ to save space,

$$W'(x) = \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n)}(x) & \cdots & y_n^{(n)}(x) \end{pmatrix}.$$

Now use the differential equation, $Ly = 0$ which is satisfied by each of these functions, y_i and properties of determinants presented above to verify that $W' + a_{n-1}(x) W = 0$. Give an explicit solution of this linear differential equation, Abel's formula, and use your answer to verify that the Wronskian of these solutions to the equation, $Ly = 0$ either vanishes identically on (a, b) or never.

11. Two $n \times n$ matrices, A and B, are similar if $B = S^{-1}AS$ for some invertible $n \times n$ matrix, S. Show that if two matrices are similar, they have the same characteristic polynomials.

12. Suppose the characteristic polynomial of an $n \times n$ matrix, A is of the form

$$t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0$$

and that $a_0 \neq 0$. Find a formula A^{-1} in terms of powers of the matrix, A. Show that A^{-1} exists if and only if $a_0 \neq 0$.

13. In constitutive modeling of the stress and strain tensors, one sometimes considers sums of the form $\sum_{k=0}^{\infty} a_k A^k$ where A is a 3×3 matrix. Show using the Cayley Hamilton theorem that if such a thing makes any sense, you can always obtain it as a finite sum having no more than n terms.

Row Operations

7.1 Elementary Matrices

The elementary matrices result from doing a row operation to the identity matrix.

Definition 7.1.1 *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 it.*

The elementary matrices are given in the following definition.

Definition 7.1.2 *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].

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.

The ij^{th} entry of the elementary matrix which results from switching row k with row l, P^{kl}

$$\left(P^{kl}\right)_{ij} = \delta_{\theta(i)j}$$

where $\theta(i) = i$ for all $i \notin \{k, l\}$ and $\theta(k) = l$ while $\theta(l) = k$. The ij^{th} entry of the elementary matrix which results from adding c times the p^{th} row to the k^{th} row, E^{cp+k}

$$\left(E^{cp+k}\right)_{ij} = \delta_{ij} + c\delta_{ik}\delta_{pj}$$

and the ij^{th} entry of the elementary matrix which results from multiplying the k^{th} row by the nonzero constant, c, E^{ck}

$$\left(E^{ck}\right)_{ij} = \delta_{ij} + (c-1)\,\delta_{ik}\delta_{kj}, \text{ no sum on } k$$

[1] More generally, a permutation matrix is a matrix which comes by permuting the rows of the identity matrix, not just switching two rows.

Theorem 7.1.3 *Let P^{kl} be the elementary matrix which is obtained from switching the k^{th} and the l^{th} rows of the identity matrix. Let E^{ck} denote the elementary matrix which results from multiplying the k^{th} row by the nonzero scalar, c, and let E^{cp+k} denote the elementary matrix obtained from replacing the k^{th} row with c times the p^{th} row added to the k^{th} row. Then if A is an $m \times n$ matrix, multiplication on the left by any of these elementary matrices produces the corresponding row operation on A.*

Proof: First consider P^{kl}.

$$\left(P^{kl}A\right)_{is} = \sum_j \delta_{\theta(i)j} A_{js} = A_{\theta(i)s} = \begin{cases} A_{is} \text{ if } i \notin \{k,l\} \\ A_{ks} \text{ if } i = l \\ A_{ls} \text{ if } i = k \end{cases}.$$

Next consider E^{ck}

$$\begin{aligned} \left(E^{ck}A\right)_{is} &= \sum_j \left(\delta_{ij} + (c-1)\,\delta_{ik}\delta_{kj}\right) A_{js} \\ &= A_{is} + (c-1)\,\delta_{ik}A_{ks}, \text{ no sum on } k \\ &= \begin{cases} cA_{ks} \text{ if } i = k \\ A_{is} \text{ if } i \neq k \end{cases} \end{aligned}$$

Finally consider the case of E^{cp+k}.

$$\begin{aligned} \left(E^{cp+k}A\right)_{is} &= \sum_j \left(\delta_{ij} + c\delta_{ik}\delta_{pj}\right) A_{js} \\ &= A_{is} + c\delta_{ik}A_{ps} \\ &= \begin{cases} A_{is} \text{ if } i \neq k \\ A_{ks} + cA_{ps} \text{ if } i = k \end{cases} \end{aligned}$$

This proves the theorem.

The following corollary follows.

Corollary 7.1.4 *Let A be an $m \times n$ matrix and let R denote the row reduced echelon form obtained from A by row operations. Then there exists a sequence of elementary matrices, $E_1, \cdots, E_p$ such that*

$$\left(E_p E_{p-1} \cdots E_1\right) A = R.$$

The following theorem is also very important.

Theorem 7.1.5 *Let P^{kl}, E^{ck}, and E^{cp+k} be defined in Theorem 7.1.3. Then*

$$\left(P^{kl}\right)^{-1} = P^{kl}, \left(E^{ck}\right)^{-1} = E^{c^{-1}k}, \left(E^{cp+k}\right)^{-1} = E^{-cp+k}.$$

In particular, the inverse of an elementary matrix is an elementary matrix.

Proof: To see the first claim,

$$\begin{aligned} \left(P^{kl}P^{kl}\right)_{is} &= \sum_j \delta_{\theta(i)j}\delta_{\theta(j)s} \\ &= \delta_{\theta(\theta(i))s} = \delta_{is} \end{aligned}$$

the is^{th} entry of I. Thus $P^{kl}P^{kl} = I$ and so $\left(P^{kl}\right)^{-1} = P^{kl}$ as claimed.

Consider the next claim. $E^{c^{-1}k}E^{ck} = I$ because E^{ck} is just the identity in which the k^{th} row is multiplied by c. Then $E^{c^{-1}k}$ multiplies that row by c^{-1} which brings the row back to where it was.

Now consider the last claim. Consider the k^{th} row of $E^{-cp+k}E^{cp+k}$. The k^{th} row of E^{cp+k} is c times the p^{th} row of I added to the k^{th} row of I. This did not change any row but the k^{th} row. Therefore, using Theorem 7.1.3, multiplying on the left by E^{-cp+k} has the effect of taking $-c$ times the p^{th} row of I and adding this to the row just obtained. In other words, E^{-cp+k} undoes what was just done and restores I. This proves the theorem.

Corollary 7.1.6 *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 Corollary 7.1.4 there is a sequence of elementary matrices, $E_1, \cdots, E_p$ such that

$$(E_p E_{p-1} \cdots E_1) A = I.$$

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 7.1.5 each of these in this product is an elementary matrix.

7.2 The Rank Of A Matrix

To begin with there is a definition which includes some terminology.

Definition 7.2.1 *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.*

There are three definitions of the rank of a matrix which are useful and the concept of rank is defined in the following definition.

Definition 7.2.2 *A submatrix of a matrix A is a 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. A given row, $\mathbf{a}_s$ of a matrix, A is a linear combination of rows $\mathbf{a}_{i_1}, \cdots, \mathbf{a}_{i_r}$ if there are scalars, c_j such that $\mathbf{a}_s = \sum_{j=1}^{r} c_j \mathbf{a}_{i_j}$. The row rank of a matrix is the smallest number, r such that every row is a linear combination of some r rows. The column rank of a matrix is the smallest number, r, such that every column is a linear combination of some r columns. Thus the row rank is the dimension of the row space and the column rank is the dimension of the column space. The rank of a matrix, A is denoted by* $\text{rank}(A)$.

The following theorem is proved in the section on the theory of the determinant and is restated here for convenience.

Theorem 7.2.3 *Let A be an $m \times n$ matrix. Then the row rank, column rank and determinant rank are all the same.*

It turns out that row operations are the key to the practical computation of the rank of a matrix.

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

Lemma 7.2.4 *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: 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 7.1.3 on Page 114 $\mathbf{b}_k = E\mathbf{a}_k$ 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.

Corollary 7.2.5 *Let A and B be two $m \times n$ matrices such that B is obtained by applying a row operation to A. Then the two matrices have the same rank.*

Proof: Suppose the column rank of B is r. This means there are r columns whose span yields all the columns of B. By Lemma 7.2.4 every column of A is a linear combination of the corresponding columns in A. Therefore, the rank of A is no larger than the rank of B. But A may also be obtained from B by a row operation. (Why?) Therefore, the same reasoning implies the rank of B is no larger than the rank of A. This proves the corollary.

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

Example 7.2.6 *Find the rank of the following matrix and identify columns whose linear combinations yield all the other columns.*

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 1 & 3 & 6 & 0 & 2 \\ 3 & 7 & 8 & 6 & 6 \end{pmatrix} \tag{7.1}$$

Take (-1) times the first row and add to the second and then take (-3) times the first row and add to the third. This yields

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 0 & 1 & 5 & -3 & 0 \\ 0 & 1 & 5 & -3 & 0 \end{pmatrix}$$

By the above corollary, this matrix has the same rank as the first matrix. Now take (-1) times the second row and add to the third row yielding

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 0 & 1 & 5 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Next take (-2) times the second row and add to the first row. to obtain

$$\begin{pmatrix} 1 & 0 & -9 & 9 & 2 \\ 0 & 1 & 5 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{7.2}$$

Each of these row operations did not change the rank of the matrix. It is clear that linear combinations of the first two columns yield every other column so the rank of the matrix is no larger than 2. However, it is also clear that the determinant rank is at least 2 because, deleting every column other than the first two and every zero row yields the 2×2 identity matrix having determinant 1.

By Lemma 7.2.4 the first two columns of the original matrix yield all other columns as linear combinations.

Example 7.2.7 *Find the rank of the following matrix and identify columns whose linear combinations yield all the other columns.*

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 1 & 2 & 6 & 0 & 2 \\ 3 & 6 & 8 & 6 & 6 \end{pmatrix} \tag{7.3}$$

Take (-1) times the first row and add to the second and then take (-3) times the first row and add to the last row. This yields

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 0 & 0 & 5 & -3 & 0 \\ 0 & 0 & 5 & -3 & 0 \end{pmatrix}$$

Now multiply the second row by $1/5$ and add 5 times it to the last row.

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 0 & 0 & 1 & -3/5 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Add (-1) times the second row to the first.

$$\begin{pmatrix} 1 & 2 & 0 & \frac{18}{5} & 2 \\ 0 & 0 & 1 & -3/5 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{7.4}$$

The determinant rank is at least 2 because deleting the second, third and fifth columns as well as the last row yields the 2×2 identity matrix. On the other hand, the rank is no more than two because clearly every column can be obtained as a linear combination of the first and third columns. Also, by Lemma 7.2.4 every column of the original matrix is a linear combination of the first and third columns of that matrix.

The matrix, 7.4 is the row reduced echelon form for the matrix, 7.3 and 7.2 is the row reduced echelon form for 7.1.

7.3 The Row Reduced Echelon Form

Definition 7.3.1 *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 7.3.2 *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. 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 $\mathbb{F}^{m-1}$. Use this 1 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.

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 7.3.3 *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 7.3.4 *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 7.3.5 *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 $i_1, i_2, \cdots, i_{j-1}$. By Lemma 7.2.4 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 7.2.4. 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 7.2.4, 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.

Corollary 7.3.6 *The rank of a matrix equals the number of nonzero pivot columns. Furthermore, every column is contained in the span of the pivot columns.*

Proof: Row rank, determinant rank, and column rank are all the same so it suffices to consider only column rank. Write the row reduced echelon form for the matrix. From Corollary 7.2.5 this row reduced matrix has the same rank as the original matrix. Deleting all the zero rows and all the columns in the row reduced echelon form which do not correspond to a pivot column, yields an $r \times r$ identity submatrix in which r is the number of pivot columns. Thus the rank is at least r. Now from the construction of the row reduced echelon form, every column is a linear combination of these r columns. Therefore, the rank is no more than r. This proves the corollary.

Definition 7.3.7 *Let A be an $m \times n$ matrix having rank, r. Then the nullity of A is defined to be $n - r$. Also define* $\ker(A) \equiv \{\mathbf{x} \in \mathbb{F}^n : A\mathbf{x} = \mathbf{0}\}$.

Observation 7.3.8 *Note that* $\ker(A)$ *is a subspace because if a, b are scalars and $\mathbf{x}, \mathbf{y}$ are vectors in* $\ker(A)$*, then*

$$A(a\mathbf{x} + b\mathbf{y}) = aA\mathbf{x} + bA\mathbf{y} = \mathbf{0} + \mathbf{0} = \mathbf{0}$$

Recall that the dimension of the column space of a matrix equals its rank and since the column space is just $A(\mathbb{F}^n)$, the rank is just the dimension of $A(\mathbb{F}^n)$. The next theorem shows that the nullilty equals the dimension of $\ker(A)$.

Theorem 7.3.9 *Let A be an $m \times n$ matrix. Then* $\operatorname{rank}(A) + \dim(\ker(A)) = n$..

Proof: Since $\ker(A)$ is a subspace, there exists a basis for $\ker(A)$, $\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$. Now this basis may be extended to a basis of $\mathbb{F}^n$, $\{\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{y}_1, \cdots, \mathbf{y}_{n-k}\}$. If $\mathbf{z} \in A(\mathbb{F}^n)$, then there exist scalars, $c_i, i = 1, \cdots, k$ and $d_i, i = 1, \cdots, n-k$ such that

$$\begin{aligned} \mathbf{z} &= A\left(\sum_{i=1}^{k} c_i \mathbf{x}_i + \sum_{i=1}^{n-k} d_i \mathbf{y}_i\right) \\ &= \sum_{i=1}^{k} c_i A\mathbf{x}_i + \sum_{i=1}^{n-k} d_i A\mathbf{y}_i = \sum_{i=1}^{n-k} d_i A\mathbf{y}_i \end{aligned}$$

and this shows $\operatorname{span}(A\mathbf{y}_1, \cdots, A\mathbf{y}_{n-k}) = A(\mathbb{F}^n)$. Are the vectors, $\{A\mathbf{y}_1, \cdots, A\mathbf{y}_{n-k}\}$ independent? Suppose

$$\sum_{i=1}^{n-k} c_i A\mathbf{y}_i = \mathbf{0}.$$

Then since A is linear, it follows

$$A\left(\sum_{i=1}^{n-k} c_i \mathbf{y}_i\right) = \mathbf{0}$$

showing that $\sum_{i=1}^{n-k} c_i \mathbf{y}_i \in \ker(A)$. Therefore, there exists constants, $d_i, i = 1, \cdots, k$ such that

$$\sum_{i=1}^{n-k} c_i \mathbf{y}_i = \sum_{j=1}^{k} d_j \mathbf{x}_j. \tag{7.5}$$

If any of these constants, d_i or c_i is not equal to zero then

$$\mathbf{0} = \sum_{j=1}^{k} d_j \mathbf{x}_j + \sum_{i=1}^{k} (-c_i)\, \mathbf{y}_i$$

and this would be a nontrivial linear combination of the vectors, $\{\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{y}_1, \cdots, \mathbf{y}_{n-k}\}$ which equals zero contrary to the fact that $\{\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{y}_1, \cdots, \mathbf{y}_{n-k}\}$ is a basis. Therefore, all the constants, d_i and c_i in 7.5 must equal zero. It follows the vectors, $\{A\mathbf{y}_1, \cdots, A\mathbf{y}_{n-k}\}$ are linearly independent and so they must be a basis $A(\mathbb{F}^n)$. Therefore, $\operatorname{rank}(A) + \dim(\ker(A)) = n - k + k = n$. This proves the theorem.

7.4 Exercises

1. Find the rank and nullity 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.

   (a) $\begin{pmatrix} 0 & 1 & 0 & 2 & 1 & 2 & 2 \\ 0 & 3 & 2 & 12 & 1 & 6 & 8 \\ 0 & 1 & 1 & 5 & 0 & 2 & 3 \\ 0 & 2 & 1 & 7 & 0 & 3 & 4 \end{pmatrix}$

   (b) $\begin{pmatrix} 0 & 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 3 & 2 & 6 & 0 & 5 & 4 \\ 0 & 1 & 1 & 2 & 0 & 2 & 2 \\ 0 & 2 & 1 & 4 & 0 & 3 & 2 \end{pmatrix}$

(c) $\begin{pmatrix} 0 & 1 & 0 & 2 & 1 & 1 & 2 \\ 0 & 3 & 2 & 6 & 1 & 5 & 1 \\ 0 & 1 & 1 & 2 & 0 & 2 & 1 \\ 0 & 2 & 1 & 4 & 0 & 3 & 1 \end{pmatrix}$

2. Suppose A is an $m \times n$ matrix. Explain why the rank of A is always no larger than $\min(m,n)$.

3. Suppose A is an $m \times n$ matrix in which $m \leq n$. Suppose also that the rank of A equals m. Show that A maps $\mathbb{F}^n$ onto $\mathbb{F}^m$. **Hint:** The vectors $\mathbf{e}_1, \cdots, \mathbf{e}_m$ occur as columns in the row reduced echelon form for A.

4. Suppose A is an $m \times n$ matrix in which $m \geq n$. Suppose also that the rank of A equals n. Show that A is one to one. **Hint:** If not, there exists a vector, $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{0}$, and this implies at least one column of A is a linear combination of the others. Show this would require the column rank to be less than n.

5. Explain why an $n \times n$ matrix, A is both one to one and onto if and only if its rank is n.

6. Suppose A is an $m \times n$ matrix and $\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}$ is a linearly independent set of vectors in $A(\mathbb{F}^n) \subseteq \mathbb{F}^m$. Now suppose $A(\mathbf{z}_i) = \mathbf{w}_i$. Show $\{\mathbf{z}_1, \cdots, \mathbf{z}_k\}$ is also independent.

7. Suppose A is an $m \times n$ matrix and B is an $n \times p$ matrix. Show that

$$\dim(\ker(AB)) \leq \dim(\ker(A)) + \dim(\ker(B)).$$

Hint: Consider the subspace, $B(\mathbb{F}^p) \cap \ker(A)$ and suppose a basis for this subspace is $\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}$. Now suppose $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ is a basis for $\ker(B)$. Let $\{\mathbf{z}_1, \cdots, \mathbf{z}_k\}$ be such that $B\mathbf{z}_i = \mathbf{w}_i$ and argue that

$$\ker(AB) \subseteq \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_r, \mathbf{z}_1, \cdots, \mathbf{z}_k).$$

Here is how you do this. Suppose $AB\mathbf{x} = \mathbf{0}$. Then $B\mathbf{x} \in \ker(A) \cap B(\mathbb{F}^p)$ and so $B\mathbf{x} = \sum_{i=1}^k B\mathbf{z}_i$ showing that

$$\mathbf{x} - \sum_{i=1}^k \mathbf{z}_i \in \ker(B).$$

7.5 *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. 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, $A\mathbf{x} = \mathbf{y}$. The method lacks generality but is of interest just the same.

Example 7.5.1 *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 I mean above by saying the method lacks generality.

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 $i < j$.

7.6 Finding The LU Decomposition

There is 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 and is called the multiplier method. It is due to Dolittle.

Example 7.6.1 *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 method for updating the matrix on the left. The 2 in the second entry of the first column is there because -2 times the first row of A added to the second row of A produced a 0. 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, 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, stop because the matrix on the right is upper triangular. An LU decomposition is the above.

The justification for this gimmick will be given later.

Example 7.6.2 *Find an LU decomposition for* $A = \begin{pmatrix} 1 & 2 & 1 & 2 & 1 \\ 2 & 0 & 2 & 1 & 1 \\ 2 & 3 & 1 & 3 & 2 \\ 1 & 0 & 1 & 1 & 2 \end{pmatrix}$.

This time everything is done at once for a whole column. This saves trouble. 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.

7.7 Solving Linear 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 7.7.1 *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

$$\mathbf{x} = \begin{pmatrix} -\frac{3}{5} + \frac{7}{5}t \\ \frac{9}{5} - \frac{11}{5}t \\ t \\ -1 \end{pmatrix}, t \in \mathbb{R}.$$

Work this out by hand and you will see the advantage of working only with triangular matrices.

It may seem like a trivial thing but it is used because it cuts down on the number of operations involved in finding a solution to a system of equations enough that it makes a difference for large systems.

7.8 The PLU Decomposition

As indicated above, some matrices don't have an LU decomposition. Here is an example.

$$M = \begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 2 & 3 & 0 \\ 4 & 3 & 1 & 1 \end{pmatrix} \tag{7.6}$$

In this case, there is another decomposition which is useful called a PLU decomposition. Here P is a permutation matrix.

Example 7.8.1 *Find a PLU decomposition for the above matrix in 7.6.*

Proceed as before trying to find the row echelon form of the matrix. First add -1 times the first row to the second row and then add -4 times the first to the third. This yields

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 4 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & 0 & 0 & -2 \\ 0 & -5 & -11 & -7 \end{pmatrix}$$

There is no way to do only row operations involving replacing a row with itself added to a multiple of another row to the second matrix in such a way as to obtain an upper triangular matrix. Therefore, consider M with the bottom two rows switched.

$$M' = \begin{pmatrix} 1 & 2 & 3 & 2 \\ 4 & 3 & 1 & 1 \\ 1 & 2 & 3 & 0 \end{pmatrix}.$$

Now try again with this matrix. First take -1 times the first row and add to the bottom row and then take -4 times the first row and add to the second row. This yields

$$\begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix}$$

The second matrix is upper triangular and so the LU decomposition of the matrix, M' is

$$\begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix}.$$

Thus $M' = PM = LU$ where L and U are given above. Therefore, $M = P^2M = PLU$ and so

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 2 & 3 & 0 \\ 4 & 3 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix}$$

This process can always be followed and so there always exists a PLU decomposition of a given matrix even though there isn't always an LU decomposition.

Example 7.8.2 *Use the PLU decomposition of $M \equiv \begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 2 & 3 & 0 \\ 4 & 3 & 1 & 1 \end{pmatrix}$ to solve the system $M\mathbf{x} = \mathbf{b}$ where $\mathbf{b} = (1,2,3)^T$.*

Let $U\mathbf{x} = \mathbf{y}$ and consider $PL\mathbf{y} = \mathbf{b}$. In other words, solve,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Then multiplying both sides by P gives

$$\begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$$

and so

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

Now $U\mathbf{x} = \mathbf{y}$ and so it only remains to solve

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -11 & -7 \\ 0 & 0 & 0 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

which yields

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \frac{1}{5} + \frac{7}{5}t \\ \frac{9}{10} - \frac{11}{5}t \\ t \\ -\frac{1}{2} \end{pmatrix} : t \in \mathbb{R}.$$

7.9 Justification For The Multiplier Method

Why does the multiplier method work for finding the LU decomposition? Suppose A is a matrix which has the property that the row reduced echelon form for A may be achieved using only the row operations which involve replacing a row with itself added to a multiple of another row. It is not ever necessary to switch rows. Thus every row which is replaced using this row operation in obtaining the echelon form may be modified by using a row which is above it. Furthermore, in the multiplier method for finding the LU decomposition, we zero out the elements below the pivot entry in first column and then the next and so on when scanning from the left. In terms of elementary matrices, this means the row operations used to reduce A to upper triangular form correspond to multiplication on the left by lower triangular matrices having all ones down the main diagonal.and the sequence of elementary matrices which row reduces A has the property that in scanning the list of elementary matrices from the right to the left, this list consists of several matrices which involve only changes from the identity in the first column, then several which involve only changes from the identity in the second column and so forth. More precisely, $E_p \cdots E_1 A = U$ where U is upper triangular, each E_i is a lower triangular elementary matrix having all ones down the main diagonal, for some r_i, each of $E_{r_1} \cdots E_1$ differs from the identity only in the first column, each of $E_{r_2} \cdots E_{r_1+1}$ differs from the identity only in the second column and so forth. Therefore, $A = \overbrace{E_1^{-1} \cdots E_{p-1}^{-1} E_p^{-1}}^{\text{Will be } L} U$. You multiply the inverses in the reverse order. Now each of the E_i^{-1} is also lower triangular with 1 down the main diagonal. Therefore their product has this property. Recall also that if E_i equals the identity matrix except for having an a in the j^{th} column somewhere below the main diagonal, E_i^{-1} is obtained by replacing the a in E_i with $-a$ thus explaining why we replace with -1 times the multiplier in computing L. In the case where A is a $3 \times m$ matrix, $E_1^{-1} \cdots E_{p-1}^{-1} E_p^{-1}$ is of the form

$$\begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ b & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & c & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix}.$$

Note that scanning from left to right, the first two in the product involve changes in the identity only in the first column while in the third matrix, the change is only in the second. If the entries in the first column had been zeroed out in a different order, the following would have resulted.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ b & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & c & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix}$$

However, it is important to be working from the left to the right, one column at a time.

A similar observation holds in any dimension. Multiplying the elementary matrices which involve a change only in the j^{th} column you obtain A equal to an upper triangular, $n \times m$ matrix, U which is multiplied by a sequence of lower triangular matrices on its left which is of the following form in which the a_{ij} are negatives of multipliers used in row reducing to

an upper triangular matrix.

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ a_{11} & 1 & & & & \vdots \\ \vdots & 0 & \ddots & & & \vdots \\ \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ a_{1,n-1} & 0 & 0 & \cdots & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & & & & \vdots \\ \vdots & a_{21} & \ddots & & & \vdots \\ \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & 0 \\ 0 & a_{2,n-2} & 0 & \cdots & \cdots & 1 \end{pmatrix} \cdots$$

$$\cdots \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & & & & \vdots \\ \vdots & 0 & \ddots & & & \vdots \\ \vdots & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & 1 & 0 \\ 0 & 0 & 0 & \cdots & a_{n,n-1} & 1 \end{pmatrix}$$

From the matrix multiplication, this product equals

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ a_{11} & 1 & & & & \vdots \\ a_{12} & a_{21} & \ddots & & & \vdots \\ \vdots & a_{22} & a_{31} & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & 1 & 0 \\ a_{1,n-1} & a_{2,n-2} & a_{3,n-3} & \cdots & a_{n,n-1} & 1 \end{pmatrix}$$

Notice how the end result of the matrix multiplication made no change in the a_{ij}. It just filled in the empty spaces with the a_{ij} which occured in one of the matrices in the product. This is why, in computing L, it is sufficient to begin with the left column and work column by column toward the right, replacing entries with the negative of the multiplier used in the row operation which produces a zero in that entry.

7.10 Exercises

1. Find a LU decomposition of $\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 3 \\ 1 & 2 & 3 \end{pmatrix}$.

2. Find a LU decomposition of $\begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 3 & 2 & 1 \\ 5 & 0 & 1 & 3 \end{pmatrix}$.

3. Find a PLU decomposition of $\begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 2 & 1 & 1 \end{pmatrix}$.

4. Find a PLU decomposition of $\begin{pmatrix} 1 & 2 & 1 & 2 & 1 \\ 2 & 4 & 2 & 4 & 1 \\ 1 & 2 & 1 & 3 & 2 \end{pmatrix}$.

5. Find a PLU decomposition of $\begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 2 & 4 & 1 \\ 3 & 2 & 1 \end{pmatrix}$.

6. Is there only one LU decomposition for a given matrix? **Hint:** Consider the equation

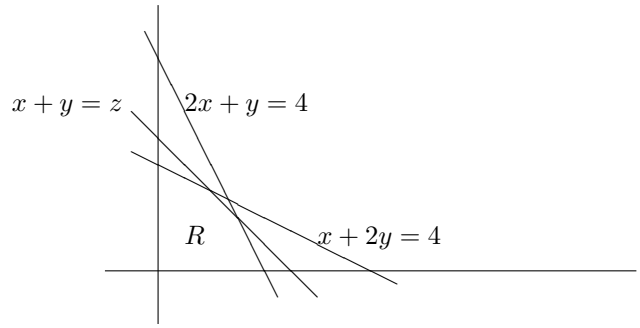
$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Linear Programming

8.1 Simple Geometric Considerations

One of the most important uses of row operations is in solving linear program problems which involve maximizing a linear function subject to inequality constraints determined from linear equations. Here is an example. A certain hamburger store has 9000 hamburger pattys to use in one week and a limitless supply of special sauce, lettuce, tomatos, onions, and buns. They sell two types of hamburgers, the big stack and the basic burger. It has also been determined that the employees cannot prepare more than 9000 of either type in one week. The big stack, popular with the teen agers from the local high school, involves two pattys, lots of delicious sauce, condiments galore, and a divider between the two pattys. The basic burger, very popular with children, involves only one patty and some pickles and ketchup. Demand for the basic burger is twice what it is for the big stack. What is the maximum number of hamburgers which could be sold in one week given the above limitations?

Let x be the number of basic burgers and y the number of big stacks which could be sold in a week. Thus it is desired to maximize $z = x + y$ subject to the above constraints. The total number of pattys is 9000 and so the number of pattys used is $x+2y$. This number must satisfy $x+2y \leq 9000$ because there are only 9000 pattys available. Because of the limitation on the number the employees can prepare and the demand, it follows $2x + y \leq 9000$. You never sell a negative number of hamburgers and so $x, y \geq 0$. In simpler terms the problem reduces to maximizing $z = x + y$ subject to the two constraints, $x + 2y \leq 9000$ and $2x + y \leq 9000$. This problem is pretty easy to solve geometrically. Consider the following picture in which R labels the region described by the above inequalities and the line $z = x+y$ is shown for a particular value of z.



As you make z larger this line moves away from the origin, always having the same slope

and the desired solution would consist of a point in the region, R which makes z as large as possible or equivalently one for which the line is as far as possible from the origin. Clearly this point is the point of intersection of the two lines, $(3000, 3000)$ and so the maximum value of the given function is 6000. Of course this type of procedure is fine for a situation in which there are only two variables but what about a similar problem in which there are very many variables. In reality, this hamburger store makes many more types of burgers than those two and there are many considerations other than demand and available pattys. Each will likely give you a constraint which must be considered in order to solve a more realistic problem and the end result will likely be a problem in many dimensions, probably many more than three so your ability to draw a picture will get you nowhere for such a problem. Another method is needed. This method is the topic of this section. I will illustrate with this particular problem. Let $x_1 = x$ and $y = x_2$. Also let x_3 and x_4 be nonnegative variables such that

$$x_1 + 2x_2 + x_3 = 9000,\ 2x_1 + x_2 + x_4 = 9000.$$

To say that x_3 and x_4 are nonnegative is the same as saying $x_1 + 2x_2 \leq 9000$ and $2x_1 + x_2 \leq 9000$ and these variables are called slack variables at this point. They are called this because they "take up the slack". I will discuss these more later. First a general situation is considered.

8.2 The Simplex Tableau

Here is some notation.

Definition 8.2.1 *Let* $\mathbf{x}, \mathbf{y}$ *be vectors in* $\mathbb{R}^q$. *Then* $\mathbf{x} \leq \mathbf{y}$ *means for each* $i, x_i \leq y_i$.

The problem is as follows:

Let A be an $m \times (m+n)$ real matrix of rank m. It is desired to find $\mathbf{x} \in \mathbb{R}^{n+m}$ such that $\mathbf{x}$ satisfies the constraints,

$$\mathbf{x} \geq \mathbf{0},\ A\mathbf{x} = \mathbf{b} \tag{8.1}$$

and out of all such $\mathbf{x}$,

$$z \equiv \sum_{i=1}^{m+n} c_i x_i$$

is as large (or small) as possible. This is usually refered to as maximizing or minimizing z subject to the above constraints. First I will consider the constraints.

Let $A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_{n+m} \end{pmatrix}$. First you find a vector, $\mathbf{x}^0 \geq \mathbf{0}$, $A\mathbf{x}^0 = \mathbf{b}$ such that n of the components of this vector equal 0. Letting $i_1, \cdots, i_n$ be the positions of $\mathbf{x}^0$ for which $x_i^0 = 0$, suppose also that $\{\mathbf{a}_{j_1}, \cdots, \mathbf{a}_{j_m}\}$ is linearly independent for j_i the other positions of $\mathbf{x}^0$. Geometrically, this means that $\mathbf{x}^0$ is a corner of the feasible region, those $\mathbf{x}$ which satisfy the constraints. This is called a basic feasible solution. Also define

$$\begin{aligned} \mathbf{c}_B &\equiv (c_{j_1} \cdots, c_{j_m}),\ \mathbf{c}_F \equiv (c_{i_1}, \cdots, c_{i_n}) \\ \mathbf{x}_B &\equiv (x_{j_1}, \cdots, x_{j_m}),\ \mathbf{x}_F \equiv (x_{i_1}, \cdots, x_{i_n}). \end{aligned}$$

and

$$z^0 \equiv z\left(\mathbf{x}^0\right) = \begin{pmatrix} \mathbf{c}_B & \mathbf{c}_F \end{pmatrix} \begin{pmatrix} \mathbf{x}_B^0 \\ \mathbf{x}_F^0 \end{pmatrix} = \mathbf{c}_B \mathbf{x}_B^0$$

since $\mathbf{x}_F^0 = \mathbf{0}$. The variables which are the components of the vector $\mathbf{x}_B$ are called the **basic variables** and the variables which are the entries of $\mathbf{x}_F$ are called the **free variables.** You

set $\mathbf{x}_F = \mathbf{0}$. Now $\left(\mathbf{x}^0, z^0\right)^T$ is a solution to

$$\begin{pmatrix} A & \mathbf{0} \\ -\mathbf{c} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ z \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ 0 \end{pmatrix}$$

along with the constraints $\mathbf{x} \geq \mathbf{0}$. Writing the above in augmented matrix form yields

$$\begin{pmatrix} A & \mathbf{0} & \mathbf{b} \\ -\mathbf{c} & 1 & 0 \end{pmatrix} \tag{8.2}$$

Permute the columns and variables on the left if necessary to write the above in the form

$$\begin{pmatrix} B & F & \mathbf{0} \\ -\mathbf{c}_B & -\mathbf{c}_F & 1 \end{pmatrix} \begin{pmatrix} \mathbf{x}_B \\ \mathbf{x}_F \\ z \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ 0 \end{pmatrix} \tag{8.3}$$

or equivalently in the augmented matrix form keeping track of the variables on the bottom as

$$\begin{pmatrix} B & F & \mathbf{0} & \mathbf{b} \\ -\mathbf{c}_B & -\mathbf{c}_F & 1 & 0 \\ \mathbf{x}_B & \mathbf{x}_F & 0 & 0 \end{pmatrix}. \tag{8.4}$$

Here B pertains to the variables $x_{i_1}, \cdots, x_{j_m}$ and is an $m \times m$ matrix with linearly independent columns, $\{\mathbf{a}_{j_1}, \cdots, \mathbf{a}_{j_m}\}$, and F is an $m \times n$ matrix. Now it is assumed that

$$\begin{pmatrix} B & F \end{pmatrix} \begin{pmatrix} \mathbf{x}_B^0 \\ \mathbf{x}_F^0 \end{pmatrix} = \begin{pmatrix} B & F \end{pmatrix} \begin{pmatrix} \mathbf{x}_B^0 \\ \mathbf{0} \end{pmatrix} = B\mathbf{x}_B^0 = \mathbf{b}$$

and since B is assumed to have rank m, it follows

$$\mathbf{x}_B^0 = B^{-1}\mathbf{b} \geq \mathbf{0}. \tag{8.5}$$

This is very important to observe. $B^{-1}\mathbf{b} \geq \mathbf{0}$!

Do row operations on the top part of the matrix,

$$\begin{pmatrix} B & F & \mathbf{0} & \mathbf{b} \\ -\mathbf{c}_B & -\mathbf{c}_F & 1 & 0 \end{pmatrix} \tag{8.6}$$

and obtain its row reduced echelon form. Then after these row operations the above becomes

$$\begin{pmatrix} I & B^{-1}F & \mathbf{0} & B^{-1}\mathbf{b} \\ -\mathbf{c}_B & -\mathbf{c}_F & 1 & 0 \end{pmatrix}. \tag{8.7}$$

where $B^{-1}\mathbf{b} \geq 0$. Next do another row operation in order to get a $\mathbf{0}$ where you see a $-\mathbf{c}_B$. Thus

$$\begin{pmatrix} I & B^{-1}F & \mathbf{0} & B^{-1}\mathbf{b} \\ \mathbf{0} & \mathbf{c}_B B^{-1}F' - \mathbf{c}_F & 1 & \mathbf{c}_B B^{-1}\mathbf{b} \end{pmatrix} \tag{8.8}$$

$$= \begin{pmatrix} I & B^{-1}F & \mathbf{0} & B^{-1}\mathbf{b} \\ \mathbf{0} & \mathbf{c}_B B^{-1}F' - \mathbf{c}_F & 1 & \mathbf{c}_B \mathbf{x}_B^0 \end{pmatrix}$$

$$= \begin{pmatrix} I & B^{-1}F & \mathbf{0} & B^{-1}\mathbf{b} \\ \mathbf{0} & \mathbf{c}_B B^{-1}F - \mathbf{c}_F & 1 & z^0 \end{pmatrix} \tag{8.9}$$

The reason there is a z^0 on the bottom right corner is that $\mathbf{x}_F = 0$ and $\left(\mathbf{x}_B^0, \mathbf{x}_F^0, z^0\right)^T$ is a solution of the system of equations represented by the above augmented matrix because it is

a solution to the system of equations corresponding to the system of equations represented by 8.6 and row operations leave solution sets unchanged. Note how attractive this is. The z_0 is the value of z at the point $\mathbf{x}^0$. The augmented matrix of 8.9 is called the simplex tableau and it is the beginning point for the simplex algorithm to be described a little later. It is very convenient to express the simplex tableau in the above form in which the variables are possibly permuted in order to have $\begin{pmatrix} I \\ \mathbf{0} \end{pmatrix}$ on the left side. However, as far as the simplex algorithm is concerned it is not necessary to be permuting the variables in this manner. Starting with 8.9 you could permute the variables and columns to obtain an augmented matrix in which the variables are in their original order. What is really required for the simplex tableau?

It is an augmented $m+1 \times m+n+2$ matrix which represents a system of equations which has the same set of solutions, $(\mathbf{x},z)^T$ as the system whose augmented matrix is

$$\begin{pmatrix} A & \mathbf{0} & \mathbf{b} \\ -\mathbf{c} & 1 & 0 \end{pmatrix}$$

(Possibly the variables for $\mathbf{x}$ are taken in another order.) There are m linearly independent columns in the first $m+n$ columns for which there is only one nonzero entry, a 1 in one of the first m rows, the "simple columns", the other first $m+n$ columns being the "nonsimple columns". As in the above, the variables corresponding to the simple columns are $\mathbf{x}_B$, the basic variables and those corresponding to the nonsimple columns are $\mathbf{x}_F$, the free variables. Also, the top m entries of the last column on the right are nonnegative. This is the description of a simplex tableau.

In a simplex tableau it is easy to spot a basic feasible solution. You can see one quickly by setting the variables, $\mathbf{x}_F$ corresponding to the nonsimple columns equal to zero. Then the other variables, corresponding to the simple columns are each equal to a nonnegative entry in the far right column. Lets call this an "obvious basic feasible solution". If a solution is obtained by setting the variables corresponding to the nonsimple columns equal to zero and the variables corresponding to the simple columns equal to zero this will be referred to as an "obvious" solution. Lets also call the first $m+n$ entries in the bottom row the "bottom left row". In a simplex tableau, the entry in the bottom right corner gives the value of the variable being maximized or minimized when the obvious basic feasible solution is chosen.

The following is a special case of the general theory presented above and shows how such a special case can be fit into the above framework. The following example is rather typical of the sorts of problems considered. It involves inequality constraints instead of $A\mathbf{x} = \mathbf{b}$. This is handled by adding in "slack variables" as explained below.

Example 8.2.2 *Consider $z = x_1 - x_2$ subject to the constraints, $x_1+2x_2 \leq 10, x_1+2x_2 \geq 2$, and $2x_1 + x_2 \leq 6, x_i \geq 0$. Find a simplex tableau for a problem of the form $\mathbf{x} \geq \mathbf{0}, A\mathbf{x} = \mathbf{b}$ which is equivalent to the above problem.*

You add in slack variables. These are positive variables, one for each of the first three constraints, which change the first three inequalities into equations. Thus the first three inequalities become $x_1+2x_2+x_3 = 10, x_1+2x_2-x_4 = 2$, and $2x_1+x_2+x_5 = 6, x_1, x_2, x_3, x_4, x_5 \geq 0$. Now it is necessary to find a basic feasible solution. You mainly need to find a positive solution to the equations,

$$\begin{array}{c} x_1 + 2x_2 + x_3 = 10 \\ x_1 + 2x_2 - x_4 = 2 \\ 2x_1 + x_2 + x_5 = 6 \end{array} .$$

the solution set for the above system is given by

$$x_2 = \frac{2}{3}x_4 - \frac{2}{3} + \frac{1}{3}x_5, x_1 = -\frac{1}{3}x_4 + \frac{10}{3} - \frac{2}{3}x_5, x_3 = -x_4 + 8.$$

An easy way to get a basic feasible solution is to let $x_4 = 8$ and $x_5 = 1$. Then a feasible solution is

$$(x_1, x_2, x_3, x_4, x_5) = (0, 5, 0, 8, 1)\,.$$

It follows $z^0 = -5$ and the matrix 8.2, $\begin{pmatrix} A & \mathbf{0} & \mathbf{b} \\ -\mathbf{c} & 1 & 0 \end{pmatrix}$ with the variables kept track of on the bottom is

$$\begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 0 & 10 \\ 1 & 2 & 0 & -1 & 0 & 0 & 2 \\ 2 & 1 & 0 & 0 & 1 & 0 & 6 \\ -1 & 1 & 0 & 0 & 0 & 1 & 0 \\ x_1 & x_2 & x_3 & x_4 & x_5 & 0 & 0 \end{pmatrix}$$

and the first thing to do is to permute the columns so that the list of variables on the bottom will have x_1 and x_3 at the end.

$$\begin{pmatrix} 2 & 0 & 0 & 1 & 1 & 0 & 10 \\ 2 & -1 & 0 & 1 & 0 & 0 & 2 \\ 1 & 0 & 1 & 2 & 0 & 0 & 6 \\ 1 & 0 & 0 & -1 & 0 & 1 & 0 \\ x_2 & x_4 & x_5 & x_1 & x_3 & 0 & 0 \end{pmatrix}$$

Next, as described above, take the row reduced echelon form of the top three lines of the above matrix. This yields

$$\begin{pmatrix} 1 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 5 \\ 0 & 1 & 0 & 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & \frac{3}{2} & -\frac{1}{2} & 0 & 1 \end{pmatrix}.$$

Now do row operations to

$$\begin{pmatrix} 1 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 5 \\ 0 & 1 & 0 & 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & \frac{3}{2} & -\frac{1}{2} & 0 & 1 \\ 1 & 0 & 0 & -1 & 0 & 1 & 0 \end{pmatrix}$$

to finally obtain

$$\begin{pmatrix} 1 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 5 \\ 0 & 1 & 0 & 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & \frac{3}{2} & -\frac{1}{2} & 0 & 1 \\ 0 & 0 & 0 & -\frac{3}{2} & -\frac{1}{2} & 1 & -5 \end{pmatrix}$$

and this is a simplex tableau. The variables are $x_2, x_4, x_5, x_1, x_3, z$.

It isn't as hard as it may appear from the above. Lets not permute the variables and simply find an acceptable simplex tableau as described above.

Example 8.2.3 *Consider $z = x_1 - x_2$ subject to the constraints, $x_1 + 2x_2 \leq 10, x_1 + 2x_2 \geq 2$, and $2x_1 + x_2 \leq 6, x_i \geq 0$. Find a simplex tableau.*

Adding in slack variables, an augmented matrix which is descriptive of the constraints is

$$\begin{pmatrix} 1 & 2 & 1 & 0 & 0 & 10 \\ 1 & 2 & 0 & -1 & 0 & 6 \\ 2 & 1 & 0 & 0 & 1 & 6 \end{pmatrix}$$

The obvious solution is not feasible because of that -1 in the fourth column. Consider the second column and select the 2 as a pivot to zero out that which is above and below the 2. This is because that 2 satisfies the criterion for being chosen as a pivot.

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

This one is good. The obvious solution is now feasible. You can now assemble the simplex tableau. The first step is to include a column and row for z. This yields

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

Now you need to get zeros in the right places so the simple columns will be preserved as simple columns. This means you need to zero out the 1 in the third column on the bottom. A simplex tableau is now

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

Note it is not the same one obtained earlier. There is no reason a simplex tableau should be unique. In fact, it follows from the above general description that you have one for each basic feasible point of the region determined by the constraints.

8.3 The Simplex Algorithm

8.3.1 Maximums

The simplex algorithm takes you from one basic feasible solution to another while maximizing or minimizing the function you are trying to maximize or minimize. Algebraically, it takes you from one simplex tableau to another in which the lower right corner either increases in the case of maximization or decreases in the case of minimization.

I will continue writing the simplex tableau in such a way that the simple columns having only one entry nonzero are on the left. As explained above, this amounts to permuting the variables. I will do this because it is possible to describe what is going on without onerous notation. However, in the examples, I won't worry so much about it. Thus, from a basic feasible solution, a simplex tableau of the following form has been obtained in which the columns for the basic variables, $\mathbf{x}_B$ are listed first and $\mathbf{b} \geq \mathbf{0}$.

$$\begin{pmatrix} I & F & \mathbf{0} & \mathbf{b} \\ \mathbf{0} & \mathbf{c} & 1 & z^0 \end{pmatrix} \tag{8.10}$$

Let $x_i^0 = b_i$ for $i = 1, \cdots, m$ and $x_i^0 = 0$ for $i > m$. Then $\left(\mathbf{x}^0, z^0\right)$ is a solution to the above system and since $\mathbf{b} \geq \mathbf{0}$, it follows $\left(\mathbf{x}^0, z^0\right)$ is a basic feasible solution.

If $c_i < 0$ for some i, and if $F_{ji} \leq 0$ so that a whole column of $\begin{pmatrix} F \\ \mathbf{c} \end{pmatrix}$ is ≤ 0 with the bottom entry < 0, then letting x_i be the variable corresponding to that column, you could

leave all the other entries of $\mathbf{x}_F$ equal to zero but change x_i to be positive. Let the new vector be denoted by $\mathbf{x}'_F$ and letting $\mathbf{x}'_B = \mathbf{b} - F\mathbf{x}'_F$ it follows

$$\begin{aligned}(\mathbf{x}'_B)_k &= b_k - \sum_j F_{kj} (\mathbf{x}_F)_j \\ &= b_k - F_{ki} x_i \geq 0\end{aligned}$$

Now this shows $(\mathbf{x}'_B, \mathbf{x}'_F)$ is feasible whenever $x_i > 0$ and so you could let x_i become arbitrarily large and positive and conclude there is no maximum for z because

$$z = -\mathbf{c}\mathbf{x}'_F + z^0 = (-c_i) x_i + z^0 \tag{8.11}$$

If this happens in a simplex tableau, you can say there is no maximum and stop.

What if $\mathbf{c} \geq \mathbf{0}$? Then $z = z^0 - \mathbf{c}\mathbf{x}_F$ and to satisfy the constraints, $\mathbf{x}_F \geq 0$. Therefore, in this case, z^0 is the largest possible value of z and so the maximum has been found. You stop when this occurs. Next I explain what to do if neither of the above stopping conditions hold.

The only case which remains is that some $c_i < 0$ and some $F_{ji} > 0$. You pick a column in $\begin{pmatrix} F \\ \mathbf{c} \end{pmatrix}$ in which $c_i < 0$, usually the one for which c_i is the largest in absolute value. You pick $F_{ji} > 0$ as a pivot entry, divide the j^{th} row by F_{ji} and then use to obtain zeros above F_{ji} and below F_{ji}, thus obtaining a new simple column. This row operation also makes exactly one of the other simple columns into a nonsimple column. (In terms of variables, it is said that a free variable becomes a basic variable and a basic variable becomes a free variable.) Now permuting the columns and variables, yields

$$\begin{pmatrix} I & F' & \mathbf{0} & \mathbf{b}' \\ \mathbf{0} & \mathbf{c}' & 1 & z^{0\prime} \end{pmatrix}$$

where $z^{0\prime} \geq z^0$ because $z^{0\prime} = z^0 - c_i \left(\frac{b_j}{F_{ji}}\right)$ and $c_i < 0$. If $\mathbf{b}' \geq 0$, you are in the same position you were at the beginning but now z^0 is larger. Now here is the **important** thing. You don't pick just any F_{ji} when you do these row operations. You **pick the positive one for which the row operation results in $\mathbf{b}' \geq \mathbf{0}$**. Otherwise the obvious basic feasible solution obtained by letting $\mathbf{x}'_F = \mathbf{0}$ will fail to satisfy the constraint that $\mathbf{x} \geq \mathbf{0}$.

How is this done? You need

$$b'_p \equiv b_p - \frac{F_{pi} b_j}{F_{ji}} \geq 0 \tag{8.12}$$

for each $p = 1, \cdots, m$ or equivalently,

$$b_p \geq \frac{F_{pi} b_j}{F_{ji}}. \tag{8.13}$$

Now if $F_{pi} \leq 0$ the above holds. Therefore, you only need to check F_{pi} for $F_{pi} > 0$. The pivot, F_{ji} is the one which makes the quotients of the form

$$\frac{b_p}{F_{pi}}$$

for all positive F_{pi} the smallest. Having gotten a new simplex tableau, you do the same thing to it which was just done and continue. As long as $\mathbf{b} > \mathbf{0}$, so you don't encounter the degenerate case, the values for z associated with setting $\mathbf{x}_F = \mathbf{0}$ keep getting strictly larger every time the process is repeated. You keep going until you find $\mathbf{c} \geq \mathbf{0}$. Then you stop. You are at a maximum. Problems can occur in the process in the so called degenerate case when at some stage of the process some $b_j = 0$. In this case you can cycle through different values for $\mathbf{x}$ with no improvement in z. This case will not be discussed here.

8.3.2 Minimums

How does it differ if you are finding a minimum? From a basic feasible solution, a simplex tableau of the following form has been obtained in which the simple columns for the basic variables, $\mathbf{x}_B$ are listed first and $\mathbf{b} \geq \mathbf{0}$.

$$\begin{pmatrix} I & F & \mathbf{0} & \mathbf{b} \\ \mathbf{0} & \mathbf{c} & 1 & z^0 \end{pmatrix} \tag{8.14}$$

Let $x_i^0 = b_i$ for $i = 1, \cdots, m$ and $x_i^0 = 0$ for $i > m$. Then $\left(\mathbf{x}^0, z^0\right)$ is a solution to the above system and since $\mathbf{b} \geq \mathbf{0}$, it follows $\left(\mathbf{x}^0, z^0\right)$ is a basic feasible solution. So far, there is no change.

Suppose first that some $c_i > 0$ and $F_{ji} \leq 0$ for each j. Then let $\mathbf{x}_F'$ consist of changing x_i by making it positive but leaving the other entries of $\mathbf{x}_F$ equal to 0. Then from the bottom row,

$$z = -\mathbf{c}\mathbf{x}_F' + z^0 = -c_i x_i + z^0$$

and you let $\mathbf{x}_B' = \mathbf{b} - F\mathbf{x}_F' \geq \mathbf{0}$. Thus the constraints continue to hold when x_i is made increasingly positive and it follows from the above equation that there is no minimum for z. You stop when this happens.

Next suppose $\mathbf{c} \leq \mathbf{0}$. Then in this case, $z = z^0 - \mathbf{c}\mathbf{x}_F$ and from the constraints, $\mathbf{x}_F \geq \mathbf{0}$ and so $-\mathbf{c}\mathbf{x}_F \geq 0$ and so z^0 is the minimum value and you stop since this is what you are looking for.

What do you do in the case where some $c_i > 0$ and some $F_{ji} > 0$? In this case, you use the simplex algorithm as in the case of maximums to obtain a new simplex tableau in which $z^{0\prime}$ is smaller. You choose F_{ji} the same way to be the positive entry of the i^{th} column such that $b_p/F_{pi} \geq b_j/F_{ji}$ for all positive entries, F_{pi} and do the same row operations. Now this time,

$$z^{0\prime} = z^0 - c_i \left(\frac{b_j}{F_{ji}}\right) < z^0$$

As in the case of maximums no problem can occur and the process will converge unless you have the degenerate case in which some $b_j = 0$. As in the earlier case, this is most unfortunate when it occurs. You see what happens of course. z^0 does not change and the algorithm just delivers different values of the variables forever with no improvement.

To summarize the geometrical significance of the simplex algorithm, it takes you from one corner of the feasible region to another. You go in one direction to find the maximum and in another to find the minimum. For the maximum you try to get rid of negative entries of $\mathbf{c}$ and for minimums you try to eliminate positive entries of $\mathbf{c}$ where the method of elimination involves the auspicious use of an appropriate pivot entry and row operations.

Now return to Example 8.2.2. It will be modified to be a maximization problem.

Example 8.3.1 *Maximize $z = x_1 - x_2$ subject to the constraints, $x_1 + 2x_2 \leq 10, x_1 + 2x_2 \geq 2$, and $2x_1 + x_2 \leq 6, x_i \geq 0$.*

Recall this is the same as maximizing $z = x_1 - x_2$ subject to

$$\begin{pmatrix} 1 & 2 & 1 & 0 & 0 \\ 1 & 2 & 0 & -1 & 0 \\ 2 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 10 \\ 2 \\ 6 \end{pmatrix}, \mathbf{x} \geq \mathbf{0},$$

the variables, x_3, x_4, x_5 being slack variables. Recall the simplex tableau was

$$\begin{pmatrix} 1 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 5 \\ 0 & 1 & 0 & 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & \frac{3}{2} & -\frac{1}{2} & 0 & 1 \\ 0 & 0 & 0 & -\frac{3}{2} & -\frac{1}{2} & 1 & -5 \end{pmatrix}$$

with the variables ordered as x_2, x_4, x_5, x_1, x_3 and so $\mathbf{x}_B = (x_2, x_4, x_5)$ and $\mathbf{x}_F = (x_1, x_3)$.

Apply the simplex algorithm to the fourth column because $-\frac{3}{2} < 0$ and this is the most negative entry in the bottom row. The pivot is $3/2$ because $1/(3/2) = 2/3 < 5/(1/2)$. Dividing this row by $3/2$ and then using this to zero out the other elements in that column, the new simplex tableau is

$$\begin{pmatrix} 1 & 0 & -\frac{1}{3} & 0 & \frac{2}{3} & 0 & \frac{14}{3} \\ 0 & 1 & 0 & 0 & 1 & 0 & 8 \\ 0 & 0 & \frac{2}{3} & 1 & -\frac{1}{3} & 0 & \frac{2}{3} \\ 0 & 0 & 1 & 0 & -1 & 1 & -4 \end{pmatrix}.$$

Now there is still a negative number in the bottom left row. Therefore, the process should be continued. This time the pivot is the $2/3$ in the top of the column. Dividing the top row by $2/3$ and then using this to zero out the entries below it,

$$\begin{pmatrix} \frac{3}{2} & 0 & -\frac{1}{2} & 0 & 1 & 0 & 7 \\ -\frac{3}{2} & 1 & \frac{1}{2} & 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} & 1 & 0 & 0 & 3 \\ \frac{3}{2} & 0 & \frac{1}{2} & 0 & 0 & 1 & 3 \end{pmatrix}.$$

Now all the numbers on the bottom left row are nonnegative so the process stops. Now recall the variables and columns were ordered as x_2, x_4, x_5, x_1, x_3. The solution in terms of x_1 and x_2 is $x_2 = 0$ and $x_1 = 3$ and $z = 3$. Note that in the above, I did not worry about permuting the columns to keep those which go with the basic varibles on the left.

Here is a bucolic example.

Example 8.3.2 *Consider the following table.*

	F_1	F_2	F_3	F_4
iron	*1*	*2*	*1*	*3*
protein	*5*	*3*	*2*	*1*
folic acid	*1*	*2*	*2*	*1*
copper	*2*	*1*	*1*	*1*
calcium	*1*	*1*	*1*	*1*

This information is available to a pig farmer and F_i denotes a particular feed. The numbers in the table contain the number of units of a particular nutrient contained in one pound of the given feed. Thus F_2 has 2 units of iron in one pound. Now suppose the cost of each feed in cents per pound is given in the following table.

F_1	F_2	F_3	F_4
2	*3*	*2*	*3*

A typical pig needs 5 units of iron, 8 of protein, 6 of folic acid, 7 of copper and 4 of calcium. (The units may change from nutrient to nutrient.) How many pounds of each feed per pig should the pig farmer use in order to minimize his cost?

His problem is to minimize $C \equiv 2x_1 + 3x_2 + 2x_3 + 3x_4$ subject to the constraints

$$\begin{aligned} x_1 + 2x_2 + x_3 + 3x_4 &\geq 5, \\ 5x_1 + 3x_2 + 2x_3 + x_4 &\geq 8, \\ x_1 + 2x_2 + 2x_3 + x_4 &\geq 6, \\ 2x_1 + x_2 + x_3 + x_4 &\geq 7, \\ x_1 + x_2 + x_3 + x_4 &\geq 4. \end{aligned}$$

where each $x_i \geq 0$. Add in the slack variables,

$$\begin{aligned} x_1 + 2x_2 + x_3 + 3x_4 - x_5 &= 5 \\ 5x_1 + 3x_2 + 2x_3 + x_4 - x_6 &= 8 \\ x_1 + 2x_2 + 2x_3 + x_4 - x_7 &= 6 \\ 2x_1 + x_2 + x_3 + x_4 - x_8 &= 7 \\ x_1 + x_2 + x_3 + x_4 - x_9 &= 4 \end{aligned}$$

The augmented matrix for this system is

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

How in the world can you find a basic feasible solution? Remember the simplex algorithm is designed to keep the entries in the right column nonnegative so you use this algorithm a few times till the obvious solution is a basic feasible solution.

Consider the first column. The pivot is the 5. Using the row operations described in the algorithm, you get

$$\begin{pmatrix} 0 & \frac{7}{5} & \frac{3}{5} & \frac{14}{5} & -1 & \frac{1}{5} & 0 & 0 & 0 & \frac{17}{5} \\ 1 & \frac{3}{5} & \frac{2}{5} & \frac{1}{5} & 0 & -\frac{1}{5} & 0 & 0 & 0 & \frac{8}{5} \\ 0 & \frac{7}{5} & \frac{8}{5} & \frac{4}{5} & 0 & \frac{1}{5} & -1 & 0 & 0 & \frac{22}{5} \\ 0 & -\frac{1}{5} & \frac{1}{5} & \frac{3}{5} & 0 & \frac{2}{5} & 0 & -1 & 0 & \frac{19}{5} \\ 0 & \frac{2}{5} & \frac{3}{5} & \frac{4}{5} & 0 & \frac{1}{5} & 0 & 0 & -1 & \frac{12}{5} \end{pmatrix}$$

Now go to the second column. The pivot in this column is the 7/5. This is in a different row than the pivot in the first column so I will use it to zero out everything below it. This will get rid of the zeros in the fifth column and introduce zeros in the second. This yields

$$\begin{pmatrix} 0 & 1 & \frac{3}{7} & 2 & -\frac{5}{7} & \frac{1}{7} & 0 & 0 & 0 & \frac{17}{7} \\ 1 & 0 & \frac{1}{7} & -1 & \frac{3}{7} & -\frac{2}{7} & 0 & 0 & 0 & \frac{1}{7} \\ 0 & 0 & 1 & -2 & 1 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & \frac{2}{7} & 1 & -\frac{1}{7} & \frac{3}{7} & 0 & -1 & 0 & \frac{30}{7} \\ 0 & 0 & \frac{3}{7} & 0 & \frac{2}{7} & \frac{1}{7} & 0 & 0 & -1 & \frac{10}{7} \end{pmatrix}$$

Now consider another column, this time the fourth. I will pick this one because it has some negative numbers in it so there are fewer entries to check in looking for a pivot. Unfortunately, the pivot is the top 2 and I don't want to pivot on this because it would destroy the zeros in the second column. Consider the fifth column. It is also not a good choice because the pivot is the second entry from the top and this would destroy the zeros

in the first column. Consider the sixth column. I can use either of the two bottom entries as the pivot. The matrix is

$$\begin{pmatrix} 0 & 1 & 0 & 2 & -1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & -1 & 1 & 0 & 0 & 0 & -2 & 3 \\ 0 & 0 & 1 & -2 & 1 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 & -1 & 0 & 0 & -1 & 3 & 0 \\ 0 & 0 & 3 & 0 & 2 & 1 & 0 & 0 & -7 & 10 \end{pmatrix}$$

Next consider the third column. The pivot is the 1 in the third row. This yields

$$\begin{pmatrix} 0 & 1 & 0 & 2 & -1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & -2 & 2 \\ 0 & 0 & 1 & -2 & 1 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & -1 & -1 & 3 & 1 \\ 0 & 0 & 0 & 6 & -1 & 1 & 3 & 0 & -7 & 7 \end{pmatrix}.$$

There are still 5 columns which consist entirely of zeros except for one entry. Four of them have that entry equal to 1 but one still has a -1 in it, the -1 being in the fourth column. I need to do the row operations on a nonsimple column which has the pivot in the fourth row. Such a column is the second to the last. The pivot is the 3. The new matrix is

$$\begin{pmatrix} 0 & 1 & 0 & \frac{7}{3} & -1 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{2}{3} \\ 1 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & -\frac{2}{3} & 0 & \frac{8}{3} \\ 0 & 0 & 1 & -2 & 1 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{1}{3} & 0 & 0 & -\frac{1}{3} & -\frac{1}{3} & 1 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{11}{3} & -1 & 1 & \frac{2}{3} & -\frac{7}{3} & 0 & \frac{28}{3} \end{pmatrix}. \tag{8.15}$$

Now the obvious basic solution is feasible. You let $x_4 = 0 = x_5 = x_7 = x_8$ and $x_1 = 8/3, x_2 = 2/3, x_3 = 1$, and $x_6 = 28/3$. You don't need to worry too much about this. It is the above matrix which is desired. Now you can assemble the simplex tableau and begin the algorithm. Remember $C \equiv 2x_1 + 3x_2 + 2x_3 + 3x_4$. First add the row and column which deal with C. This yields

$$\begin{pmatrix} 0 & 1 & 0 & \frac{7}{3} & -1 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ 1 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & -\frac{2}{3} & 0 & 0 & \frac{8}{3} \\ 0 & 0 & 1 & -2 & 1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{1}{3} & 0 & 0 & -\frac{1}{3} & -\frac{1}{3} & 1 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{11}{3} & -1 & 1 & \frac{2}{3} & -\frac{7}{3} & 0 & 0 & \frac{28}{3} \\ -2 & -3 & -2 & -3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{8.16}$$

Now you do row operations to keep the simple columns of 8.15 simple in 8.16. Of course you could permute the columns if you wanted but this is not necessary.

This yields the following for a simplex tableau. Now it is a matter of getting rid of the positive entries in the bottom row because you are trying to minimize.

$$\begin{pmatrix} 0 & 1 & 0 & \frac{7}{3} & -1 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ 1 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & -\frac{2}{3} & 0 & 0 & \frac{8}{3} \\ 0 & 0 & 1 & -2 & 1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{1}{3} & 0 & 0 & -\frac{1}{3} & -\frac{1}{3} & 1 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{11}{3} & -1 & 1 & \frac{2}{3} & -\frac{7}{3} & 0 & 0 & \frac{28}{3} \\ 0 & 0 & 0 & \frac{2}{3} & -1 & 0 & -\frac{1}{3} & -\frac{1}{3} & 0 & 1 & \frac{28}{3} \end{pmatrix}$$

The most positive of them is the 2/3 and so I will apply the algorithm to this one first. The pivot is the 7/3. After doing the row operation the next tableau is

$$\begin{pmatrix} 0 & \frac{3}{7} & 0 & 1 & -\frac{3}{7} & 0 & \frac{1}{7} & \frac{1}{7} & 0 & 0 & \frac{2}{7} \\ 1 & -\frac{1}{7} & 0 & 0 & \frac{1}{7} & 0 & \frac{2}{7} & -\frac{5}{7} & 0 & 0 & \frac{18}{7} \\ 0 & \frac{6}{7} & 1 & 0 & \frac{1}{7} & 0 & -\frac{5}{7} & \frac{2}{7} & 0 & 0 & \frac{11}{7} \\ 0 & \frac{1}{7} & 0 & 0 & -\frac{1}{7} & 0 & -\frac{2}{7} & -\frac{2}{7} & 1 & 0 & \frac{3}{7} \\ 0 & -\frac{11}{7} & 0 & 0 & \frac{4}{7} & 1 & \frac{1}{7} & -\frac{20}{7} & 0 & 0 & \frac{58}{7} \\ 0 & -\frac{2}{7} & 0 & 0 & -\frac{5}{7} & 0 & -\frac{3}{7} & -\frac{3}{7} & 0 & 1 & \frac{64}{7} \end{pmatrix}$$

and you see that all the entries are negative and so the minimum is 64/7 and it occurs when $x_1 = 18/7, x_2 = 0, x_3 = 11/7, x_4 = 2/7$.

There is no maximum for the above problem. However, I will pretend I don't know this and attempt to use the simplex algorithm. You set up the simiplex tableau the same way. Recall it is

$$\begin{pmatrix} 0 & 1 & 0 & \frac{7}{3} & -1 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ 1 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & -\frac{2}{3} & 0 & 0 & \frac{8}{3} \\ 0 & 0 & 1 & -2 & 1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{1}{3} & 0 & 0 & -\frac{1}{3} & -\frac{1}{3} & 1 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{11}{3} & -1 & 1 & \frac{2}{3} & -\frac{7}{3} & 0 & 0 & \frac{28}{3} \\ 0 & 0 & 0 & \frac{2}{3} & -1 & 0 & -\frac{1}{3} & -\frac{1}{3} & 0 & 1 & \frac{28}{3} \end{pmatrix}$$

Now to maximize, you try to get rid of the negative entries in the bottom left row. The most negative entry is the -1 in the fifth column. The pivot is the 1 in the third row of this column. The new tableau is

$$\begin{pmatrix} 0 & 1 & 1 & \frac{1}{3} & 0 & 0 & -\frac{2}{3} & \frac{1}{3} & 0 & 0 & \frac{5}{3} \\ 1 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & -\frac{2}{3} & 0 & 0 & \frac{8}{3} \\ 0 & 0 & 1 & -2 & 1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{1}{3} & 0 & 0 & -\frac{1}{3} & -\frac{1}{3} & 1 & 0 & \frac{1}{3} \\ 0 & 0 & 1 & \frac{5}{3} & 0 & 1 & -\frac{1}{3} & -\frac{7}{3} & 0 & 0 & \frac{31}{3} \\ 0 & 0 & 1 & -\frac{4}{3} & 0 & 0 & -\frac{4}{3} & -\frac{1}{3} & 0 & 1 & \frac{31}{3} \end{pmatrix}.$$

Consider the fourth column. The pivot is the top 1/3. The new tableau is

$$\begin{pmatrix} 0 & 3 & 3 & 1 & 0 & 0 & -2 & 1 & 0 & 0 & 5 \\ 1 & -1 & -1 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 1 \\ 0 & 6 & 7 & 0 & 1 & 0 & -5 & 2 & 0 & 0 & 11 \\ 0 & 1 & 1 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 2 \\ 0 & -5 & -4 & 0 & 0 & 1 & 3 & -4 & 0 & 0 & 2 \\ 0 & 4 & 5 & 0 & 0 & 0 & -4 & 1 & 0 & 1 & 17 \end{pmatrix}$$

There is still a negative in the bottom, the -4. The pivot in that column is the 3. The algorithm yields

$$\begin{pmatrix} 0 & -\frac{1}{3} & \frac{1}{3} & 1 & 0 & \frac{2}{3} & 0 & -\frac{5}{3} & 0 & 0 & \frac{19}{3} \\ 1 & \frac{2}{3} & \frac{1}{3} & 0 & 0 & -\frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & -\frac{7}{3} & \frac{1}{3} & 0 & 1 & \frac{5}{3} & 0 & -\frac{14}{3} & 0 & 0 & \frac{43}{3} \\ 0 & -\frac{2}{3} & -\frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & -\frac{4}{3} & 1 & 0 & \frac{8}{3} \\ 0 & -\frac{5}{3} & -\frac{4}{3} & 0 & 0 & \frac{1}{3} & 1 & -\frac{4}{3} & 0 & 0 & \frac{2}{3} \\ 0 & -\frac{8}{3} & -\frac{1}{3} & 0 & 0 & \frac{4}{3} & 0 & -\frac{13}{3} & 0 & 1 & \frac{59}{3} \end{pmatrix}$$

Note how z keeps getting larger. Consider the column having the $-13/3$ in it. The pivot is

the single positive entry, 1/3. The next tableau is

$$\begin{pmatrix} 5 & 3 & 2 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 8 \\ 3 & 2 & 1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 1 \\ 14 & 7 & 5 & 0 & 1 & -3 & 0 & 0 & 0 & 0 & 19 \\ 4 & 2 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 4 \\ 4 & 1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 2 \\ 13 & 6 & 4 & 0 & 0 & -3 & 0 & 0 & 0 & 1 & 24 \end{pmatrix}.$$

There is a column consisting of all negative entries. There is therefore, no maximum. Note also how there is no way to pick the pivot in that column.

Example 8.3.3 *Minimize* $z = x_1 - 3x_2 + x_3$ *subject to the constraints* $x_1 + x_2 + x_3 \leq 10, x_1 + x_2 + x_3 \geq 2$, $x_1 + x_2 + 3x_3 \leq 8$ *and* $x_1 + 2x_2 + x_3 \leq 7$ *with all variables nonnegative.*

There exists an answer because the region defined by the contraints is closed and bounded. Adding in slack variables you get the following augmented matrix corresponding to the constraints.

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 10 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 2 \\ 1 & 1 & 3 & 0 & 0 & 1 & 0 & 8 \\ 1 & 2 & 1 & 0 & 0 & 0 & 1 & 7 \end{pmatrix}$$

Of course there is a problem with the obvious solution obtained by setting to zero all variables corresponding to a nonsimple column because of the simple column which has the -1 in it. Therefore, I will use the simplex algorithm to make this column non simple. The third column has the 1 in the second row as the pivot so I will use this column. This yields

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 8 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 2 \\ -2 & -2 & 0 & 0 & 3 & 1 & 0 & 2 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 5 \end{pmatrix} \tag{8.17}$$

and the obvious solution is feasible. Now it is time to assemble the simplex tableau. First add in the bottom row and second to last column corresponding to the the equation for z. This yields

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 8 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & 2 \\ -2 & -2 & 0 & 0 & 3 & 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 5 \\ -1 & 3 & -1 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Next you need to zero out the entries in the bottom row which are below one of the simple columns in 8.17. This yields the simplex tableau

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 8 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & 2 \\ -2 & -2 & 0 & 0 & 3 & 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 5 \\ 0 & 4 & 0 & 0 & -1 & 0 & 0 & 1 & 2 \end{pmatrix}.$$

The desire is to minimize this so you need to get rid of the positive entries in the left bottom row. There is only one such entry, the 4. In that column the pivot is the 1 in the second

row of this column. Thus the next tableau is

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 8 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 2 & 0 & 1 & 1 & 0 & 0 & 6 \\ -1 & 0 & -1 & 0 & 2 & 0 & 1 & 0 & 3 \\ -4 & 0 & -4 & 0 & 3 & 0 & 0 & 1 & -6 \end{pmatrix}$$

There is still a positive number there, the 3. The pivot in this column is the 2. Apply the algorithm again. This yields

$$\begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 1 & 0 & 0 & -\frac{1}{2} & 0 & \frac{13}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{7}{2} \\ \frac{1}{2} & 0 & \frac{5}{2} & 0 & 0 & 1 & -\frac{1}{2} & 0 & \frac{3}{2} \\ -\frac{1}{2} & 0 & -\frac{1}{2} & 0 & 1 & 0 & \frac{1}{2} & 0 & \frac{3}{2} \\ -\frac{5}{2} & 0 & -\frac{5}{2} & 0 & 0 & 0 & -\frac{3}{2} & 1 & -\frac{21}{2} \end{pmatrix}.$$

Now all the entries in the left bottom row are nonpositive so the process has stopped. The minimum is $-21/2$. It occurs when $x_1 = 0$, $x_2 = 7/2, x_3 = 0$.

Now consider the same problem but change the word, minimize to the word, maximize.

Example 8.3.4 *Maximize* $z = x_1 - 3x_2 + x_3$ *subject to the constraints* $x_1 + x_2 + x_3 \leq 10, x_1 + x_2 + x_3 \geq 2$, $x_1 + x_2 + 3x_3 \leq 8$ *and* $x_1 + 2x_2 + x_3 \leq 7$ *with all variables nonnegative.*

The first part of it is the same. You wind up with the same simplex tableau,

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 8 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & 2 \\ -2 & -2 & 0 & 0 & 3 & 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 5 \\ 0 & 4 & 0 & 0 & -1 & 0 & 0 & 1 & 2 \end{pmatrix}$$

but this time, you apply the algorithm to get rid of the negative entries in the left bottom row. There is a -1. Use this column. The pivot is the 3. The next tableau is

$$\begin{pmatrix} \frac{2}{3} & \frac{2}{3} & 0 & 1 & 0 & -\frac{1}{3} & 0 & 0 & \frac{22}{3} \\ \frac{1}{3} & \frac{1}{3} & 1 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{8}{3} \\ -\frac{2}{3} & -\frac{2}{3} & 0 & 0 & 1 & \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ \frac{2}{3} & \frac{5}{3} & 0 & 0 & 0 & -\frac{1}{3} & 1 & 0 & \frac{13}{3} \\ -\frac{2}{3} & \frac{10}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 1 & \frac{8}{3} \end{pmatrix}$$

There is still a negative entry, the $-2/3$. This will be the new pivot column. The pivot is the $2/3$ on the fourth row. This yields

$$\begin{pmatrix} 0 & -1 & 0 & 1 & 0 & 0 & -1 & 0 & 3 \\ 0 & -\frac{1}{2} & 1 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 5 \\ 1 & \frac{5}{2} & 0 & 0 & 0 & -\frac{1}{2} & \frac{3}{2} & 0 & \frac{13}{2} \\ 0 & 5 & 0 & 0 & 0 & 0 & 1 & 1 & 7 \end{pmatrix}$$

and the process stops. The maximum for z is 7 and it occurs when $x_1 = 13/2, x_2 = 0, x_3 = 1/2$.

8.4 Finding A Basic Feasible Solution

By now it should be fairly clear that finding a basic feasible solution can create considerable difficulty. Indeed, given a system of linear inequalities along with the requirement that each variable be nonnegative, do there even exist points satisfying all these inequalities? If you have many variables, you can't answer this by drawing a picture. Is there some other way to do this which is more systematic than what was presented above? The answer is yes. It is called the method of artificial variables. I will illustrate this method with an example.

Example 8.4.1 *Find a basic feasible solution to the system* $2x_1+x_2-x_3 \geq 3, x_1+x_2+x_3 \geq 2, x_1+x_2+x_3 \leq 7$ *and* $\mathbf{x} \geq \mathbf{0}$.

If you write the appropriate augmented matrix with the slack variables,

$$\begin{pmatrix} 2 & 1 & -1 & -1 & 0 & 0 & 3 \\ 1 & 1 & 1 & 0 & -1 & 0 & 2 \\ 1 & 1 & 1 & 0 & 0 & 1 & 7 \end{pmatrix} \tag{8.18}$$

The obvious solution is not feasible. This is why it would be hard to get started with the simplex method. What is the problem? It is those -1 entries in the fourth and fifth columns. To get around this, you add in artificial variables to get an augmented matrix of the form

$$\begin{pmatrix} 2 & 1 & -1 & -1 & 0 & 0 & 1 & 0 & 3 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 1 & 2 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 7 \end{pmatrix} \tag{8.19}$$

Thus the variables are $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$. Suppose you can find a feasible solution to the system of equations represented by the above augmented matrix. Thus all variables are nonnegative. Suppose also that it can be done in such a way that x_8 and x_7 happen to be 0. Then it will follow that $x_1, \cdots, x_6$ is a feasible solution for 8.18. Conversely, if you can find a feasible solution for 8.18, then letting x_7 and x_8 both equal zero, you have obtained a feasible solution to 8.19. Since all variables are nonnegative, x_7 and x_8 both equalling zero is equivalent to saying the minimum of $z = x_7 + x_8$ subject to the constraints represented by the above augmented matrix equals zero. This has proved the following simple observation.

Observation 8.4.2 *There exists a feasible solution to the constraints represented by the augmented matrix of 8.18 and* $\mathbf{x} \geq \mathbf{0}$ *if and only if the minimum of* $x_7 + x_8$ *subject to the constraints of 8.19 and* $\mathbf{x} \geq \mathbf{0}$ *exists and equals* 0.

Of course a similar observation would hold in other similar situations. Now the point of all this is that it is trivial to see a feasible solution to 8.19, namely $x_6 = 7, x_7 = 3, x_8 = 2$ and all the other variables may be set to equal zero. Therefore, it is easy to find an initial simplex tableau for the minimization problem just described. First add the column and row for z

$$\begin{pmatrix} 2 & 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 3 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 1 & 0 & 2 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 7 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 0 \end{pmatrix}$$

Next it is necessary to make the last two columns on the bottom left row into simple columns. Performing the row operation, this yields an initial simplex tableau,

$$\begin{pmatrix} 2 & 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 3 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 1 & 0 & 2 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 7 \\ 3 & 2 & 0 & -1 & -1 & 0 & 0 & 0 & 1 & 5 \end{pmatrix}$$

Now the algorithm involves getting rid of the positive entries on the left bottom row. Begin with the first column. The pivot is the 2. An application of the simplex algorithm yields the new tableau

$$\begin{pmatrix} 1 & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{3}{2} \\ 0 & \frac{1}{2} & \frac{3}{2} & \frac{1}{2} & -1 & 0 & -\frac{1}{2} & 1 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{3}{2} & \frac{1}{2} & 0 & 1 & -\frac{1}{2} & 0 & 0 & \frac{11}{2} \\ 0 & \frac{1}{2} & \frac{3}{2} & \frac{1}{2} & -1 & 0 & -\frac{3}{2} & 0 & 1 & \frac{1}{2} \end{pmatrix}$$

Now go to the third column. The pivot is the 3/2 in the second row. An application of the simplex algorithm yields

$$\begin{pmatrix} 1 & \frac{2}{3} & 0 & -\frac{1}{3} & -\frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{5}{3} \\ 0 & \frac{1}{3} & 1 & \frac{1}{3} & -\frac{2}{3} & 0 & -\frac{1}{3} & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & -1 & 0 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 0 \end{pmatrix} \tag{8.20}$$

and you see there are only nonpositive numbers on the bottom left column so the process stops and yields 0 for the minimum of $z = x_7 + x_8$. As for the other variables, $x_1 = 5/3, x_2 = 0, x_3 = 1/3, x_4 = 0, x_5 = 0, x_6 = 5$. Now as explained in the above observation, this is a basic feasible solution for the original system 8.18.

Now consider a maximization problem associated with the above constraints.

Example 8.4.3 *Maximize* $x_1 - x_2 + 2x_3$ *subject to the constraints,* $2x_1 + x_2 - x_3 \geq 3, x_1 + x_2 + x_3 \geq 2, x_1 + x_2 + x_3 \leq 7$ *and* $\mathbf{x} \geq \mathbf{0}$.

From 8.20 you can immediately assemble an initial simplex tableau. You begin with the first 6 columns and top 3 rows in 8.20. Then add in the column and row for z. This yields

$$\begin{pmatrix} 1 & \frac{2}{3} & 0 & -\frac{1}{3} & -\frac{1}{3} & 0 & 0 & \frac{5}{3} \\ 0 & \frac{1}{3} & 1 & \frac{1}{3} & -\frac{2}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 5 \\ -1 & 1 & -2 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

and you first do row operations to make the first and third columns simple columns. Thus the next simplex tableau is

$$\begin{pmatrix} 1 & \frac{2}{3} & 0 & -\frac{1}{3} & -\frac{1}{3} & 0 & 0 & \frac{5}{3} \\ 0 & \frac{1}{3} & 1 & \frac{1}{3} & -\frac{2}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 5 \\ 0 & \frac{7}{3} & 0 & \frac{1}{3} & -\frac{5}{3} & 0 & 1 & \frac{7}{3} \end{pmatrix}$$

You are trying to get rid of negative entries in the bottom left row. There is only one, the $-5/3$. The pivot is the 1. The next simplex tableau is then

$$\begin{pmatrix} 1 & \frac{2}{3} & 0 & -\frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{10}{3} \\ 0 & \frac{1}{3} & 1 & \frac{1}{3} & 0 & \frac{2}{3} & 0 & \frac{11}{3} \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 5 \\ 0 & \frac{7}{3} & 0 & \frac{1}{3} & 0 & \frac{5}{3} & 1 & \frac{32}{3} \end{pmatrix}$$

and so the maximum value of z is 32/3 and it occurs when $x_1 = 10/3, x_2 = 0$ and $x_3 = 11/3$.

8.5 Duality

You can solve minimization problems by solving maximization problems. You can also go the other direction and solve maximization problems by minimization problems. Sometimes this makes things much easier. To be more specific, the two problems to be considered are

A.) Minimize $z = \mathbf{c}\mathbf{x}$ subject to $\mathbf{x} \geq \mathbf{0}$ and $A\mathbf{x} \geq \mathbf{b}$ and
B.) Maximize $w = \mathbf{y}\mathbf{b}$ such that $\mathbf{y} \geq \mathbf{0}$ and $\mathbf{y}A \leq \mathbf{c}$,

$$\left(\text{equivalently } A^T\mathbf{y}^T \geq \mathbf{c}^T \text{ and } w = \mathbf{b}^T\mathbf{y}^T\right).$$

In these problems it is assumed A is an $m \times p$ matrix.

I will show how a solution of the first yields a solution of the second and then show how a solution of the second yields a solution of the first. The problems, A.) and B.) are called dual problems.

Lemma 8.5.1 *Let* $\mathbf{x}$ *be a solution of the inequalities of* A.) *and let* $\mathbf{y}$ *be a solution of the inequalities of* B.). *Then*

$$\mathbf{c}\mathbf{x} \geq \mathbf{y}\mathbf{b}.$$

and if equality holds in the above, then $\mathbf{x}$ *is the solution to* A.) *and* $\mathbf{y}$ *is a solution to* B.).

Proof: This follows immediately. Since $\mathbf{c} \geq \mathbf{y}A$,

$$\mathbf{c}\mathbf{x} \geq \mathbf{y}A\mathbf{x} \geq \mathbf{y}\mathbf{b}.$$

It follows from this lemma that if $\mathbf{y}$ satisfies the inequalitites of B.) and $\mathbf{x}$ satisfies the inequalities of A.) then if equality holds in the above lemma, it must be that $\mathbf{x}$ is a solution of A.) and $\mathbf{y}$ is a solution of B.). This proves the lemma.

Now recall that to solve either of these problems using the simplex method, you first add in slack variables. Denote by $\mathbf{x}'$ and $\mathbf{y}'$ the enlarged list of variables. Thus $\mathbf{x}'$ has at least m entries and so does $\mathbf{y}'$ and the inequalities involving A were replaced by equalities whose augmented matrices were of the form

$$\begin{pmatrix} A & -I & \mathbf{b} \end{pmatrix}, \text{ and } \begin{pmatrix} A^T & I & \mathbf{c}^T \end{pmatrix}$$

Then you included the row and column for z and w to obtain

$$\begin{pmatrix} A & -I & \mathbf{0} & \mathbf{b} \\ -\mathbf{c} & \mathbf{0} & 1 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} A^T & I & \mathbf{0} & \mathbf{c}^T \\ -\mathbf{b}^T & \mathbf{0} & 1 & 0 \end{pmatrix}. \tag{8.21}$$

Then the problems have basic feasible solutions if it is possible to permute the first $p+m$ columns in the above two matrices and obtain matrices of the form

$$\begin{pmatrix} B & F & \mathbf{0} & \mathbf{b} \\ -\mathbf{c}_B & -\mathbf{c}_F & 1 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} B_1 & F_1 & \mathbf{0} & \mathbf{c}^T \\ -\mathbf{b}_{B_1}^T & -\mathbf{b}_{F_1}^T & 1 & 0 \end{pmatrix} \tag{8.22}$$

where B, B_1 are invertible $m \times m$ and $p \times p$ matrices and denoting the variables associated with these columns by $\mathbf{x}_B, \mathbf{y}_B$ and those variables associated with F or F_1 by $\mathbf{x}_F$ and $\mathbf{y}_F$, it follows that letting $B\mathbf{x}_B = \mathbf{b}$ and $\mathbf{x}_F = \mathbf{0}$, the resulting vector, $\mathbf{x}'$ is a solution to $\mathbf{x}' \geq \mathbf{0}$ and $\begin{pmatrix} A & -I \end{pmatrix}\mathbf{x}' = \mathbf{b}$ with similar constraints holding for $\mathbf{y}'$. In other words, it is possible to obtain simplex tableaus,

$$\begin{pmatrix} I & B^{-1}F & \mathbf{0} & B^{-1}\mathbf{b} \\ \mathbf{0} & \mathbf{c}_B B^{-1}F - \mathbf{c}_F & 1 & \mathbf{c}_B B^{-1}\mathbf{b} \end{pmatrix}, \begin{pmatrix} I & B_1^{-1}F_1 & \mathbf{0} & B_1^{-1}\mathbf{c}^T \\ \mathbf{0} & \mathbf{b}_{B_1}^T B_1^{-1}F - \mathbf{b}_{F_1}^T & 1 & \mathbf{b}_{B_1}^T B_1^{-1}\mathbf{c}^T \end{pmatrix} \tag{8.23}$$

Similar considerations apply to the second problem. Thus as just described, a basic feasible solution is one which determines a simplex tableau like the above in which you get a feasible solution by setting all but the first m variables equal to zero. The simplex algorithm takes you from one basic feasible solution to another till eventually, if there is no degeneracy, you obtain a basic feasible solution which yields the solution of the problem of interest.

Theorem 8.5.2 *Suppose there exists a solution,* $\mathbf{x}$ *to* $A.$*) where* $\mathbf{x}$ *is a basic feasible solution of the inequalities of* $\mathbf{A}.$*). Then there exists a solution,* $\mathbf{y}$ *to* $B.$*) and* $\mathbf{cx} = \mathbf{by}$*. It is also possible to find* $\mathbf{y}$ *from* $\mathbf{x}$ *using a simple formula.*

Proof: Since the solution to $A.$) is basic and feasible, there exists a simplex tableau like 8.23 such that $\mathbf{x}'$ can be split into $\mathbf{x}_B$ and $\mathbf{x}_F$ such that $\mathbf{x}_F = 0$ and $\mathbf{x}_B = B^{-1}\mathbf{b}$. Now since it is a minimizer, it follows $\mathbf{c}_B B^{-1}F - \mathbf{c}_F \leq \mathbf{0}$ and the minimum value for $\mathbf{cx}$ is $\mathbf{c}_B B^{-1}\mathbf{b}$. Stating this again, $\mathbf{cx} = \mathbf{c}_B B^{-1}\mathbf{b}$. Is it posible you can take $\mathbf{y} = \mathbf{c}_B B^{-1}$? From Lemma 8.5.1 this will be so if $\mathbf{c}_B B^{-1}$ solves the constraints of problem $B.$). Is $\mathbf{c}_B B^{-1} \geq 0$? Is $\mathbf{c}_B B^{-1} A \leq \mathbf{c}$? These two conditions are satisfied if and only if $\mathbf{c}_B B^{-1} \begin{pmatrix} A & -I \end{pmatrix} \leq \begin{pmatrix} \mathbf{c} & \mathbf{0} \end{pmatrix}$. Referring to the process of permuting the columns of the first augmented matrix of 8.21 to get 8.22 and doing the same permutations on the columns of $\begin{pmatrix} A & -I \end{pmatrix}$ and $\begin{pmatrix} \mathbf{c} & \mathbf{0} \end{pmatrix}$, the desired inequality holds if and only if $\mathbf{c}_B B^{-1} \begin{pmatrix} B & F \end{pmatrix} \leq \begin{pmatrix} \mathbf{c}_B & \mathbf{c}_F \end{pmatrix}$ which is equivalent to saying $\begin{pmatrix} \mathbf{c}_B & \mathbf{c}_B B^{-1}F \end{pmatrix} \leq \begin{pmatrix} \mathbf{c}_B & \mathbf{c}_F \end{pmatrix}$ and this is true because $\mathbf{c}_B B^{-1}F - \mathbf{c}_F \leq \mathbf{0}$ due to the assumption that $\mathbf{x}$ is a minimizer. The simple formula is just

$$\mathbf{y} = \mathbf{c}_B B^{-1}.$$

This proves the theorem.

The proof of the following corollary is similar.

Corollary 8.5.3 *Suppose there exists a solution,* $\mathbf{y}$ *to* $B.$*) where* $\mathbf{y}$ *is a basic feasible solution of the inequalities of* $B.$*). Then there exists a solution,* $\mathbf{x}$ *to* $A.$*) and* $\mathbf{cx} = \mathbf{by}$*. It is also possible to find* $\mathbf{x}$ *from* $\mathbf{y}$ *using a simple formula. In this case, and referring to 8.23, the simple formula is* $\mathbf{x} = B_1^{-T}\mathbf{b}_{B_1}$.

As an example, consider the pig farmers problem. The main difficulty in this problem was finding an initial simplex tableau. Now consider the following example and marvel at how all the difficulties disappear.

Example 8.5.4 *minimize* $C \equiv 2x_1 + 3x_2 + 2x_3 + 3x_4$ *subject to the constraints*

$$\begin{aligned} x_1 + 2x_2 + x_3 + 3x_4 &\geq 5, \\ 5x_1 + 3x_2 + 2x_3 + x_4 &\geq 8, \\ x_1 + 2x_2 + 2x_3 + x_4 &\geq 6, \\ 2x_1 + x_2 + x_3 + x_4 &\geq 7, \\ x_1 + x_2 + x_3 + x_4 &\geq 4. \end{aligned}$$

where each $x_i \geq 0$.

Here the dual problem is to maximize $w = 5y_1 + 8y_2 + 6y_3 + 7y_4 + 4y_5$ subject to the constraints

$$\begin{pmatrix} 1 & 5 & 1 & 2 & 1 \\ 2 & 3 & 2 & 1 & 1 \\ 1 & 2 & 2 & 1 & 1 \\ 3 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} \leq \begin{pmatrix} 2 \\ 3 \\ 2 \\ 3 \end{pmatrix}.$$

Adding in slack variables, these inequalities are equivalent to the system of equations whose augmented matrix is

$$\begin{pmatrix} 1 & 5 & 1 & 2 & 1 & 1 & 0 & 0 & 0 & 2 \\ 2 & 3 & 2 & 1 & 1 & 0 & 1 & 0 & 0 & 3 \\ 1 & 2 & 2 & 1 & 1 & 0 & 0 & 1 & 0 & 2 \\ 3 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 3 \end{pmatrix}$$

Now the obvious solution is feasible so there is no hunting for an initial obvious feasible solution required. Now add in the row and column for w. This yields

$$\begin{pmatrix} 1 & 5 & 1 & 2 & 1 & 1 & 0 & 0 & 0 & 0 & 2 \\ 2 & 3 & 2 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 3 \\ 1 & 2 & 2 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 2 \\ 3 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 3 \\ -5 & -8 & -6 & -7 & -4 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

It is a maximization problem so you want to eliminate the negatives in the bottom left row. Pick the column having the one which is most negative, the -8. The pivot is the top 5. Then apply the simplex algorithm to obtain

$$\begin{pmatrix} \frac{1}{5} & 1 & \frac{1}{5} & \frac{2}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 & \frac{2}{5} \\ \frac{7}{5} & 0 & \frac{7}{5} & -\frac{1}{5} & \frac{2}{5} & -\frac{3}{5} & 1 & 0 & 0 & 0 & \frac{9}{5} \\ \frac{3}{5} & 0 & \frac{8}{5} & \frac{1}{5} & \frac{3}{5} & -\frac{2}{5} & 0 & 1 & 0 & 0 & \frac{8}{5} \\ \frac{14}{5} & 0 & \frac{4}{5} & \frac{3}{5} & \frac{4}{5} & -\frac{1}{5} & 0 & 0 & 1 & 0 & \frac{13}{5} \\ -\frac{17}{5} & 0 & -\frac{22}{5} & -\frac{19}{5} & -\frac{12}{5} & \frac{8}{5} & 0 & 0 & 0 & 1 & \frac{16}{5} \end{pmatrix}.$$

There are still negative entries in the bottom left row. Do the simplex algorithm to the column which has the $-\frac{22}{5}$. The pivot is the $\frac{8}{5}$. This yields

$$\begin{pmatrix} \frac{1}{8} & 1 & 0 & \frac{3}{8} & \frac{1}{8} & \frac{1}{4} & 0 & -\frac{1}{8} & 0 & 0 & \frac{1}{4} \\ \frac{7}{8} & 0 & 0 & -\frac{3}{8} & -\frac{1}{8} & -\frac{1}{4} & 1 & -\frac{7}{8} & 0 & 0 & \frac{3}{4} \\ \frac{3}{8} & 0 & 1 & \frac{1}{8} & \frac{3}{8} & -\frac{1}{4} & 0 & \frac{5}{8} & 0 & 0 & \frac{3}{4} \\ \frac{5}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & -\frac{1}{2} & 1 & 0 & 2 \\ -\frac{7}{4} & 0 & 0 & -\frac{13}{4} & -\frac{3}{4} & \frac{1}{2} & 0 & \frac{11}{4} & 0 & 1 & \frac{13}{2} \end{pmatrix}$$

and there are still negative numbers. Pick the column which has the $-13/4$. The pivot is the $3/8$ in the top. This yields

$$\begin{pmatrix} \frac{1}{3} & \frac{8}{3} & 0 & 1 & \frac{1}{3} & \frac{2}{3} & 0 & -\frac{1}{3} & 0 & 0 & \frac{2}{3} \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 1 \\ \frac{1}{3} & -\frac{1}{3} & 1 & 0 & \frac{1}{3} & -\frac{1}{3} & 0 & \frac{2}{3} & 0 & 0 & \frac{2}{3} \\ \frac{7}{3} & -\frac{4}{3} & 0 & 0 & \frac{1}{3} & -\frac{1}{3} & 0 & -\frac{1}{3} & 1 & 0 & \frac{5}{3} \\ -\frac{2}{3} & \frac{26}{3} & 0 & 0 & \frac{1}{3} & \frac{8}{3} & 0 & \frac{5}{3} & 0 & 1 & \frac{26}{3} \end{pmatrix}$$

which has only one negative entry on the bottom left. The pivot for this first column is the $\frac{7}{3}$. The next tableau is

$$\begin{pmatrix} 0 & \frac{20}{7} & 0 & 1 & \frac{2}{7} & \frac{5}{7} & 0 & -\frac{2}{7} & -\frac{1}{7} & 0 & \frac{3}{7} \\ 0 & \frac{11}{7} & 0 & 0 & -\frac{1}{7} & \frac{1}{7} & 1 & -\frac{6}{7} & -\frac{3}{7} & 0 & \frac{2}{7} \\ 0 & -\frac{1}{7} & 1 & 0 & \frac{2}{7} & -\frac{2}{7} & 0 & \frac{5}{7} & -\frac{1}{7} & 0 & \frac{3}{7} \\ 1 & -\frac{4}{7} & 0 & 0 & \frac{1}{7} & -\frac{1}{7} & 0 & -\frac{1}{7} & \frac{3}{7} & 0 & \frac{5}{7} \\ 0 & \frac{58}{7} & 0 & 0 & \frac{3}{7} & \frac{18}{7} & 0 & \frac{11}{7} & \frac{2}{7} & 1 & \frac{64}{7} \end{pmatrix}$$

and all the entries in the left bottom row are nonnegative so the answer is $64/7$. This is the same as obtained before. So what values for $\mathbf{x}$ are needed? Here the basic variables are y_1, y_3, y_4, y_7. Consider the original augmented matrix, one step before the simplex tableau.

$$\begin{pmatrix} 1 & 5 & 1 & 2 & 1 & 1 & 0 & 0 & 0 & 0 & 2 \\ 2 & 3 & 2 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 3 \\ 1 & 2 & 2 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 2 \\ 3 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 3 \\ -5 & -8 & -6 & -7 & -4 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Permute the columns to put the columns associated with these basic variables first. Thus

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

The matrix, B is

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

and so B^{-T} equals

$$\begin{pmatrix} -\frac{1}{7} & -\frac{2}{7} & \frac{5}{7} & \frac{1}{7} \\ 0 & 0 & 0 & 1 \\ -\frac{1}{7} & \frac{5}{7} & -\frac{2}{7} & -\frac{6}{7} \\ \frac{3}{7} & -\frac{1}{7} & -\frac{1}{7} & -\frac{3}{7} \end{pmatrix}$$

Also $\mathbf{b}_B^T = \begin{pmatrix} 5 & 6 & 7 & 0 \end{pmatrix}$ and so from Corollary 8.5.3,

$$\mathbf{x} = \begin{pmatrix} -\frac{1}{7} & -\frac{2}{7} & \frac{5}{7} & \frac{1}{7} \\ 0 & 0 & 0 & 1 \\ -\frac{1}{7} & \frac{5}{7} & -\frac{2}{7} & -\frac{6}{7} \\ \frac{3}{7} & -\frac{1}{7} & -\frac{1}{7} & -\frac{3}{7} \end{pmatrix} \begin{pmatrix} 5 \\ 6 \\ 7 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{18}{7} \\ 0 \\ \frac{11}{7} \\ \frac{2}{7} \end{pmatrix}$$

which agrees with the original way of doing the problem.

Two good books which give more discussion of linear programming are Strang [13] and Nobel and Daniels [10]. Also listed in these books are other references which may prove useful if you are interested in seeing more on these topics. There is a great deal more which can be said about linear programming.

8.6 Exercises

1. Maximize and minimize $z = x_1 - 2x_2 + x_3$ subject to the constraints $x_1 + x_2 + x_3 \leq 10$, $x_1 + x_2 + x_3 \geq 2$, and $x_1 + 2x_2 + x_3 \leq 7$ if possible. All variables are nonnegative.

2. Maximize and minimize the following is possible. All variables are nonnegative.

   (a) $z = x_1 - 2x_2$ subject to the constraints $x_1 + x_2 + x_3 \leq 10$, $x_1 + x_2 + x_3 \geq 1$, and $x_1 + 2x_2 + x_3 \leq 7$

   (b) $z = x_1 - 2x_2 - 3x_3$ subject to the constraints $x_1 + x_2 + x_3 \leq 8$, $x_1 + x_2 + 3x_3 \geq 1$, and $x_1 + x_2 + x_3 \leq 7$

   (c) $z = 2x_1 + x_2$ subject to the constraints $x_1 - x_2 + x_3 \leq 10$, $x_1 + x_2 + x_3 \geq 1$, and $x_1 + 2x_2 + x_3 \leq 7$

   (d) $z = x_1 + 2x_2$ subject to the constraints $x_1 - x_2 + x_3 \leq 10$, $x_1 + x_2 + x_3 \geq 1$, and $x_1 + 2x_2 + x_3 \leq 7$

3. Consider contradictory constraints, $x_1 + x_2 \geq 12$ and $x_1 + 2x_2 \leq 5$. You know these two contradict but show they contradict using the simplex algorithm.

4. Find a solution to the following inequalities $x, y \geq 0$ and if it is possible to do so. If it is not possible, prove it is not possible.

   (a)
   $$\begin{aligned} 6x + 3y &\geq 4 \\ 8x + 4y &\leq 5 \end{aligned}$$

   (b)
   $$\begin{aligned} 6x_1 + 4x_3 &\leq 11 \\ 5x_1 + 4x_2 + 4x_3 &\geq 8 \\ 6x_1 + 6x_2 + 5x_3 &\leq 11 \end{aligned}$$

   (c)
   $$\begin{aligned} 6x_1 + 4x_3 &\leq 11 \\ 5x_1 + 4x_2 + 4x_3 &\geq 9 \\ 6x_1 + 6x_2 + 5x_3 &\leq 9 \end{aligned}$$

   (d)
   $$\begin{aligned} x_1 - x_2 + x_3 &\leq 2 \\ x_1 + 2x_2 &\geq 4 \\ 3x_1 + 2x_3 &\leq 7 \end{aligned}$$

   (e)
   $$\begin{aligned} 5x_1 - 2x_2 + 4x_3 &\leq 1 \\ 6x_1 - 3x_2 + 5x_3 &\geq 2 \\ 5x_1 - 2x_2 + 4x_3 &\leq 5 \end{aligned}$$

5. Minimize $z = x_1 + x_2$ subject to $x_1 + x_2 \geq 2,\ x_1 + 3x_2 \leq 20,\ x_1 + x_2 \leq 18$. Change to a maximization problem and solve as follows: Let $y_i = M - x_i$. Formulate in terms of y_1, y_2.

Spectral Theory

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.

9.1 Eigenvalues And Eigenvectors Of A Matrix

The field of scalars in spectral theory is best taken to equal $\mathbb{C}$ although I will sometimes refer to it as $\mathbb{F}$.

Definition 9.1.1 *Let M be an $n \times n$ matrix and let $\mathbf{x} \in \mathbb{C}^n$ be a nonzero vector for which*

$$M\mathbf{x} = \lambda\mathbf{x} \tag{9.1}$$

for some scalar, λ. Then $\mathbf{x}$ is called an eigenvector and λ is called an eigenvalue (characteristic value) of the matrix, M.

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.

Eigenvectors are vectors which are shrunk, stretched or reflected upon multiplication by a matrix. How can they be identified? Suppose $\mathbf{x}$ satisfies 9.1. Then

$$(\lambda I - M)\mathbf{x} = \mathbf{0}$$

for some $\mathbf{x} \neq \mathbf{0}$. Therefore, the matrix $M - \lambda I$ cannot have an inverse and so by Theorem 6.3.15

$$\det(\lambda I - M) = 0. \tag{9.2}$$

In other words, λ must be a zero of the characteristic polynomial. Since M is an $n \times n$ matrix, it follows from the theorem on expanding a matrix by its cofactor that this is a polynomial equation of degree n. As such, it has a solution, $\lambda \in \mathbb{C}$. Is it actually an eigenvalue? The answer is yes and this follows from Theorem 6.3.23 on Page 106. Since $\det(\lambda I - M) = 0$ the matrix, $\lambda I - M$ cannot be one to one and so there exists a nonzero vector, $\mathbf{x}$ such that $(\lambda I - M)\mathbf{x} = \mathbf{0}$. This proves the following corollary.

Corollary 9.1.2 *Let M be an $n \times n$ matrix and $\det(M - \lambda I) = 0$. Then there exists $\mathbf{x} \in \mathbb{C}^n$ such that $(M - \lambda I)\mathbf{x} = \mathbf{0}$.*

As an example, consider the following.

Example 9.1.3 *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

$$\det\left(\lambda\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix}\right) = 0$$

When you expand this determinant, you find the equation is

$$(\lambda - 5)\left(\lambda^2 - 20\lambda + 100\right) = 0$$

and so the eigenvalues are

$$5, 10, 10.$$

I 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,

$$\left(5\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

That is you need to find the solution to

$$\begin{pmatrix} 0 & 10 & 5 \\ -2 & -9 & -2 \\ 4 & 8 & -1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

By now this is an old problem. You set up the augmented matrix and row reduce to get the solution. Thus the matrix you must row reduce is

$$\begin{pmatrix} 0 & 10 & 5 & 0 \\ -2 & -9 & -2 & 0 \\ 4 & 8 & -1 & 0 \end{pmatrix}. \tag{9.3}$$

The reduced row echelon form is

$$\begin{pmatrix} 1 & 0 & -\frac{5}{4} & 0 \\ 0 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and so the solution is any vector of the form

$$\begin{pmatrix} \frac{5}{4}z \\ \frac{-1}{2}z \\ z \end{pmatrix} = z\begin{pmatrix} \frac{5}{4} \\ \frac{-1}{2} \\ 1 \end{pmatrix}$$

where $z \in \mathbb{F}$. You would obtain the same collection of vectors if you replaced z with $4z$. Thus a simpler description for the solutions to this system of equations whose augmented matrix is in 9.3 is

$$z \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix} \tag{9.4}$$

where $z \in \mathbb{F}$. Now you need to remember that you can't take $z = 0$ 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 9.4 results in an eigenvector. It is a good idea to check your work! To do so, I will take the original matrix and multiply by this vector and see if I get 5 times this vector.

$$\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix} = \begin{pmatrix} 25 \\ -10 \\ 20 \end{pmatrix} = 5 \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix}$$

so it appears this is correct. Always check your work on these problems if you care about getting the answer right.

The variable, z is called a free variable or sometimes a parameter. The set of vectors in 9.4 is called the eigenspace and it equals $\ker(\lambda I - A)$. You should observe that in this case the eigenspace has dimension 1 because there is one vector which spans the eigenspace. 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,

$$\left(10 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} \right) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

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,

$$\begin{pmatrix} 5 & 10 & 5 & 0 \\ -2 & -4 & -2 & 0 \\ 4 & 8 & 4 & 0 \end{pmatrix}$$

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

$$\begin{pmatrix} -2y-z \\ y \\ z \end{pmatrix} = y\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + z\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

You can't pick z and y both equal to zero because this would result in the zero vector and

Eigenvectors are never equal to zero!

However, every other choice of z and y 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.

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 value of the eigenvalues. The eigenvalue problem is really not about row operations. A general rule to remember about the eigenvalue problem is this.

If it is not long and hard it is usually wrong!

The eigenvalue problem is the hardest problem in algebra and people still do research on ways to find eigenvalues. Now 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 is anything but easy because for an $n \times n$ matrix, it involves solving a polynomial equation of degree n and none of us are very good at doing this. 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, $(\lambda I - M)\,\mathbf{x} = \mathbf{0}$ will be the zero solution as explained above and

Eigenvectors are never equal to zero!

Here is another example.

Example 9.1.4 *Let*

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

First find the eigenvalues.

$$\det\left(\lambda\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 2 & 2 & -2 \\ 1 & 3 & -1 \\ -1 & 1 & 1 \end{pmatrix}\right) = 0$$

This is $\lambda^3 - 6\lambda^2 + 8\lambda = 0$ 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

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

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

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

where $z \neq 0$.

Next find the eigenvectors for $\lambda = 2$. The augmented matrix for the system of equations needed to find these eigenvectors is

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

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

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

where $z \neq 0$.

Finally find the eigenvectors for $\lambda = 4$. The augmented matrix for the system of equations needed to find these eigenvectors is

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

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

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

where $y \neq 0$.

Example 9.1.5 *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 I 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

$$\left(3\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 2 & -2 & -1 \\ -2 & -1 & -2 \\ 14 & 25 & 14 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The augmented matrix is

$$\begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 4 & 2 & 0 \\ -14 & -25 & -11 & 0 \end{pmatrix}$$

and the row reduced echelon form is

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

so the eigenvectors are nonzero vectors of the form

$$\begin{pmatrix} z \\ -z \\ z \end{pmatrix} = z\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

Next consider the eigenvectors for $\lambda = 6$. This requires you to solve

$$\left(6\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 2 & -2 & -1 \\ -2 & -1 & -2 \\ 14 & 25 & 14 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

and the augmented matrix for this system of equations is

$$\begin{pmatrix} 4 & 2 & 1 & 0 \\ 2 & 7 & 2 & 0 \\ -14 & -25 & -8 & 0 \end{pmatrix}$$

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & \frac{1}{8} & 0 \\ 0 & 1 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and so the eigenvectors for $\lambda = 6$ are of the form

$$z\begin{pmatrix} -\frac{1}{8} \\ -\frac{1}{4} \\ 1 \end{pmatrix}$$

or written more simply,

$$z\begin{pmatrix} -1 \\ -2 \\ 8 \end{pmatrix}$$

where $z \in \mathbb{F}$.

Note that in this example the eigenspace for the eigenvalue, $\lambda = 6$ is of dimension 1 because there is only one parameter which can be chosen. However, this eigenvalue is of multiplicity two as a root to the characteristic equation.

Definition 9.1.6 *If A is an $n \times n$ matrix with the property that some eigenvalue has algebraic multiplicity as a root of the characteristic equation which is greater than the dimension of the eigenspace associated with this eigenvalue, then the matrix is called defective.*

There may be repeated roots to the characteristic equation, 9.2 and it is not known whether the dimension of the eigenspace equals the multiplicity of the eigenvalue. However, the following theorem is available.

Theorem 9.1.7 *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, $\{\mathbf{v}_i\}_{i=1}^r$ is linearly independent.*

Proof: If the conclusion of this theorem is not true, then there exist non zero scalars, c_{k_j} such that

$$\sum_{j=1}^{m} c_{k_j}\mathbf{v}_{k_j} = \mathbf{0}. \tag{9.5}$$

Since any nonempty set of non negative integers has a smallest integer in the set, take m is as small as possible for this to take place. Then solving for $\mathbf{v}_{k_1}$

$$\mathbf{v}_{k_1} = \sum_{k_j \neq k_1} d_{k_j}\mathbf{v}_{k_j} \tag{9.6}$$

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 9.6 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 9.5 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.

In words, this theorem says that eigenvectors associated with distinct eigenvalues are linearly independent.

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 9.1.8 *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

$$\det\left(\lambda\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix}\right) = 0.$$

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,

$$\begin{pmatrix} 1+i & 0 & 0 & 0 \\ 0 & i & 1 & 0 \\ 0 & -1 & i & 0 \end{pmatrix}$$

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

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

Now multiply the second row by $-i$ to obtain

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -i & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Therefore, the eigenvectors are of the form

$$z\begin{pmatrix} 0 \\ i \\ 1 \end{pmatrix}.$$

You should find the eigenvectors for $\lambda = 2 - i$. These are

$$z\begin{pmatrix} 0 \\ -i \\ 1 \end{pmatrix}.$$

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.

9.2 Some Applications Of Eigenvalues And Eigenvectors

Recall that $n \times n$ matrices can be considered as linear transformations. If F is a 3×3 real matrix having positive determinant, it can be shown that $F = RU$ where R is a rotation matrix and U is a symmetric real matrix having positive eigenvalues. An application of this wonderful result, known to mathematicians as the right polar decomposition, is to continuum mechanics where a chunk of material is identified with a set of points in three dimensional space.

The linear transformation, F in this context is called the deformation gradient and it describes the local deformation of the material. Thus it is possible to consider this deformation in terms of two processes, one which distorts the material and the other which just rotates it. It is the matrix, U which is responsible for stretching and compressing. This is why in continuum mechanics, the stress is often taken to depend on U which is known in this context as the right Cauchy Green strain tensor. This process of writing a matrix as a product of two such matrices, one of which preserves distance and the other which distorts is also important in applications to geometric measure theory an interesting field of study in mathematics and to the study of quadratic forms which occur in many applications such as statistics. Here I am emphasizing the application to mechanics in which the eigenvectors of U determine the principle directions, those directions in which the material is stretched or compressed to the maximum extent.

Example 9.2.1 *Find the principle directions determined by the matrix,*

$$\begin{pmatrix} \frac{29}{11} & \frac{6}{11} & \frac{6}{11} \\ \frac{6}{11} & \frac{41}{44} & \frac{19}{44} \\ \frac{6}{11} & \frac{19}{44} & \frac{41}{44} \end{pmatrix}$$

The eigenvalues are $3, 1,$ *and* $\frac{1}{2}$.

It is nice to be given the eigenvalues. The largest eigenvalue is 3 which means that in the direction determined by the eigenvector associated with 3 the stretch is three times as large. The smallest eigenvalue is $1/2$ and so in the direction determined by the eigenvector for $1/2$ the material is compressed, becoming locally half as long. It remains to find these directions. First consider the eigenvector for 3. It is necessary to solve

$$\left(3\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{29}{11} & \frac{6}{11} & \frac{6}{11} \\ \frac{6}{11} & \frac{41}{44} & \frac{19}{44} \\ \frac{6}{11} & \frac{19}{44} & \frac{41}{44} \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Thus the augmented matrix for this system of equations is

$$\begin{pmatrix} \frac{4}{11} & -\frac{6}{11} & -\frac{6}{11} & 0 \\ -\frac{6}{11} & \frac{91}{44} & -\frac{19}{44} & 0 \\ -\frac{6}{11} & -\frac{19}{44} & \frac{91}{44} & 0 \end{pmatrix}$$

The row reduced echelon form is

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

and so the principle direction for the eigenvalue, 3 in which the material is stretched to the maximum extent is

$$\begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}.$$

A direction vector in this direction is

$$\begin{pmatrix} 3/\sqrt{11} \\ 1/\sqrt{11} \\ 1/\sqrt{11} \end{pmatrix}.$$

You should show that the direction in which the material is compressed the most is in the direction

$$\begin{pmatrix} 0 \\ -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

Note this is meaningful information which you would have a hard time finding without the theory of eigenvectors and eigenvalues.

Another application is to the problem of finding solutions to systems of differential equations. It turns out that vibrating systems involving masses and springs can be studied in the form

$$\mathbf{x}'' = A\mathbf{x} \tag{9.7}$$

where A is a real symmetric $n \times n$ matrix which has nonpositive eigenvalues. This is analogous to the case of the scalar equation for undamped oscillation, $x' + \omega^2 x = 0$. The main difference is that here the scalar ω^2 is replaced with the matrix, $-A$. Consider the problem of finding solutions to 9.7. You look for a solution which is in the form

$$\mathbf{x}(t) = \mathbf{v}e^{\lambda t} \tag{9.8}$$

and substitute this into 9.7. Thus

$$\mathbf{x}'' = \mathbf{v}\lambda^2 e^{\lambda t} = e^{\lambda t} A\mathbf{v}$$

and so

$$\lambda^2 \mathbf{v} = A\mathbf{v}.$$

Therefore, λ^2 needs to be an eigenvalue of A and $\mathbf{v}$ needs to be an eigenvector. Since A has nonpositive eigenvalues, $\lambda^2 = -a^2$ and so $\lambda = \pm ia$ where $-a^2$ is an eigenvalue of A. Corresponding to this you obtain solutions of the form

$$\mathbf{x}(t) = \mathbf{v}\cos(at), \mathbf{v}\sin(at).$$

Note these solutions oscillate because of the $\cos(at)$ and $\sin(at)$ in the solutions. Here is an example.

Example 9.2.2 *Find oscillatory solutions to the system of differential equations,* $\mathbf{x}'' = A\mathbf{x}$ *where*

$$A = \begin{pmatrix} -\frac{5}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{13}{6} & \frac{5}{6} \\ -\frac{1}{3} & \frac{5}{6} & -\frac{13}{6} \end{pmatrix}.$$

The eigenvalues are $-1, -2,$ *and* -3.

According to the above, you can find solutions by looking for the eigenvectors. Consider the eigenvectors for -3. The augmented matrix for finding the eigenvectors is

$$\begin{pmatrix} -\frac{4}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & -\frac{5}{6} & -\frac{5}{6} & 0 \\ \frac{1}{3} & -\frac{5}{6} & -\frac{5}{6} & 0 \end{pmatrix}$$

and its row echelon form is

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

Therefore, the eigenvectors are of the form

$$\mathbf{v} = z \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}.$$

It follows

$$\begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} \cos\left(\sqrt{3}t\right), \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} \sin\left(\sqrt{3}t\right)$$

are both solutions to the system of differential equations. You can find other oscillatory solutions in the same way by considering the other eigenvalues. You might try checking these answers to verify they work.

This is just a special case of a procedure used in differential equations to obtain closed form solutions to systems of differential equations using linear algebra. The overall philosophy is to take one of the easiest problems in analysis and change it into the eigenvalue problem which is the most difficult problem in algebra. However, when it works, it gives precise solutions in terms of known functions.

9.3 Exercises

1. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} -19 & -14 & -1 \\ 8 & 4 & 8 \\ 15 & 30 & -3 \end{pmatrix}.$$

   Determine whether the matrix is defective.

2. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} -3 & -30 & 15 \\ 0 & 12 & 0 \\ 15 & 30 & -3 \end{pmatrix}.$$

   Determine whether the matrix is defective.

3. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 8 & 4 & 5 \\ 0 & 12 & 9 \\ -2 & 2 & 10 \end{pmatrix}.$$

   Determine whether the matrix is defective.

4. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 12 & -12 & 6 \\ 0 & 18 & 0 \\ 6 & 12 & 12 \end{pmatrix}$$

5. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} -5 & -1 & 10 \\ -15 & 9 & -6 \\ 8 & -8 & 2 \end{pmatrix}.$$

Determine whether the matrix is defective.

6. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} -10 & -8 & 8 \\ -4 & -14 & -4 \\ 0 & 0 & -18 \end{pmatrix}.$$

Determine whether the matrix is defective.

7. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 1 & 26 & -17 \\ 4 & -4 & 4 \\ -9 & -18 & 9 \end{pmatrix}.$$

Determine whether the matrix is defective.

8. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 8 & 4 & 5 \\ 0 & 12 & 9 \\ -2 & 2 & 10 \end{pmatrix}.$$

Determine whether the matrix is defective.

9. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 9 & 6 & -3 \\ 0 & 6 & 0 \\ -3 & -6 & 9 \end{pmatrix}.$$

Determine whether the matrix is defective.

10. 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.

11. Find the complex eigenvalues and eigenvectors of the matrix $\begin{pmatrix} 4 & -2 & -2 \\ 0 & 2 & -2 \\ 2 & 0 & 2 \end{pmatrix}$. Determine whether the matrix is defective.

12. Find the complex eigenvalues and eigenvectors of the matrix $\begin{pmatrix} -4 & 2 & 0 \\ 2 & -4 & 0 \\ -2 & 2 & -2 \end{pmatrix}$. Determine whether the matrix is defective.

13. 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.

14. Find the complex eigenvalues and eigenvectors of the matrix $\begin{pmatrix} 4 & 2 & 0 \\ 2 & 4 & 0 \\ -2 & 2 & 6 \end{pmatrix}$. Determine whether the matrix is defective.

15. 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 $a-ib$.

16. Recall an $n \times n$ matrix is said to be symmetric if it has all real entries and if $A = A^T$. Show the eigenvectors and eigenvalues of a real symmetric matrix are real.

17. Recall an $n \times n$ matrix is said to be skew symmetric if it has all real entries and if $A = -A^T$. Show that any nonzero eigenvalues must be of the form ib where $i^2 = -1$. In words, the eigenvalues are either 0 or pure imaginary. Show also that the eigenvectors corresponding to the pure imaginary eigenvalues are imaginary in the sense that every entry is of the form ix for $x \in \mathbb{R}$.

18. Is it possible for a nonzero matrix to have only 0 as an eigenvalue?

19. Show that if $A\mathbf{x} = \lambda\mathbf{x}$ and $A\mathbf{y} = \lambda\mathbf{y}$, then whenever a, b are scalars,

$$A(a\mathbf{x}+b\mathbf{y}) = \lambda(a\mathbf{x}+b\mathbf{y}).$$

20. 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.

21. Show that the eigenvalues and eigenvectors of a real matrix occur in conjugate pairs.

22. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 7 & -2 & 0 \\ 8 & -1 & 0 \\ -2 & 4 & 6 \end{pmatrix}.$$

Can you find three independent eigenvectors?

23. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 3 & -2 & -1 \\ 0 & 5 & 1 \\ 0 & 2 & 4 \end{pmatrix}.$$

Can you find three independent eigenvectors in this case?

24. Let M be an $n \times n$ matrix and suppose $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are n eigenvectors which form a linearly independent set. Form the matrix S by making the columns these vectors. Show that S^{-1} exists and that $S^{-1}MS$ is a diagonal matrix (one having zeros everywhere except on the main diagonal) having the eigenvalues of M on the main diagonal. When this can be done the matrix is diagonalizable.

25. Show that a matrix, M is diagonalizable if and only if it has a basis of eigenvectors. **Hint:** The first part is done in Problem 3. It only remains to show that if the matrix can be diagonalized by some matrix, S giving $D = S^{-1}MS$ for D a diagonal matrix, then it has a basis of eigenvectors. Try using the columns of the matrix S.

26. Find the principle directions determined by the matrix, $\begin{pmatrix} \frac{7}{12} & -\frac{1}{4} & \frac{1}{6} \\ -\frac{1}{4} & \frac{7}{12} & -\frac{1}{6} \\ \frac{1}{6} & -\frac{1}{6} & \frac{2}{3} \end{pmatrix}$. The eigenvalues are $\frac{1}{3}, 1$, and $\frac{1}{2}$ listed according to multiplicity.

27. Find the principle directions determined by the matrix, $\begin{pmatrix} \frac{5}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{7}{6} & \frac{1}{6} \\ -\frac{1}{3} & \frac{1}{6} & \frac{7}{6} \end{pmatrix}$ The eigenvalues are $1, 2$, and 1. What is the physical interpretation of the repeated eigenvalue?

28. Find the principle directions determined by the matrix, $\begin{pmatrix} \frac{19}{54} & \frac{1}{27} & \frac{1}{27} \\ \frac{1}{27} & \frac{11}{27} & \frac{2}{27} \\ \frac{1}{27} & \frac{2}{27} & \frac{11}{27} \end{pmatrix}$ The eigenvalues are $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{3}$.

29. Find the principle directions determined by the matrix, $\begin{pmatrix} \frac{3}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{3}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{3}{2} \end{pmatrix}$ The eigenvalues are 2, $\frac{1}{2}$, and 2. What is the physical interpretation of the repeated eigenvalue?

30. Find oscillatory solutions to the system of differential equations, $\mathbf{x}'' = A\mathbf{x}$ where $A = \begin{pmatrix} -3 & -1 & -1 \\ -1 & -2 & 0 \\ -1 & 0 & -2 \end{pmatrix}$ The eigenvalues are $-1, -4$, and -2.

31. Find oscillatory solutions to the system of differential equations, $\mathbf{x}'' = A\mathbf{x}$ where $A = \begin{pmatrix} -\frac{7}{3} & -\frac{2}{3} & -\frac{2}{3} \\ -\frac{2}{3} & -\frac{11}{6} & \frac{1}{6} \\ -\frac{2}{3} & \frac{1}{6} & -\frac{11}{6} \end{pmatrix}$ The eigenvalues are $-1, -3$, and -2.

9.4 Exercises

1. Let A and B be $n \times n$ matrices and let the columns of B be

$$\mathbf{b}_1, \cdots, \mathbf{b}_n$$

and the rows of A are

$$\mathbf{a}_1^T, \cdots, \mathbf{a}_n^T.$$

Show the columns of AB are

$$A\mathbf{b}_1 \cdots A\mathbf{b}_n$$

and the rows of AB are

$$\mathbf{a}_1^T B \cdots \mathbf{a}_n^T B.$$

2. 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.

3. Let M be an $n \times n$ matrix and suppose $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are n eigenvectors which form a linearly independent set. Form the matrix S by making the columns these vectors. Show that S^{-1} exists and that $S^{-1}MS$ is a diagonal matrix (one having zeros everywhere except on the main diagonal) having the eigenvalues of M on the main diagonal. When this can be done the matrix is diagonalizable.

4. Show that a matrix, M is diagonalizable if and only if it has a basis of eigenvectors. **Hint:** The first part is done in Problem 3. It only remains to show that if the matrix can be diagonalized by some matrix, S giving $D = S^{-1}MS$ for D a diagonal matrix, then it has a basis of eigenvectors. Try using the columns of the matrix S.

5. Let

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

and let

$$B = \begin{pmatrix} \boxed{\begin{matrix} 0 & 1 \\ 1 & 1 \end{matrix}} \\ \boxed{\begin{matrix} 2 & 1 \end{matrix}} \end{pmatrix}$$

Multiply AB verifying the block multiplication formula. Here $A_{11} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $A_{12} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$, $A_{21} = \begin{pmatrix} 0 & 1 \end{pmatrix}$ and $A_{22} = (3)$.

9.5 Shur's Theorem

Every matrix is related to an upper triangular matrix in a particularly significant way. This is Shur's theorem and it is the most important theorem in the spectral theory of matrices.

Lemma 9.5.1 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a basis for $\mathbb{F}^n$. Then there exists an orthonormal basis for $\mathbb{F}^n$, $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ which has the property that for each $k \leq n$, span$(\mathbf{x}_1, \cdots, \mathbf{x}_k) =$* span $(\mathbf{u}_1, \cdots, \mathbf{u}_k)$.

Proof: Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a basis for $\mathbb{F}^n$. Let $\mathbf{u}_1 \equiv \mathbf{x}_1 / |\mathbf{x}_1|$. Thus for $k = 1$, $\mathrm{span}(\mathbf{u}_1) = \mathrm{span}(\mathbf{x}_1)$ and $\{\mathbf{u}_1\}$ is an orthonormal set. Now suppose for some $k < n$, $\mathbf{u}_1, \cdots, \mathbf{u}_k$ have been chosen such that $(\mathbf{u}_j \cdot \mathbf{u}_l) = \delta_{jl}$ and $\mathrm{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \mathrm{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$. Then define

$$\mathbf{u}_{k+1} \equiv \frac{\mathbf{x}_{k+1} - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)\, \mathbf{u}_j}{\left| \mathbf{x}_{k+1} - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)\, \mathbf{u}_j \right|}, \tag{9.9}$$

where the denominator is not equal to zero because the $\mathbf{x}_j$ form a basis and so

$$\mathbf{x}_{k+1} \notin \mathrm{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \mathrm{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$$

Thus by induction,

$$\mathbf{u}_{k+1} \in \mathrm{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{x}_{k+1}) = \mathrm{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{x}_{k+1}).$$

Also, $\mathbf{x}_{k+1} \in \mathrm{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1})$ which is seen easily by solving 9.9 for $\mathbf{x}_{k+1}$ and it follows

$$\mathrm{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{x}_{k+1}) = \mathrm{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1}).$$

If $l \leq k$,

$$\begin{aligned}
(\mathbf{u}_{k+1} \cdot \mathbf{u}_l) &= C\left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)(\mathbf{u}_j \cdot \mathbf{u}_l) \right) \\
&= C\left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)\, \delta_{lj} \right) \\
&= C\left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - (\mathbf{x}_{k+1} \cdot \mathbf{u}_l) \right) = 0.
\end{aligned}$$

The vectors, $\{\mathbf{u}_j\}_{j=1}^{n}$, generated in this way are therefore an orthonormal basis because each vector has unit length.

The process by which these vectors were generated is called the Gram Schmidt process. Recall the following definition.

Definition 9.5.2 *An $n \times n$ matrix, U, is unitary if $UU^* = I = U^*U$ where U^* is defined to be the transpose of the conjugate of U.*

Theorem 9.5.3 *Let A be an $n \times n$ matrix. Then there exists a unitary matrix, U such that*

$$U^*AU = T, \tag{9.10}$$

where T is an upper triangular matrix having the eigenvalues of A on the main diagonal listed according to multiplicity as roots of the characteristic equation.

Proof: Let $\mathbf{v}_1$ be a unit eigenvector for A. Then there exists λ_1 such that

$$A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1, \ |\mathbf{v}_1| = 1.$$

Extend $\{\mathbf{v}_1\}$ to a basis and then use Lemma 9.5.1 to obtain $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$, an orthonormal basis in $\mathbb{F}^n$. Let U_0 be a matrix whose i^{th} column is $\mathbf{v}_i$. Then from the above, it follows U_0 is unitary. Then $U_0^* A U_0$ is of the form

$$\begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$$

where A_1 is an $n-1 \times n-1$ matrix. Repeat the process for the matrix, A_1 above. There exists a unitary matrix $\widetilde{U}_1$ such that $\widetilde{U}_1^* A_1 \widetilde{U}_1$ is of the form

$$\begin{pmatrix} \lambda_2 & * & \cdots & * \\ 0 & & & \\ \vdots & & A_2 & \\ 0 & & & \end{pmatrix}.$$

Now let U_1 be the $n \times n$ matrix of the form

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix}.$$

This is also a unitary matrix because by block multiplication,

$$\begin{aligned} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix}^* \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix} &= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1^* \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1^* \widetilde{U}_1 \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix} \end{aligned}$$

Then using block multiplication, $U_1^* U_0^* A U_0 U_1$ is of the form

$$\begin{pmatrix} \lambda_1 & * & * & \cdots & * \\ 0 & \lambda_2 & * & \cdots & * \\ 0 & 0 & & & \\ \vdots & \vdots & & A_2 & \\ 0 & 0 & & & \end{pmatrix}$$

where A_2 is an $n-2 \times n-2$ matrix. Continuing in this way, there exists a unitary matrix, U given as the product of the U_i in the above construction such that

$$U^* A U = T$$

where T is some upper triangular matrix. Since the matrix is upper triangular, the characteristic equation is $\prod_{i=1}^n (\lambda - \lambda_i)$ where the λ_i are the diagonal entries of T. Therefore, the λ_i are the eigenvalues.

What if A is a real matrix and you only want to consider real unitary matrices?

Theorem 9.5.4 *Let A be a real $n \times n$ matrix. Then there exists a real unitary matrix, Q and a matrix T of the form*

$$T = \begin{pmatrix} P_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & P_r \end{pmatrix} \tag{9.11}$$

where P_i equals either a real 1×1 matrix or P_i equals a real 2×2 matrix having two complex eigenvalues of A such that $Q^T A Q = T$. The matrix, T is called the real Schur form of the matrix A.

Proof: Suppose

$$A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1, \ |\mathbf{v}_1| = 1$$

where λ_1 is real. Then let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be an orthonormal basis of vectors in $\mathbb{R}^n$. Let Q_0 be a matrix whose i^{th} column is $\mathbf{v}_i$. Then $Q_0^* A Q_0$ is of the form

$$\begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$$

where A_1 is a real $n-1 \times n-1$ matrix. This is just like the proof of Theorem 9.5.3 up to this point.

Now in case $\lambda_1 = \alpha + i\beta$, it follows since A is real that $\mathbf{v}_1 = \mathbf{z}_1 + i\mathbf{w}_1$ and that $\overline{\mathbf{v}}_1 = \mathbf{z}_1 - i\mathbf{w}_1$ is an eigenvector for the eigenvalue, $\alpha - i\beta$. Here $\mathbf{z}_1$ and $\mathbf{w}_1$ are real vectors. It is clear that $\{\mathbf{z}_1, \mathbf{w}_1\}$ is an independent set of vectors in $\mathbb{R}^n$. Indeed,$\{\mathbf{v}_1, \overline{\mathbf{v}}_1\}$ is an independent set and it follows $\text{span}(\mathbf{v}_1, \overline{\mathbf{v}}_1) = \text{span}(\mathbf{z}_1, \mathbf{w}_1)$. Now using the Gram Schmidt theorem in $\mathbb{R}^n$, there exists $\{\mathbf{u}_1, \mathbf{u}_2\}$, an orthonormal set of real vectors such that $\text{span}(\mathbf{u}_1, \mathbf{u}_2) = \text{span}(\mathbf{v}_1, \overline{\mathbf{v}}_1)$. Now let $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$ be an orthonormal basis in $\mathbb{R}^n$ and let Q_0 be a unitary matrix whose i^{th} column is $\mathbf{u}_i$. Then $A\mathbf{u}_j$ are both in $\text{span}(\mathbf{u}_1, \mathbf{u}_2)$ for $j = 1, 2$ and so $\mathbf{u}_k^T A \mathbf{u}_j = 0$ whenever $k \geq 3$. It follows that $Q_0^* A Q_0$ is of the form

$$\begin{pmatrix} * & * & \cdots & * \\ * & * & & \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$$

where A_1 is now an $n-2 \times n-2$ matrix. In this case, find $\widetilde{Q}_1$ an $n-2 \times n-2$ matrix to put A_1 in an appropriate form as above and come up with A_2 either an $n-4 \times n-4$ matrix or an $n-3 \times n-3$ matrix. Then the only other difference is to let

$$Q_1 = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & & & \\ \vdots & \vdots & & \widetilde{Q}_1 & \\ 0 & 0 & & & \end{pmatrix}$$

thus putting a 2×2 identity matrix in the upper left corner rather than a one. Repeating this process with the above modification for the case of a complex eigenvalue leads eventually to 9.11 where Q is the product of real unitary matrices Q_i above. Finally,

$$\lambda I - T = \begin{pmatrix} \lambda I_1 - P_1 & \cdots & * \\ & \ddots & \vdots \\ 0 & & \lambda I_r - P_r \end{pmatrix}$$

where I_k is the 2×2 identity matrix in the case that P_k is 2×2 and is the number 1 in the case where P_k is a 1×1 matrix. Now, it follows that $\det(\lambda I - T) = \prod_{k=1}^r \det(\lambda I_k - P_k)$. Therefore, λ is an eigenvalue of T if and only if it is an eigenvalue of some P_k. This proves the theorem since the eigenvalues of T are the same as those of A because they have the same characteristic polynomial due to the similarity of A and T.

Definition 9.5.5 *When a linear transformation, A, mapping a linear space, V to V has a basis of eigenvectors, the linear transformation is called non defective. Otherwise it is*

called defective. An $n \times n$ matrix, A, is called normal if $AA^* = A^*A$. An important class of normal matrices is that of the Hermitian or self adjoint matrices. An $n \times n$ matrix, A is self adjoint or Hermitian if $A = A^*$.

The next lemma is the basis for concluding that every normal matrix is unitarily similar to a diagonal matrix.

Lemma 9.5.6 *If T is upper triangular and normal, then T is a diagonal matrix.*

Proof: Since T is normal, $T^*T = TT^*$. Writing this in terms of components and using the description of the adjoint as the transpose of the conjugate, yields the following for the ik^{th} entry of $T^*T = TT^*$.

$$\sum_j t_{ij}t^*_{jk} = \sum_j t_{ij}\overline{t_{kj}} = \sum_j t^*_{ij}t_{jk} = \sum_j \overline{t_{ji}}t_{jk}.$$

Now use the fact that T is upper triangular and let $i = k = 1$ to obtain the following from the above.

$$\sum_j |t_{1j}|^2 = \sum_j |t_{j1}|^2 = |t_{11}|^2$$

You see, $t_{j1} = 0$ unless $j = 1$ due to the assumption that T is upper triangular. This shows T is of the form

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

Now do the same thing only this time take $i = k = 2$ and use the result just established. Thus, from the above,

$$\sum_j |t_{2j}|^2 = \sum_j |t_{j2}|^2 = |t_{22}|^2,$$

showing that $t_{2j} = 0$ if $j > 2$ which means T has the form

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

Next let $i = k = 3$ and obtain that T looks like a diagonal matrix in so far as the first 3 rows and columns are concerned. Continuing in this way it follows T is a diagonal matrix.

Theorem 9.5.7 *Let A be a normal matrix. Then there exists a unitary matrix, U such that U^*AU is a diagonal matrix.*

Proof: From Theorem 9.5.3 there exists a unitary matrix, U such that U^*AU equals an upper triangular matrix. The theorem is now proved if it is shown that the property of being normal is preserved under unitary similarity transformations. That is, verify that if A is normal and if $B = U^*AU$, then B is also normal. But this is easy.

$$\begin{aligned} B^*B &= U^*A^*UU^*AU = U^*A^*AU \\ &= U^*AA^*U = U^*AUU^*A^*U = BB^*. \end{aligned}$$

Therefore, U^*AU is a normal and upper triangular matrix and by Lemma 9.5.6 it must be a diagonal matrix. This proves the theorem.

Corollary 9.5.8 *If A is Hermitian, then all the eigenvalues of A are real and there exists an orthonormal basis of eigenvectors.*

Proof: Since A is normal, there exists unitary, U such that $U^*AU = D$, a diagonal matrix whose diagonal entries are the eigenvalues of A. Therefore, $D^* = U^*A^*U = U^*AU = D$ showing D is real.

Finally, let

$$U = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{pmatrix}$$

where the $\mathbf{u}_i$ denote the columns of U and

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

The equation, $U^*AU = D$ implies

$$\begin{aligned} AU &= \begin{pmatrix} A\mathbf{u}_1 & A\mathbf{u}_2 & \cdots & A\mathbf{u}_n \end{pmatrix} \\ &= UD = \begin{pmatrix} \lambda_1\mathbf{u}_1 & \lambda_2\mathbf{u}_2 & \cdots & \lambda_n\mathbf{u}_n \end{pmatrix} \end{aligned}$$

where the entries denote the columns of AU and UD respectively. Therefore, $A\mathbf{u}_i = \lambda_i\mathbf{u}_i$ and since the matrix is unitary, the ij^{th} entry of U^*U equals δ_{ij} and so

$$\delta_{ij} = \overline{\mathbf{u}}_i^T\mathbf{u}_j = \overline{\mathbf{u}_i^T\overline{\mathbf{u}}_j} = \overline{\mathbf{u}_i \cdot \mathbf{u}_j}.$$

This proves the corollary because it shows the vectors $\{\mathbf{u}_i\}$ form an orthonormal basis.

Corollary 9.5.9 *If A is a real symmetric matrix, then A is Hermitian and there exists a real unitary matrix, U such that $U^TAU = D$ where D is a diagonal matrix.*

Proof: This follows from Theorem 9.5.4 and Corollary 9.5.8.

9.6 Quadratic Forms

Definition 9.6.1 *A quadratic form in three dimensions is an expression of the form*

$$\begin{pmatrix} x & y & z \end{pmatrix} A \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{9.12}$$

where A is a 3×3 symmetric matrix. In higher dimensions the idea is the same except you use a larger symmetric matrix in place of A. In two dimensions A is a 2×2 matrix.

For example, consider

$$\begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} 3 & -4 & 1 \\ -4 & 0 & -4 \\ 1 & -4 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{9.13}$$

which equals $3x^2 - 8xy + 2xz - 8yz + 3z^2$. This is very awkward because of the mixed terms such as $-8xy$. The idea is to pick different axes such that if x, y, z are taken with respect

to these axes, the quadratic form is much simpler. In other words, look for new variables, x', y', and z' and a unitary matrix, U such that

$$U \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{9.14}$$

and if you write the quadratic form in terms of the primed variables, there will be no mixed terms. Any symmetric real matrix is Hermitian and is therefore normal. From Corollary 9.5.9, it follows there exists a real unitary matrix, U, (an orthogonal matrix) such that $U^T AU = D$ a diagonal matrix. Thus in the quadratic form, 9.12

$$\begin{aligned} \begin{pmatrix} x & y & z \end{pmatrix} A \begin{pmatrix} x \\ y \\ z \end{pmatrix} &= \begin{pmatrix} x' & y' & z' \end{pmatrix} U^T AU \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} \\ &= \begin{pmatrix} x' & y' & z' \end{pmatrix} D \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} \end{aligned}$$

and in terms of these new variables, the quadratic form becomes

$$\lambda_1 (x')^2 + \lambda_2 (y')^2 + \lambda_3 (z')^2$$

where $D = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3)$. Similar considerations apply equally well in any other dimension. For the given example,

$$\begin{pmatrix} -\frac{1}{2}\sqrt{2} & 0 & \frac{1}{2}\sqrt{2} \\ \frac{1}{6}\sqrt{6} & \frac{1}{3}\sqrt{6} & \frac{1}{6}\sqrt{6} \\ \frac{1}{3}\sqrt{3} & -\frac{1}{3}\sqrt{3} & \frac{1}{3}\sqrt{3} \end{pmatrix} \begin{pmatrix} 3 & -4 & 1 \\ -4 & 0 & -4 \\ 1 & -4 & 3 \end{pmatrix} \cdot$$

$$\begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 8 \end{pmatrix}$$

and so if the new variables are given by

$$\begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix},$$

it follows that in terms of the new variables the quadratic form is $2(x')^2 - 4(y')^2 + 8(z')^2$. You can work other examples the same way.

9.7 Second Derivative Test

Here is a second derivative test for functions of n variables.

Theorem 9.7.1 *Let U be an open subset of $\mathbb{C}^n$ and suppose $f : U \to \mathbb{C}^m$, $D^2\mathbf{f}(\mathbf{x})$ exists for all $\mathbf{x} \in U$ and $D^2\mathbf{f}$ is continuous at $\mathbf{x} \in U$. Then*

$$D^2\mathbf{f}(\mathbf{x})(\mathbf{u})(\mathbf{v}) = D^2\mathbf{f}(\mathbf{x})(\mathbf{v})(\mathbf{u}).$$

Proof: Let $B(\mathbf{x},r) \subseteq U$ and let $t,s \in (0,r/2]$. Now let $\mathbf{z} \in Y$ and define

$$\Delta(s,t) \equiv \mathrm{Re}\left(\frac{1}{st}\{\mathbf{f}(\mathbf{x}+t\mathbf{u}+s\mathbf{v}) - \mathbf{f}(\mathbf{x}+t\mathbf{u}) - (\mathbf{f}(\mathbf{x}+s\mathbf{v}) - \mathbf{f}(\mathbf{x}))\}, \mathbf{z}\right). \tag{9.15}$$

Let $h(t) = \mathrm{Re}(\mathbf{f}(\mathbf{x}+s\mathbf{v}+t\mathbf{u}) - \mathbf{f}(\mathbf{x}+t\mathbf{u}), \mathbf{z})$. Then by the mean value theorem,

$$\begin{aligned}\Delta(s,t) &= \frac{1}{st}(h(t) - h(0)) = \frac{1}{st}h'(\alpha t)\, t \\ &= \frac{1}{s}\mathrm{Re}(D\mathbf{f}(\mathbf{x}+s\mathbf{v}+\alpha t\mathbf{u})\,\mathbf{u} - D\mathbf{f}(\mathbf{x}+\alpha t\mathbf{u})\,\mathbf{u}, \mathbf{z}).\end{aligned}$$

Applying the mean value theorem again,

$$\Delta(s,t) = \mathrm{Re}\left(D^2\mathbf{f}(\mathbf{x}+\beta s\mathbf{v}+\alpha t\mathbf{u})(\mathbf{v})(\mathbf{u}), \mathbf{z}\right)$$

where $\alpha,\beta \in (0,1)$. If the terms $\mathbf{f}(\mathbf{x}+t\mathbf{u})$ and $\mathbf{f}(\mathbf{x}+s\mathbf{v})$ are interchanged in 9.15, $\Delta(s,t)$ is also unchanged and the above argument shows there exist $\gamma,\delta \in (0,1)$ such that

$$\Delta(s,t) = \mathrm{Re}\left(D^2\mathbf{f}(\mathbf{x}+\gamma s\mathbf{v}+\delta t\mathbf{u})(\mathbf{u})(\mathbf{v}), \mathbf{z}\right).$$

Letting $(s,t) \to (0,0)$ and using the continuity of $D^2\mathbf{f}$ at $\mathbf{x}$,

$$\lim_{(s,t)\to(0,0)} \Delta(s,t) = \mathrm{Re}\left(D^2\mathbf{f}(\mathbf{x})(\mathbf{u})(\mathbf{v}), \mathbf{z}\right) = \mathrm{Re}\left(D^2\mathbf{f}(\mathbf{x})(\mathbf{v})(\mathbf{u}), \mathbf{z}\right).$$

Since $\mathbf{z}$ is arbitrary, this demonstrates the conclusion of the theorem.

Consider the important special case of $\mathbb{R}^n$ and $\mathbb{R}$. If $\mathbf{e}_i$ are the standard basis vectors, what is

$$D^2 f(\mathbf{x})(\mathbf{e}_i)(\mathbf{e}_j)?$$

To see what this is, use the definition to write

$$\begin{aligned}D^2 f(\mathbf{x})(\mathbf{e}_i)(\mathbf{e}_j) &= t^{-1}s^{-1}D^2 f(\mathbf{x})(t\mathbf{e}_i)(s\mathbf{e}_j) \\ &= t^{-1}s^{-1}(Df(\mathbf{x}+t\mathbf{e}_i) - Df(\mathbf{x}) + o(t))(s\mathbf{e}_j) \\ &= t^{-1}s^{-1}(f(\mathbf{x}+t\mathbf{e}_i + s\mathbf{e}_j) - f(\mathbf{x}+t\mathbf{e}_i) \\ &+ o(s) - (f(\mathbf{x}+s\mathbf{e}_j) - f(\mathbf{x}) + o(s)) + o(t)\, s).\end{aligned}$$

First let $s \to 0$ to get

$$t^{-1}\left(\frac{\partial f}{\partial x_j}(\mathbf{x}+t\mathbf{e}_i) - \frac{\partial f}{\partial x_j}(\mathbf{x}) + o(t)\right)$$

and then let $t \to 0$ to obtain

$$D^2 f(\mathbf{x})(\mathbf{e}_i)(\mathbf{e}_j) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) \tag{9.16}$$

Thus the theorem asserts that in this special case the mixed partial derivatives are equal at $\mathbf{x}$ if they are defined near $\mathbf{x}$ and continuous at $\mathbf{x}$.

Definition 9.7.2 *The matrix,* $\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})\right)$ *is called the Hessian matrix.*

Now recall the Taylor formula with the Lagrange form of the remainder. See any good non reformed calculus book for a proof of this theorem. Ellis and Gulleck has a good proof.

Theorem 9.7.3 *Let $h : (-\delta, 1+\delta) \to \mathbb{R}$ have $m+1$ derivatives. Then there exists $t \in [0,1]$ such that*

$$h(1) = h(0) + \sum_{k=1}^{m} \frac{h^{(k)}(0)}{k!} + \frac{h^{(m+1)}(t)}{(m+1)!}.$$

Now let $f : U \to \mathbb{R}$ where $U \subseteq X$ a normed linear space and suppose $f \in C^m(U)$. 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

$$h'(t) = Df(\mathbf{x}+t\mathbf{v})(\mathbf{v}), \; h''(t) = D^2 f(\mathbf{x}+t\mathbf{v})(\mathbf{v})(\mathbf{v})$$

and continuing in this way,

$$h^{(k)}(t) = D^{(k)} f(\mathbf{x}+t\mathbf{v})(\mathbf{v})(\mathbf{v}) \cdots (\mathbf{v}) \equiv D^{(k)} f(\mathbf{x}+t\mathbf{v})\,\mathbf{v}^k.$$

It follows from Taylor's formula for a function of one variable,

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \sum_{k=1}^{m} \frac{D^{(k)} f(\mathbf{x})\,\mathbf{v}^k}{k!} + \frac{D^{(m+1)} f(\mathbf{x}+t\mathbf{v})\,\mathbf{v}^{m+1}}{(m+1)!}. \tag{9.17}$$

This proves the following theorem.

Theorem 9.7.4 *Let $f : U \to \mathbb{R}$ and let $f \in C^{m+1}(U)$. Then if*

$$B(\mathbf{x}, r) \subseteq U,$$

and $||\mathbf{v}|| < r$, there exists $t \in (0,1)$ such that 9.17 holds.

Now consider the case where $U \subseteq \mathbb{R}^n$ and $f : U \to \mathbb{R}$ is $C^2(U)$. Then from Taylor's theorem, if $\mathbf{v}$ is small enough, there exists $t \in (0,1)$ such that

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + Df(\mathbf{x})\,\mathbf{v} + \frac{D^2 f(\mathbf{x}+t\mathbf{v})\,\mathbf{v}^2}{2}.$$

Letting

$$\mathbf{v} = \sum_{i=1}^{n} v_i \mathbf{e}_i,$$

where $\mathbf{e}_i$ are the usual basis vectors, the second derivative term reduces to

$$\frac{1}{2} \sum_{i,j} D^2 f(\mathbf{x}+t\mathbf{v})(\mathbf{e}_i)(\mathbf{e}_j)\, v_i v_j = \frac{1}{2} \sum_{i,j} H_{ij}(\mathbf{x}+t\mathbf{v})\, v_i v_j$$

where

$$H_{ij}(\mathbf{x}+t\mathbf{v}) = D^2 f(\mathbf{x}+t\mathbf{v})(\mathbf{e}_i)(\mathbf{e}_j) = \frac{\partial^2 f(\mathbf{x}+t\mathbf{v})}{\partial x_j \partial x_i},$$

the Hessian matrix. From Theorem 9.7.1, this is a symmetric matrix. By the continuity of the second partial derivative and this,

$$f(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + Df(\mathbf{x})\,\mathbf{v} + \frac{1}{2}\mathbf{v}^T H(\mathbf{x})\,\mathbf{v} +$$

$$\frac{1}{2}\left(\mathbf{v}^T\left(H\left(\mathbf{x}+t\mathbf{v}\right)-H\left(\mathbf{x}\right)\right)\mathbf{v}\right). \tag{9.18}$$

where the last two terms involve ordinary matrix multiplication and

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

for v_i the components of $\mathbf{v}$ relative to the standard basis.

Theorem 9.7.5 *In the above situation, suppose $Df\left(\mathbf{x}\right)=0$. Then if $H\left(\mathbf{x}\right)$ has all positive eigenvalues, $\mathbf{x}$ is a local minimum. If $H\left(\mathbf{x}\right)$ has all negative eigenvalues, then $\mathbf{x}$ is a local maximum. If $H\left(\mathbf{x}\right)$ has a positive eigenvalue, then there exists a direction in which f has a local minimum at $\mathbf{x}$, while if $H\left(\mathbf{x}\right)$ has a negative eigenvalue, there exists a direction in which $H\left(\mathbf{x}\right)$ has a local maximum at $\mathbf{x}$.*

Proof: Since $Df\left(\mathbf{x}\right)=0$, formula 9.18 holds and by continuity of the second derivative, $H\left(\mathbf{x}\right)$ is a symmetric matrix. Thus, by Corollary 9.5.8 $H\left(\mathbf{x}\right)$ has all real eigenvalues. Suppose first that $H\left(\mathbf{x}\right)$ has all positive eigenvalues and that all are larger than $\delta^2>0$. Then $H\left(\mathbf{x}\right)$ has an orthonormal basis of eigenvectors, $\{\mathbf{v}_i\}_{i=1}^n$ and if $\mathbf{u}$ is an arbitrary vector, $\mathbf{u}=\sum_{j=1}^n u_j\mathbf{v}_j$ where $u_j=\mathbf{u}\cdot\mathbf{v}_j$. Thus

$$\mathbf{u}^T H\left(\mathbf{x}\right)\mathbf{u}=\left(\sum_{k=1}^n u_k\mathbf{v}_k^T\right)H\left(\mathbf{x}\right)\left(\sum_{j=1}^n u_j\mathbf{v}_j\right)$$

$$=\sum_{j=1}^n u_j^2\lambda_j\geq\delta^2\sum_{j=1}^n u_j^2=\delta^2\left|\mathbf{u}\right|^2.$$

From 9.18 and the continuity of H, if $\mathbf{v}$ is small enough,

$$f\left(\mathbf{x}+\mathbf{v}\right)\geq f\left(\mathbf{x}\right)+\frac{1}{2}\delta^2\left|\mathbf{v}\right|^2-\frac{1}{4}\delta^2\left|\mathbf{v}\right|^2=f\left(\mathbf{x}\right)+\frac{\delta^2}{4}\left|\mathbf{v}\right|^2.$$

This shows the first claim of the theorem. The second claim follows from similar reasoning. Suppose $H\left(\mathbf{x}\right)$ has a positive eigenvalue λ^2. Then let $\mathbf{v}$ be an eigenvector for this eigenvalue. From 9.18,

$$f\left(\mathbf{x}+t\mathbf{v}\right)=f\left(\mathbf{x}\right)+\frac{1}{2}t^2\mathbf{v}^T H\left(\mathbf{x}\right)\mathbf{v}+$$

$$\frac{1}{2}t^2\left(\mathbf{v}^T\left(H\left(\mathbf{x}+t\mathbf{v}\right)-H\left(\mathbf{x}\right)\right)\mathbf{v}\right)$$

which implies

$$\begin{aligned} f\left(\mathbf{x}+t\mathbf{v}\right) &= f\left(\mathbf{x}\right)+\frac{1}{2}t^2\lambda^2\left|\mathbf{v}\right|^2+\frac{1}{2}t^2\left(\mathbf{v}^T\left(H\left(\mathbf{x}+t\mathbf{v}\right)-H\left(\mathbf{x}\right)\right)\mathbf{v}\right)\\ &\geq f\left(\mathbf{x}\right)+\frac{1}{4}t^2\lambda^2\left|\mathbf{v}\right|^2 \end{aligned}$$

whenever t 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.

This theorem is an analogue of the second derivative test for higher dimensions. As in one dimension, when there is a zero eigenvalue, it may be impossible to determine from the Hessian matrix what the local qualitative behavior of the function is. For example, consider

$$f_1\left(x,y\right)=x^4+y^2,\ f_2\left(x,y\right)=-x^4+y^2.$$

Then $Df_i(0,0) = \mathbf{0}$ and for both functions, the Hessian matrix evaluated at $(0,0)$ equals

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

but the behavior of the two functions is very different near the origin. The second has a saddle point while the first has a minimum there.

9.8 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 9.8.1 *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.

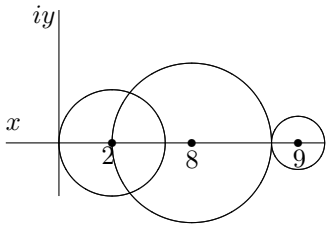
Example 9.8.2 *Here is a matrix. Estimate its eigenvalues.*

$$\begin{pmatrix} 2 & 1 & 1 \\ 3 & 5 & 0 \\ 0 & 1 & 9 \end{pmatrix}$$

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. In this case, 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. 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.

9.9 Advanced Theorems

More can be said but this requires some theory from complex variables[1]. The following is a fundamental theorem about counting zeros.

Theorem 9.9.1 *Let U be a region and let $\gamma : [a,b] \to U$ be closed, continuous, bounded variation, and the winding number, $n(\gamma, z) = 0$ for all $z \notin U$. Suppose also that f is analytic on U having zeros $a_1, \cdots, a_m$ where the zeros are repeated according to multiplicity, and suppose that none of these zeros are on $\gamma([a,b])$. Then*

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}dz = \sum_{k=1}^m n(\gamma, a_k).$$

Proof: It is given that $f(z) = \prod_{j=1}^m (z - a_j)\, g(z)$ where $g(z) \neq 0$ on U. Hence using the product rule,

$$\frac{f'(z)}{f(z)} = \sum_{j=1}^m \frac{1}{z - a_j} + \frac{g'(z)}{g(z)}$$

where $\frac{g'(z)}{g(z)}$ is analytic on U and so

$$\begin{aligned}\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}dz &= \sum_{j=1}^m n(\gamma, a_j) + \frac{1}{2\pi i}\int_\gamma \frac{g'(z)}{g(z)}dz \\ &= \sum_{j=1}^m n(\gamma, a_j).\end{aligned}$$

Therefore, this proves the theorem.

[1] If you haven't studied the theory of a complex variable, you should skip this section because you won't understand any of it.

Now let A be an $n \times n$ matrix. Recall that the eigenvalues of A are given by the zeros of the polynomial, $p_A(z) = \det(zI - A)$ where I is the $n \times n$ identity. You can argue that small changes in A will produce small changes in $p_A(z)$ and $p'_A(z)$. Let γ_k denote a very small closed circle which winds around z_k, one of the eigenvalues of A, in the counter clockwise direction so that $n(\gamma_k, z_k) = 1$. This circle is to enclose only z_k and is to have no other eigenvalue on it. Then apply Theorem 9.9.1. According to this theorem

$$\frac{1}{2\pi i}\int_\gamma \frac{p'_A(z)}{p_A(z)}dz$$

is always an integer equal to the multiplicity of z_k as a root of $p_A(t)$. Therefore, small changes in A result in no change to the above contour integral because it must be an integer and small changes in A result in small changes in the integral. Therefore whenever B is close enough to A, the two matrices have the same number of zeros inside γ_k, the zeros being counted according to multiplicity. By making the radius of the small circle equal to ε where ε is less than the minimum distance between any two distinct eigenvalues of A, this shows that if B is close enough to A, every eigenvalue of B is closer than ε to some eigenvalue of A.

Theorem 9.9.2 *If λ is an eigenvalue of A, then if all the entries of B are close enough to the corresponding entries of A, some eigenvalue of B will be within ε of λ.*

Consider the situation that $A(t)$ is an $n \times n$ matrix and that $t \to A(t)$ is continuous for $t \in [0,1]$.

Lemma 9.9.3 *Let $\lambda(t) \in \sigma(A(t))$ for $t < 1$ and let $\Sigma_t = \cup_{s \geq t}\sigma(A(s))$. Also let K_t be the connected component of $\lambda(t)$ in Σ_t. Then there exists $\eta > 0$ such that $K_t \cap \sigma(A(s)) \neq \emptyset$ for all $s \in [t, t+\eta]$.*

Proof: Denote by $D(\lambda(t), \delta)$ the disc centered at $\lambda(t)$ having radius $\delta > 0$, with other occurrences of this notation being defined similarly. Thus

$$D(\lambda(t), \delta) \equiv \{z \in \mathbb{C} : |\lambda(t) - z| \leq \delta\}.$$

Suppose $\delta > 0$ is small enough that $\lambda(t)$ is the only element of $\sigma(A(t))$ contained in $D(\lambda(t), \delta)$ and that $p_{A(t)}$ has no zeroes on the boundary of this disc. Then by continuity, and the above discussion and theorem, there exists $\eta > 0, t + \eta < 1$, such that for $s \in [t, t+\eta]$, $p_{A(s)}$ also has no zeroes on the boundary of this disc and $A(s)$ has the same number of eigenvalues, counted according to multiplicity, in the disc as $A(t)$. Thus $\sigma(A(s)) \cap D(\lambda(t), \delta) \neq \emptyset$ for all $s \in [t, t+\eta]$. Now let

$$H = \bigcup_{s \in [t, t+\eta]} \sigma(A(s)) \cap D(\lambda(t), \delta).$$

It will be shown that H is connected. Suppose not. Then $H = P \cup Q$ where P, Q are separated and $\lambda(t) \in P$. Let $s_0 \equiv \inf\{s : \lambda(s) \in Q \text{ for some } \lambda(s) \in \sigma(A(s))\}$. There exists $\lambda(s_0) \in \sigma(A(s_0)) \cap D(\lambda(t), \delta)$. If $\lambda(s_0) \notin Q$, then from the above discussion there are $\lambda(s) \in \sigma(A(s)) \cap Q$ for $s > s_0$ arbitrarily close to $\lambda(s_0)$. Therefore, $\lambda(s_0) \in Q$ which shows that $s_0 > t$ because $\lambda(t)$ is the only element of $\sigma(A(t))$ in $D(\lambda(t), \delta)$ and $\lambda(t) \in P$. Now let $s_n \uparrow s_0$. Then $\lambda(s_n) \in P$ for any $\lambda(s_n) \in \sigma(A(s_n)) \cap D(\lambda(t), \delta)$ and also it follows from the above discussion that for some choice of $s_n \to s_0$, $\lambda(s_n) \to \lambda(s_0)$ which contradicts P and Q separated and nonempty. Since P is nonempty, this shows $Q = \emptyset$. Therefore, H is connected as claimed. But $K_t \supseteq H$ and so $K_t \cap \sigma(A(s)) \neq \emptyset$ for all $s \in [t, t+\eta]$. This proves the lemma.

Theorem 9.9.4 *Suppose $A(t)$ is an $n \times n$ matrix and that $t \to A(t)$ is continuous for $t \in [0,1]$. Let $\lambda(0) \in \sigma(A(0))$ and define $\Sigma \equiv \cup_{t\in[0,1]}\sigma(A(t))$. Let $K_{\lambda(0)} = K_0$ denote the connected component of $\lambda(0)$ in Σ. Then $K_0 \cap \sigma(A(t)) \neq \emptyset$ for all $t \in [0,1]$.*

Proof: Let $S \equiv \{t \in [0,1] : K_0 \cap \sigma(A(s)) \neq \emptyset \text{ for all } s \in [0,t]\}$. Then $0 \in S$. Let $t_0 = \sup(S)$. Say $\sigma(A(t_0)) = \lambda_1(t_0), \cdots, \lambda_r(t_0)$.

Claim: At least one of these is a limit point of K_0 and consequently must be in K_0 which shows that S has a last point. Why is this claim true? Let $s_n \uparrow t_0$ so $s_n \in S$. Now let the discs, $D(\lambda_i(t_0), \delta), i = 1, \cdots, r$ be disjoint with $p_{A(t_0)}$ having no zeroes on γ_i the boundary of $D(\lambda_i(t_0), \delta)$. Then for n large enough it follows from Theorem 9.9.1 and the discussion following it that $\sigma(A(s_n))$ is contained in $\cup_{i=1}^r D(\lambda_i(t_0), \delta)$. It follows that $K_0 \cap (\sigma(A(t_0)) + D(0,\delta)) \neq \emptyset$ for all δ small enough. This requires at least one of the $\lambda_i(t_0)$ to be in $\overline{K_0}$. Therefore, $t_0 \in S$ and S has a last point.

Now by Lemma 9.9.3, if $t_0 < 1$, then $K_0 \cup K_t$ would be a strictly larger connected set containing $\lambda(0)$. (The reason this would be strictly larger is that $K_0 \cap \sigma(A(s)) = \emptyset$ for some $s \in (t, t+\eta)$ while $K_t \cap \sigma(A(s)) \neq \emptyset$ for all $s \in [t, t+\eta]$.) Therefore, $t_0 = 1$ and this proves the theorem.

Corollary 9.9.5 *Suppose one of the Gerschgorin discs, D_i is disjoint from the union of the others. Then D_i contains an eigenvalue of A. Also, if there are n disjoint Gerschgorin discs, then each one contains an eigenvalue of A.*

Proof: Denote by $A(t)$ the matrix $\left(a_{ij}^t\right)$ where if $i \neq j$, $a_{ij}^t = ta_{ij}$ and $a_{ii}^t = a_{ii}$. Thus to get $A(t)$ multiply all non diagonal terms by t. Let $t \in [0,1]$. Then $A(0) = \text{diag}(a_{11}, \cdots, a_{nn})$ and $A(1) = A$. Furthermore, the map, $t \to A(t)$ is continuous. Denote by D_j^t the Gerschgorin disc obtained from the j^{th} row for the matrix, $A(t)$. Then it is clear that $D_j^t \subseteq D_j$ the j^{th} Gerschgorin disc for A. It follows a_{ii} is the eigenvalue for $A(0)$ which is contained in the disc, consisting of the single point a_{ii} which is contained in D_i. Letting K be the connected component in Σ for Σ defined in Theorem 9.9.4 which is determined by a_{ii}, Gerschgorin's theorem implies that $K \cap \sigma(A(t)) \subseteq \cup_{j=1}^n D_j^t \subseteq \cup_{j=1}^n D_j = D_i \cup (\cup_{j\neq i} D_j)$ and also, since K is connected, there are not points of K in both D_i and $(\cup_{j\neq i} D_j)$. Since at least one point of K is in D_i,(a_{ii}), it follows all of K must be contained in D_i. Now by Theorem 9.9.4 this shows there are points of $K \cap \sigma(A)$ in D_i. The last assertion follows immediately.

This can be improved even more. This involves the following lemma.

Lemma 9.9.6 *In the situation of Theorem 9.9.4 suppose $\lambda(0) = K_0 \cap \sigma(A(0))$ and that $\lambda(0)$ is a simple root of the characteristic equation of $A(0)$. Then for all $t \in [0,1]$,*

$$\sigma(A(t)) \cap K_0 = \lambda(t)$$

where $\lambda(t)$ is a simple root of the characteristic equation of $A(t)$.

Proof: Let $S \equiv \{t \in [0,1] : K_0 \cap \sigma(A(s)) = \lambda(s), \text{ a simple eigenvalue for all } s \in [0,t]\}$. Then $0 \in S$ so it is nonempty. Let $t_0 = \sup(S)$ and suppose $\lambda_1 \neq \lambda_2$ are two elements of $\sigma(A(t_0)) \cap K_0$. Then choosing $\eta > 0$ small enough, and letting D_i be disjoint discs containing λ_i respectively, similar arguments to those of Lemma 9.9.3 can be used to conclude

$$H_i \equiv \cup_{s\in[t_0-\eta,t_0]} \sigma(A(s)) \cap D_i$$

is a connected and nonempty set for $i = 1, 2$ which would require that $H_i \subseteq K_0$. But then there would be two different eigenvalues of $A(s)$ contained in K_0, contrary to the definition of t_0. Therefore, there is at most one eigenvalue, $\lambda(t_0) \in K_0 \cap \sigma(A(t_0))$. Could it be a repeated root of the characteristic equation? Suppose $\lambda(t_0)$ is a repeated root of

the characteristic equation. As before, choose a small disc, D centered at $\lambda(t_0)$ and η small enough that

$$H \equiv \cup_{s\in[t_0-\eta,t_0]}\sigma(A(s)) \cap D$$

is a nonempty connected set containing either multiple eigenvalues of $A(s)$ or else a single repeated root to the characteristic equation of $A(s)$. But since H is connected and contains $\lambda(t_0)$ it must be contained in K_0 which contradicts the condition for $s \in S$ for all these $s \in [t_0-\eta, t_0]$. Therefore, $t_0 \in S$ as hoped. If $t_0 < 1$, there exists a small disc centered at $\lambda(t_0)$ and $\eta > 0$ such that for all $s \in [t_0, t_0+\eta]$, $A(s)$ has only simple eigenvalues in D and the only eigenvalues of $A(s)$ which could be in K_0 are in D. (This last assertion follows from noting that $\lambda(t_0)$ is the only eigenvalue of $A(t_0)$ in K_0 and so the others are at a positive distance from K_0. For s close enough to t_0, the eigenvalues of $A(s)$ are either close to these eigenvalues of $A(t_0)$ at a positive distance from K_0 or they are close to the eigenvalue, $\lambda(t_0)$ in which case it can be assumed they are in D.) But this shows that t_0 is not really an upper bound to S. Therefore, $t_0 = 1$ and the lemma is proved.

With this lemma, the conclusion of the above corollary can be sharpened.

Corollary 9.9.7 *Suppose one of the Gerschgorin discs, D_i is disjoint from the union of the others. Then D_i contains exactly one eigenvalue of A and this eigenvalue is a simple root to the characteristic polynomial of A.*

Proof: In the proof of Corollary 9.9.5, note that a_{ii} is a simple root of $A(0)$ since otherwise the i^{th} Gerschgorin disc would not be disjoint from the others. Also, K, the connected component determined by a_{ii} must be contained in D_i because it is connected and by Gerschgorin's theorem above, $K \cap \sigma(A(t))$ must be contained in the union of the Gerschgorin discs. Since all the other eigenvalues of $A(0)$, the a_{jj}, are outside D_i, it follows that $K \cap \sigma(A(0)) = a_{ii}$. Therefore, by Lemma 9.9.6, $K \cap \sigma(A(1)) = K \cap \sigma(A)$ consists of a single simple eigenvalue. This proves the corollary.

Example 9.9.8 *Consider the matrix,*

$$\begin{pmatrix} 5 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

The Gerschgorin discs are $D(5,1), D(1,2)$, and $D(0,1)$. Observe $D(5,1)$ is disjoint from the other discs. Therefore, there should be an eigenvalue in $D(5,1)$. The actual eigenvalues are not easy to find. They are the roots of the characteristic equation, $t^3 - 6t^2 + 3t + 5 = 0$. The numerical values of these are $-.669\,66, 1.423\,1$, and $5.246\,55$, verifying the predictions of Gerschgorin's theorem.

Vector Spaces

10.1 Vector Space Axioms

It is time to consider the idea of a Vector space.

Definition 10.1.1 *A vector space is an Abelian group of "vectors" satisfying the axioms of an Abelian group,*

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

the commutative law of addition,

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

the associative law for addition,

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

the existence of an additive identity,

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

the existence of an additive inverse, along with a field of "scalars", $\mathbb{F}$ which are allowed to multiply the vectors according to the following rules. (The Greek letters denote scalars.)

$$\alpha(\mathbf{v}+\mathbf{w})=\alpha\mathbf{v}+\alpha\mathbf{w}, \tag{10.1}$$

$$(\alpha+\beta)\mathbf{v}=\alpha\mathbf{v}+\beta\mathbf{v}, \tag{10.2}$$

$$\alpha(\beta\mathbf{v})=\alpha\beta(\mathbf{v}), \tag{10.3}$$

$$1\mathbf{v}=\mathbf{v}. \tag{10.4}$$

The field of scalars is usually $\mathbb{R}$ or $\mathbb{C}$ and the vector space will be called real or complex depending on whether the field is $\mathbb{R}$ or $\mathbb{C}$. However, other fields are also possible. For example, one could use the field of rational numbers or even the field of the integers mod p for p a prime. A vector space is also called a linear space.

For example, $\mathbb{R}^n$ with the usual conventions is an example of a real vector space and $\mathbb{C}^n$ is an example of a complex vector space. Up to now, the discussion has been for $\mathbb{R}^n$ or $\mathbb{C}^n$ and all that is taking place is an increase in generality and abstraction.

10.2 Subspaces And Bases

10.2.1 Basic Definitions

Definition 10.2.1 *If $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\} \subseteq V$, a vector space, then*

$$\text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_n) \equiv \left\{ \sum_{i=1}^{n} \alpha_i \mathbf{v}_i : \alpha_i \in \mathbb{F} \right\}.$$

A subset, $W \subseteq V$ is said to be a subspace if it is also a vector space with the same field of scalars. Thus $W \subseteq V$ is a subspace if $ax + by \in W$ whenever $a, b \in \mathbb{F}$ and $x, y \in W$. The span of a set of vectors as just described is an example of a subspace.

Definition 10.2.2 *If $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\} \subseteq V$, the set of vectors is linearly independent if*

$$\sum_{i=1}^{n} \alpha_i \mathbf{v}_i = \mathbf{0}$$

implies

$$\alpha_1 = \cdots = \alpha_n = 0$$

and $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is called a basis for V if

$$\text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_n) = V$$

and $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is linearly independent. The set of vectors is linearly dependent if it is not linearly independent.

10.2.2 A Fundamental Theorem

The next theorem is called the exchange theorem. It is very important that you understand this theorem. It is so important that I have given three proofs of it. The first two proofs amount to the same thing but are worded slightly differently.

Theorem 10.2.3 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ be a linearly independent set of vectors such that each $\mathbf{x}_i$ is in the span$\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$. Then $r \leq s$.*

Proof: Define span$\{\mathbf{y}_1, \cdots, \mathbf{y}_s\} \equiv V$, it follows there exist scalars, $c_1, \cdots, c_s$ such that

$$\mathbf{x}_1 = \sum_{i=1}^{s} c_i \mathbf{y}_i. \tag{10.5}$$

Not all of these scalars can equal zero because if this were the case, it would follow that $\mathbf{x}_1 = 0$ and so $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ would not be linearly independent. Indeed, if $\mathbf{x}_1 = 0$, $1\mathbf{x}_1 + \sum_{i=2}^{r} 0\mathbf{x}_i = \mathbf{x}_1 = 0$ and so there would exist a nontrivial linear combination of the vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ which equals zero.

Say $c_k \neq 0$. Then solve (10.5) for $\mathbf{y}_k$ and obtain

$$\mathbf{y}_k \in \text{span}\left(\mathbf{x}_1, \overbrace{\mathbf{y}_1, \cdots, \mathbf{y}_{k-1}, \mathbf{y}_{k+1}, \cdots, \mathbf{y}_s}^{\text{s-1 vectors here}} \right).$$

Define $\{\mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$ by

$$\{\mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\} \equiv \{\mathbf{y}_1, \cdots, \mathbf{y}_{k-1}, \mathbf{y}_{k+1}, \cdots, \mathbf{y}_s\}$$

Therefore, span $\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\} = V$ because if $\mathbf{v} \in V$, there exist constants $c_1, \cdots, c_s$ such that

$$\mathbf{v} = \sum_{i=1}^{s-1} c_i \mathbf{z}_i + c_s \mathbf{y}_k.$$

Now replace the $\mathbf{y}_k$ in the above with a linear combination of the vectors, $\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$ to obtain $\mathbf{v} \in \text{span}\{\mathbf{x}_1, \mathbf{z}_1, \cdots, \mathbf{z}_{s-1}\}$. The vector $\mathbf{y}_k$, in the list $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$, has now been replaced with the vector $\mathbf{x}_1$ and the resulting modified list of vectors has the same span as the original list of vectors, $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$.

Now suppose that $r > s$ and that span $\{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{z}_1, \cdots, \mathbf{z}_p\} = V$ where the vectors, $\mathbf{z}_1, \cdots, \mathbf{z}_p$ are each taken from the set, $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$ and $l + p = s$. This has now been done for $l = 1$ above. Then since $r > s$, it follows that $l \leq s < r$ and so $l + 1 \leq r$. Therefore, $\mathbf{x}_{l+1}$ is a vector not in the list, $\{\mathbf{x}_1, \cdots, \mathbf{x}_l\}$ and since span $\{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{z}_1, \cdots, \mathbf{z}_p\} = V$ there exist scalars, c_i and d_j such that

$$\mathbf{x}_{l+1} = \sum_{i=1}^{l} c_i \mathbf{x}_i + \sum_{j=1}^{p} d_j \mathbf{z}_j. \tag{10.6}$$

Now not all the d_j can equal zero because if this were so, it would follow that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ would be a linearly dependent set because one of the vectors would equal a linear combination of the others. Therefore, (10.6) can be solved for one of the $\mathbf{z}_i$, say $\mathbf{z}_k$, in terms of $\mathbf{x}_{l+1}$ and the other $\mathbf{z}_i$ and just as in the above argument, replace that $\mathbf{z}_i$ with $\mathbf{x}_{l+1}$ to obtain

$$\text{span}\left(\mathbf{x}_1, \cdots \mathbf{x}_l, \mathbf{x}_{l+1}, \overbrace{\mathbf{z}_1, \cdots \mathbf{z}_{k-1}, \mathbf{z}_{k+1}, \cdots, \mathbf{z}_p}^{\text{p-1 vectors here}}\right) = V.$$

Continue this way, eventually obtaining

$$\text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_s) = V.$$

But then $\mathbf{x}_r \in \text{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_s\}$ contrary to the assumption that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is linearly independent. Therefore, $r \leq s$ as claimed.

Theorem 10.2.4 *If*

$$\text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_r) \subseteq \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_s)$$

and $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ *are linearly independent, then* $r \leq s$.

Proof: Let $V \equiv \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_s)$ and suppose $r > s$. Let $A_l \equiv \{\mathbf{u}_1, \cdots, \mathbf{u}_l\}, A_0 = \emptyset$, and let B_{s-l} denote a subset of the vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_s\}$ which contains $s - l$ vectors and has the property that span $(A_l, B_{s-l}) = V$. Note that the assumption of the theorem says span $(A_0, B_s) = V$.

Now an exchange operation is given for span $(A_l, B_{s-l}) = V$. Since $r > s$, it follows $l < r$. Letting

$$B_{s-l} \equiv \{\mathbf{z}_1, \cdots, \mathbf{z}_{s-l}\} \subseteq \{\mathbf{v}_1, \cdots, \mathbf{v}_s\},$$

it follows there exist constants, c_i and d_i such that

$$\mathbf{u}_{l+1} = \sum_{i=1}^{l} c_i \mathbf{u}_i + \sum_{i=1}^{s-l} d_i \mathbf{z}_i,$$

and not all the d_i can equal zero. (If they were all equal to zero, it would follow that the set, $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ would be dependent since one of the vectors in it would be a linear combination of the others.)

Let $d_k \neq 0$. Then $\mathbf{z}_k$ can be solved for as follows.

$$\mathbf{z}_k = \frac{1}{d_k}\mathbf{u}_{l+1} - \sum_{i=1}^{l} \frac{c_i}{d_k}\mathbf{u}_i - \sum_{i \neq k} \frac{d_i}{d_k}\mathbf{z}_i.$$

This implies $V = \text{span}\,(A_{l+1}, B_{s-l-1})$, where $B_{s-l-1} \equiv B_{s-l} \setminus \{\mathbf{z}_k\}$, a set obtained by deleting $\mathbf{z}_k$ from B_{k-l}. You see, the process exchanged a vector in B_{s-l} with one from $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ and kept the span the same. Starting with $V = \text{span}\,(A_0, B_s)$, do the exchange operation until $V = \text{span}\,(A_{s-1}, \mathbf{z})$ where $\mathbf{z} \in \{\mathbf{v}_1, \cdots, \mathbf{v}_s\}$. Then one more application of the exchange operation yields $V = \text{span}\,(A_s)$. But this implies $\mathbf{u}_r \in \text{span}\,(A_s) = \text{span}\,(\mathbf{u}_1, \cdots, \mathbf{u}_s)$, contradicting the linear independence of $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$. It follows that $r \leq s$ as claimed.

Here is yet another proof in case you didn't like either of the last two.

Theorem 10.2.5 *If*

$$\text{span}\,(\mathbf{u}_1, \cdots, \mathbf{u}_r) \subseteq \text{span}\,(\mathbf{v}_1, \cdots, \mathbf{v}_s)$$

and $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ *are linearly independent, then* $r \leq s$.

Proof: Suppose $r > s$. Since each $\mathbf{u}_k \in \text{span}\,(\mathbf{v}_1, \cdots, \mathbf{v}_s)$, it follows

$$\text{span}\,(\mathbf{u}_1, \cdots, \mathbf{u}_s, \mathbf{v}_1, \cdots, \mathbf{v}_s) = \text{span}\,(\mathbf{v}_1, \cdots, \mathbf{v}_s).$$

Let $\{\mathbf{v}_{k_1}, \cdots, \mathbf{v}_{k_j}\}$ denote a subset of the set $\{\mathbf{v}_1 \cdots, \mathbf{v}_s\}$ which has j elements in it. If $j = 0$, this means no vectors from $\{\mathbf{v}_1 \cdots, \mathbf{v}_s\}$ are included. Let j be the smallest nonnegative integer such that

$$\text{span}\left(\mathbf{u}_1, \cdots, \mathbf{u}_s, \mathbf{v}_{k_1}, \cdots, \mathbf{v}_{k_j}\right) = \text{span}\,(\mathbf{v}_1, \cdots, \mathbf{v}_s) \tag{10.7}$$

Claim: $j = 0$.

Proof of claim: Suppose $j \geq 1$. Then since $s < r$, there exist scalars, a_k and b_i such that

$$\mathbf{u}_{s+1} = \sum_{k=1}^{s} a_k \mathbf{u}_k + \sum_{i=1}^{j} b_i \mathbf{v}_{k_i}.$$

By linear independence of the $\mathbf{u}_k$, not all the b_i can equal zero. Therefore, one of the $\mathbf{v}_{k_i}$ is in the span of the other vectors in the above sum. Thus there exist $l_1, \cdots, l_{j-1}$ such that

$$\mathbf{v}_{k_i} \in \text{span}\left(\mathbf{u}_1, \cdots, \mathbf{u}_s, \mathbf{u}_{s+1}, \mathbf{v}_{l_1}, \cdots, \mathbf{v}_{l_{j-1}}\right)$$

and so from 10.7,

$$\text{span}\left(\mathbf{u}_1, \cdots, \mathbf{u}_s, \mathbf{u}_{s+1}, \mathbf{v}_{l_1}, \cdots, \mathbf{v}_{l_{j-1}}\right) = \text{span}\,(\mathbf{v}_1, \cdots, \mathbf{v}_s)$$

contrary to the definition of j. Therefore, $j = 0$ and this proves the claim.

It follows from the claim that $\text{span}\,(\mathbf{u}_1, \cdots, \mathbf{u}_s) = \text{span}\,(\mathbf{v}_1, \cdots, \mathbf{v}_s)$ which implies

$$\mathbf{u}_{s+1} \in \text{span}\,(\mathbf{u}_1, \cdots, \mathbf{u}_s)$$

contrary to the assumption the $\mathbf{u}_k$ are linearly independent. Therefore, $r \leq s$ as claimed.

Corollary 10.2.6 *If* $\{\mathbf{u}_1, \cdots, \mathbf{u}_m\}$ *and* $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ *are two bases for* V, *then* $m = n$.

Proof: By Theorem 10.2.4 or Theorem 10.2.5, $m \leq n$ and $n \leq m$.

Definition 10.2.7 *A vector space V is of dimension n if it has a basis consisting of n vectors. This is well defined thanks to Corollary 10.2.6. It is always assumed here that $n < \infty$ in this case, such a vector space is said to be finite dimensional.*

Theorem 10.2.8 *If $V = \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_n)$ then some subset of $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ is a basis for V. Also, if $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\} \subseteq V$ is linearly independent and the vector space is finite dimensional, then the set, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$, can be enlarged to obtain a basis of V.*

Proof: Let

$$S = \{E \subseteq \{\mathbf{u}_1, \cdots, \mathbf{u}_n\} \text{ such that } \text{span}(E) = V\}.$$

For $E \in S$, let $|E|$ denote the number of elements of E. Let

$$m \equiv \min\{|E| \text{ such that } E \in S\}.$$

Thus there exist vectors

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_m\} \subseteq \{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$$

such that

$$\text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_m) = V$$

and m is as small as possible for this to happen. If this set is linearly independent, it follows it is a basis for V and the theorem is proved. On the other hand, if the set is not linearly independent, then there exist scalars,

$$c_1, \cdots, c_m$$

such that

$$\mathbf{0} = \sum_{i=1}^{m} c_i \mathbf{v}_i$$

and not all the c_i are equal to zero. Suppose $c_k \neq 0$. Then the vector, $\mathbf{v}_k$ may be solved for in terms of the other vectors. Consequently,

$$V = \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_{k+1}, \cdots, \mathbf{v}_m)$$

contradicting the definition of m. This proves the first part of the theorem.

To obtain the second part, begin with $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ and suppose a basis for V is $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$. If

$$\text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k) = V,$$

then $k = n$. If not, there exists a vector,

$$\mathbf{u}_{k+1} \notin \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k).$$

Then $\{\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1}\}$ is also linearly independent. Continue adding vectors in this way until n linearly independent vectors have been obtained. Then $\text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_n) = V$ because if it did not do so, there would exist $\mathbf{u}_{n+1}$ as just described and $\{\mathbf{u}_1, \cdots, \mathbf{u}_{n+1}\}$ would be a linearly independent set of vectors having $n+1$ elements even though $\{\mathbf{v}_1, \cdots, v_n\}$ is a basis. This would contradict Theorems 10.2.4 and 10.2.3. Therefore, this list is a basis and this proves the theorem.

It is useful to emphasize some of the ideas used in the above proof.

Lemma 10.2.9 *Suppose $\mathbf{v} \notin \text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ and $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ is linearly independent. Then $\{\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{v}\}$ is also linearly independent.*

Proof: Suppose $\sum_{i=1}^{k} c_i \mathbf{u}_i + d\mathbf{v} = 0$. It is required to verify that each $c_i = 0$ and that $d = 0$. But if $d \neq 0$, then you can solve for $\mathbf{v}$ as a linear combination of the vectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$,

$$\mathbf{v} = -\sum_{i=1}^{k} \left(\frac{c_i}{d}\right) \mathbf{u}_i$$

contrary to assumption. Therefore, $d = 0$. But then $\sum_{i=1}^{k} c_i \mathbf{u}_i = 0$ and the linear independence of $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ implies each $c_i = 0$ also. This proves the lemma.

10.2.3 The Basis Of A Subspace

Theorem 10.2.10 *Let V be a nonzero subspace of a finite dimentional vector space, W of dimension, n. Then V has a basis with no more than n vectors.*

Proof: Let $\mathbf{v}_1 \in V$ where $\mathbf{v}_1 \neq 0$. If $\operatorname{span}\{\mathbf{v}_1\} = V$, stop. $\{\mathbf{v}_1\}$ is a basis for V. Otherwise, there exists $\mathbf{v}_2 \in V$ which is not in $\operatorname{span}\{\mathbf{v}_1\}$. By Lemma 10.2.9 $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a linearly independent set of vectors. If $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\} = V$ stop, $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a basis for V. If $\operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\} \neq V$, then there exists $\mathbf{v}_3 \notin \operatorname{span}\{\mathbf{v}_1, \mathbf{v}_2\}$ and $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is a larger linearly independent set of vectors. Continuing this way, the process must stop before $n + 1$ steps because if not, it would be possible to obtain $n + 1$ linearly independent vectors contrary to the exchange theorem, Theorems 10.2.3 and 10.2.4. This proves the theorem.

10.3 Exercises

1. Determine which matrices are in row reduced echelon form.

   (a) $\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 7 \end{pmatrix}$

   (b) $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

   (c) $\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 5 \\ 0 & 0 & 1 & 2 & 0 & 4 \\ 0 & 0 & 0 & 0 & 1 & 3 \end{pmatrix}$

2. Row reduce the following matrices to obtain the row reduced echelon form. List the pivot columns in the original matrix.

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

   (b) $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & -2 \\ 3 & 0 & 0 \\ 3 & 2 & 1 \end{pmatrix}$

   (c) $\begin{pmatrix} 1 & 2 & 1 & 3 \\ -3 & 2 & 1 & 0 \\ 3 & 2 & 1 & 1 \end{pmatrix}$

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

   (b) $\begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 1 \\ 2 & 1 & 0 \\ 0 & 2 & 0 \end{pmatrix}$

   (c) $\begin{pmatrix} 0 & 1 & 0 & 2 & 1 & 2 & 2 \\ 0 & 3 & 2 & 12 & 1 & 6 & 8 \\ 0 & 1 & 1 & 5 & 0 & 2 & 3 \\ 0 & 2 & 1 & 7 & 0 & 3 & 4 \end{pmatrix}$

   (d) $\begin{pmatrix} 0 & 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 3 & 2 & 6 & 0 & 5 & 4 \\ 0 & 1 & 1 & 2 & 0 & 2 & 2 \\ 0 & 2 & 1 & 4 & 0 & 3 & 2 \end{pmatrix}$

   (e) $\begin{pmatrix} 0 & 1 & 0 & 2 & 1 & 1 & 2 \\ 0 & 3 & 2 & 6 & 1 & 5 & 1 \\ 0 & 1 & 1 & 2 & 0 & 2 & 1 \\ 0 & 2 & 1 & 4 & 0 & 3 & 1 \end{pmatrix}$

4. Suppose A is an $m \times n$ matrix. Explain why the rank of A is always no larger than $\min(m, n)$.

5. Let H denote span $\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \end{pmatrix} \right)$. Find the dimension of H and determine a basis.

6. Let H denote span $\left(\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right)$. Find the dimension of H and determine a basis.

7. Let H denote span $\left(\begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right)$. Find the dimension of H and determine a basis.

8. Let $M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 = u_1 = 0\}$. Is M a subspace? Explain.

9. Let $M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 \geq u_1\}$. Is M a subspace? Explain.

10. Let $\mathbf{w} \in \mathbb{R}^4$ and let $M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \mathbf{w} \cdot \mathbf{u} = 0\}$. Is M a subspace? Explain.

11. Let $M = \{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_i \geq 0 \text{ for each } i = 1, 2, 3, 4\}$. Is M a subspace? Explain.

12. Let $\mathbf{w}, \mathbf{w}_1$ be given vectors in $\mathbb{R}^4$ and define

$$M = \left\{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \mathbf{w} \cdot \mathbf{u} = 0 \text{ and } \mathbf{w}_1 \cdot \mathbf{u} = 0\right\}.$$

Is M a subspace? Explain.

13. Let $M = \left\{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : |u_1| \leq 4\right\}$. Is M a subspace? Explain.

14. Let $M = \left\{\mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \sin(u_1) = 1\right\}$. Is M a subspace? Explain.

15. Study the definition of span. Explain what is meant by the span of a set of vectors. Include pictures.

16. Suppose $\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ is a set of vectors from $\mathbb{F}^n$. Show that $\mathbf{0}$ is in $\text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k)$.

17. Study the definition of linear independence. Explain in your own words what is meant by linear independence and linear dependence. Illustrate with pictures.

18. Here are three vectors. Determine whether they are linearly independent or linearly dependent.

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

19. Here are three vectors. Determine whether they are linearly independent or linearly dependent.

$$\begin{pmatrix} 4 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}$$

20. Here are three vectors. Determine whether they are linearly independent or linearly dependent.

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 5 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}$$

21. Here are four vectors. Determine whether they span $\mathbb{R}^3$. Are these vectors linearly independent?

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 3 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$$

22. Here are four vectors. Determine whether they span $\mathbb{R}^3$. Are these vectors linearly independent?

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 3 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$$

23. Determine whether the following vectors are a basis for $\mathbb{R}^3$. If they are, explain why they are and if they are not, give a reason and tell whether they span $\mathbb{R}^3$.

$$\begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 3 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix}$$

24. Determine whether the following vectors are a basis for $\mathbb{R}^3$. If they are, explain why they are and if they are not, give a reason and tell whether they span $\mathbb{R}^3$.

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

25. Determine whether the following vectors are a basis for $\mathbb{R}^3$. If they are, explain why they are and if they are not, give a reason and tell whether they span $\mathbb{R}^3$.

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

26. Determine whether the following vectors are a basis for $\mathbb{R}^3$. If they are, explain why they are and if they are not, give a reason and tell whether they span $\mathbb{R}^3$.

$$\begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

27. Consider the vectors of the form

$$\left\{ \begin{pmatrix} 2t+3s \\ s-t \\ t+s \end{pmatrix} : s,t \in \mathbb{R} \right\}.$$

Is this set of vectors a subspace of $\mathbb{R}^3$? If so, explain why, give a basis for the subspace and find its dimension.

28. Consider the vectors of the form

$$\left\{ \begin{pmatrix} 2t+3s+u \\ s-t \\ t+s \\ u \end{pmatrix} : s,t,u \in \mathbb{R} \right\}.$$

Is this set of vectors a subspace of $\mathbb{R}^4$? If so, explain why, give a basis for the subspace and find its dimension.

29. Consider the vectors of the form

$$\left\{ \begin{pmatrix} 2t+u \\ t+3u \\ t+s+v \\ u \end{pmatrix} : s,t,u,v \in \mathbb{R} \right\}.$$

Is this set of vectors a subspace of $\mathbb{R}^4$? If so, explain why, give a basis for the subspace and find its dimension.

30. If you have 5 vectors in $\mathbb{F}^5$ and the vectors are linearly independent, can it always be concluded they span $\mathbb{F}^5$? Explain.

31. If you have 6 vectors in $\mathbb{F}^5$, is it possible they are linearly independent? Explain.

32. Suppose A is an $m \times n$ matrix and $\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}$ is a linearly independent set of vectors in $A(\mathbb{F}^n) \subseteq \mathbb{F}^m$. Now suppose $A(\mathbf{z}_i) = \mathbf{w}_i$. Show $\{\mathbf{z}_1, \cdots, \mathbf{z}_k\}$ is also independent.

33. Suppose V, W are subspaces of $\mathbb{F}^n$. Show $V \cap W$ defined to be all vectors which are in both V and W is a subspace also.

34. Suppose V and W both have dimension equal to 7 and they are subspaces of $\mathbb{F}^{10}$. What are the possibilities for the dimension of $V \cap W$? **Hint:** Remember that a linear independent set can be extended to form a basis.

35. Suppose V has dimension p and W has dimension q and they are each contained in a subspace, U which has dimension equal to n where $n > \max(p, q)$. What are the possibilities for the dimension of $V \cap W$? **Hint:** Remember that a linear independent set can be extended to form a basis.

36. If $\mathbf{b} \neq \mathbf{0}$, can the solution set of $A\mathbf{x} = \mathbf{b}$ be a plane through the origin? Explain.

37. Suppose a system of equations has fewer equations than variables and you have found a solution to this system of equations. Is it possible that your solution is the only one? Explain.

38. Suppose a system of linear equations has a 2×4 augmented matrix and the last column is a pivot column. Could the system of linear equations be consistent? Explain.

39. Suppose the coefficient matrix of a system of n equations with n variables has the property that every column is a pivot column. Does it follow that the system of equations must have a solution? If so, must the solution be unique? Explain.

40. Suppose there is a unique solution to a system of linear equations. What must be true of the pivot columns in the augmented matrix.

41. State whether each of the following sets of data are possible for the matrix equation $A\mathbf{x} = \mathbf{b}$. If possible, describe the solution set. That is, tell whether there exists a unique solution no solution or infinitely many solutions.

    (a) A is a 5×6 matrix, $\text{rank}(A) = 4$ and $\text{rank}(A|\mathbf{b}) = 4$. **Hint:** This says $\mathbf{b}$ is in the span of four of the columns. Thus the columns are not independent.

    (b) A is a 3×4 matrix, $\text{rank}(A) = 3$ and $\text{rank}(A|\mathbf{b}) = 2$.

    (c) A is a 4×2 matrix, $\text{rank}(A) = 4$ and $\text{rank}(A|\mathbf{b}) = 4$. **Hint:** This says $\mathbf{b}$ is in the span of the columns and the columns must be independent.

    (d) A is a 5×5 matrix, $\text{rank}(A) = 4$ and $\text{rank}(A|\mathbf{b}) = 5$. **Hint:** This says $\mathbf{b}$ is not in the span of the columns.

    (e) A is a 4×2 matrix, $\text{rank}(A) = 2$ and $\text{rank}(A|\mathbf{b}) = 2$.

42. Suppose A is an $m \times n$ matrix in which $m \leq n$. Suppose also that the rank of A equals m. Show that A maps $\mathbb{F}^n$ onto $\mathbb{F}^m$. **Hint:** The vectors $\mathbf{e}_1, \cdots, \mathbf{e}_m$ occur as columns in the row reduced echelon form for A.

43. Suppose A is an $m \times n$ matrix in which $m \geq n$. Suppose also that the rank of A equals n. Show that A is one to one. **Hint:** If not, there exists a vector, $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{0}$, and this implies at least one column of A is a linear combination of the others. Show this would require the column rank to be less than n.

44. Explain why an $n \times n$ matrix, A is both one to one and onto if and only if its rank is n.

45. Suppose A is an $m \times n$ matrix and B is an $n \times p$ matrix. Show that

$$\dim(\ker(AB)) \leq \dim(\ker(A)) + \dim(\ker(B)).$$

Hint: Consider the subspace, $B(\mathbb{F}^p) \cap \ker(A)$ and suppose a basis for this subspace is $\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}$. Now suppose $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ is a basis for $\ker(B)$. Let $\{\mathbf{z}_1, \cdots, \mathbf{z}_k\}$ be such that $B\mathbf{z}_i = \mathbf{w}_i$ and argue that

$$\ker(AB) \subseteq \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_r, \mathbf{z}_1, \cdots, \mathbf{z}_k).$$

Here is how you do this. Suppose $AB\mathbf{x} = \mathbf{0}$. Then $B\mathbf{x} \in \ker(A) \cap B(\mathbb{F}^p)$ and so $B\mathbf{x} = \sum_{i=1}^{k} B\mathbf{z}_i$ showing that

$$\mathbf{x} - \sum_{i=1}^{k} \mathbf{z}_i \in \ker(B).$$

46. Explain why $A\mathbf{x} = \mathbf{0}$ always has a solution even when A^{-1} does not exist.

    (a) What can you conclude about A if the solution is unique?

    (b) What can you conclude about A if the solution is not unique?

47. Suppose $\det(A - \lambda I) = 0$. Show using Theorem 6.1.17 there exists $\mathbf{x} \neq \mathbf{0}$ such that $(A - \lambda I)\mathbf{x} = \mathbf{0}$.

48. Let A be an $n \times n$ matrix and let $\mathbf{x}$ be a nonzero vector such that $A\mathbf{x} = \lambda\mathbf{x}$ for some scalar, λ. When this occurs, the vector, $\mathbf{x}$ is called an **eigenvector** and the scalar, λ is called an **eigenvalue**. It turns out that not every number is an eigenvalue. Only certain ones are. Why? **Hint:** Show that if $A\mathbf{x} = \lambda\mathbf{x}$, then $(A - \lambda I)\mathbf{x} = \mathbf{0}$. Explain why this shows that $(A - \lambda I)$ is not one to one and not onto. Now use Theorem 6.1.17 to argue $\det(A - \lambda I) = 0$. What sort of equation is this? How many solutions does it have?

49. 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 $A\mathbf{x} = \mathbf{0}$. Do this using the Fredholm alternative.

50. Let A be an $m \times n$ real matrix and let $\mathbf{b} \in \mathbb{R}^m$. Show there exists a solution, $\mathbf{x}$ to the system

$$A^T A\mathbf{x} = A^T\mathbf{b}$$

Next show that if $\mathbf{x}, \mathbf{x}_1$ are two solutions, then $A\mathbf{x} = A\mathbf{x}_1$. **Hint:** First show that $\left(A^T A\right)^T = A^T A$. Next show if $\mathbf{x} \in \ker\left(A^T A\right)$, then $A\mathbf{x} = \mathbf{0}$. Finally apply the Fredholm alternative. This will give existence of a solution.

51. Show that in the context of Problem 50 that if $\mathbf{x}$ is the solution there, then $|\mathbf{b} - A\mathbf{x}| \leq |\mathbf{b} - A\mathbf{y}|$ for every $\mathbf{y}$. Thus $A\mathbf{x}$ is the point of $A(\mathbb{R}^n)$ which is closest to $\mathbf{b}$ of every point in $A(\mathbb{R}^n)$.

Linear Transformations

11.1 Matrix Multiplication As A Linear Transformation

Definition 11.1.1 *Let V and W be two finite dimensional vector spaces. A function, L which maps V to W is called a linear transformation and $L \in \mathcal{L}(V,W)$ if for all scalars α and β, and vectors $\mathbf{v}, \mathbf{w}$,*

$$L(\alpha\mathbf{v}+\beta\mathbf{w}) = \alpha L(\mathbf{v}) + \beta L(\mathbf{w}).$$

An example of a linear transformation is familiar matrix multiplication. Let $A = (a_{ij})$ be an $m \times n$ matrix. Then an example of a linear transformation $L : \mathbb{F}^n \to \mathbb{F}^m$ is given by

$$(L\mathbf{v})_i \equiv \sum_{j=1}^{n} a_{ij} v_j.$$

Here

$$\mathbf{v} \equiv \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{F}^n.$$

11.2 $\mathcal{L}(V,W)$ As A Vector Space

Definition 11.2.1 *Given $L, M \in \mathcal{L}(V,W)$ define a new element of $\mathcal{L}(V,W)$, denoted by $L+M$ according to the rule*

$$(L+M)\mathbf{v} \equiv L\mathbf{v} + M\mathbf{v}.$$

For α a scalar and $L \in \mathcal{L}(V,W)$, define $\alpha L \in \mathcal{L}(V,W)$ by

$$\alpha L(\mathbf{v}) \equiv \alpha(L\mathbf{v}).$$

You should verify that all the axioms of a vector space hold for $\mathcal{L}(V,W)$ with the above definitions of vector addition and scalar multiplication. What about the dimension of $\mathcal{L}(V,W)$?

Theorem 11.2.2 *Let V and W be finite dimensional normed linear spaces of dimension n and m respectively Then* $\dim(\mathcal{L}(V,W)) = mn$.

Proof: Let the two sets of bases be

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_n\} \text{ and } \{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$$

for X and Y respectively. Let $E_{ik} \in \mathcal{L}(V,W)$ be the linear transformation defined on the basis, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$, by

$$E_{ik}\mathbf{v}_j \equiv \mathbf{w}_i \delta_{jk}$$

where $\delta_{ik} = 1$ if $i = k$ and 0 if $i \neq k$. Then let $L \in \mathcal{L}(V,W)$. Since $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ is a basis, there exist constants, d_{jk} such that

$$L\mathbf{v}_r = \sum_{j=1}^{m} d_{jr}\mathbf{w}_j$$

Also

$$\sum_{j=1}^{m}\sum_{k=1}^{n} d_{jk}E_{jk}(\mathbf{v}_r) = \sum_{j=1}^{m} d_{jr}\mathbf{w}_j.$$

It follows that

$$L = \sum_{j=1}^{m}\sum_{k=1}^{n} d_{jk}E_{jk}$$

because the two linear transformations agree on a basis. Since L is arbitrary this shows

$$\{E_{ik} : i = 1, \cdots, m,\ k = 1, \cdots, n\}$$

spans $\mathcal{L}(V,W)$.

If

$$\sum_{i,k} d_{ik}E_{ik} = \mathbf{0},$$

then

$$\mathbf{0} = \sum_{i,k} d_{ik}E_{ik}(\mathbf{v}_l) = \sum_{i=1}^{m} d_{il}\mathbf{w}_i$$

and so, since $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ is a basis, $d_{il} = 0$ for each $i = 1, \cdots, m$. Since l is arbitrary, this shows $d_{il} = 0$ for all i and l. Thus these linear transformations form a basis and this shows the dimension of $\mathcal{L}(V,W)$ is mn as claimed.

11.3 Eigenvalues And Eigenvectors Of Linear Transformations

Let V be a finite dimensional vector space. For example, it could be a subspace of $\mathbb{C}^n$. Also suppose $A \in \mathcal{L}(V,V)$. Does A have eigenvalues and eigenvectors just like the case where A is a $n \times n$ matrix?

Theorem 11.3.1 *Let V be a nonzero finite dimensional complex vector space of dimension n. Suppose also the field of scalars equals $\mathbb{C}$.*[1] *Suppose $A \in \mathcal{L}(V,V)$. Then there exists $v \neq 0$ and $\lambda \in \mathbb{C}$ such that*

$$Av = \lambda v.$$

[1] All that is really needed is that the minimal polynomial can be completely factored in the given field. The complex numbers have this property from the fundamental theorem of algebra.

Proof: Consider the linear transformations, $I, A, A^2, \cdots, A^{n^2}$. There are n^2+1 of these transformations and so by Theorem 11.2.2 the set is linearly dependent. Thus there exist constants, $c_i \in \mathbb{C}$ such that

$$c_0 I + \sum_{k=1}^{n^2} c_k A^k = 0.$$

This implies there exists a polynomial, $q(\lambda)$ which has the property that $q(A) = 0$. In fact, $q(\lambda) \equiv c_0 + \sum_{k=1}^{n^2} c_k \lambda^k$. Dividing by the leading term, it can be assumed this polynomial is of the form $\lambda^m + c_{m-1}\lambda^{m-1} + \cdots + c_1\lambda + c_0$, a monic polynomial. Now consider all such monic polynomials, q such that $q(A) = 0$ and pick one which has the smallest degree. This is called the minimial polynomial and will be denoted here by $p(\lambda)$. By the fundamental theorem of algebra, $p(\lambda)$ is of the form

$$p(\lambda) = \prod_{k=1}^{p} (\lambda - \lambda_k).$$

Thus, since p has minimial degree,

$$\prod_{k=1}^{p} (A - \lambda_k I) = 0, \text{ but } \prod_{k=1}^{p-1} (A - \lambda_k I) \neq 0.$$

Therefore, there exists $u \neq 0$ such that

$$v \equiv \left(\prod_{k=1}^{p-1} (A - \lambda_k I) \right)(u) \neq 0.$$

But then

$$(A - \lambda_p I)\, v = (A - \lambda_p I) \left(\prod_{k=1}^{p-1} (A - \lambda_k I) \right)(u) = 0.$$

This proves the theorem.

Corollary 11.3.2 *In the above theorem, each of the scalars, λ_k has the property that there exists a nonzero v such that $(A - \lambda_i I)\, v = 0$. Furthermore the λ_i are the only scalars with this property.*

Proof: For the first claim, just factor out $(A - \lambda_i I)$ instead of $(A - \lambda_p I)$. Next suppose $(A - \mu I)\, v = 0$ for some μ and $v \neq 0$. Then

$$\begin{aligned}
0 &= \prod_{k=1}^{p} (A - \lambda_k I)\, v = \prod_{k=1}^{p-1} (A - \lambda_k I)\, (Av - \lambda_p v) \\
&= (\mu - \lambda_p) \left(\prod_{k=1}^{p-1} (A - \lambda_k I) \right) v \\
&= (\mu - \lambda_p) \left(\prod_{k=1}^{p-2} (A - \lambda_k I) \right) (Av - \lambda_{p-1} v) \\
&= (\mu - \lambda_p)\, (\mu - \lambda_{p-1}) \left(\prod_{k=1}^{p-2} (A - \lambda_k I) \right)
\end{aligned}$$

continuing this way yields

$$= \prod_{k=1}^{p} (\mu - \lambda_k)\, v,$$

a contradiction unless $\mu = \lambda_k$ for some k.

Therefore, these are eigenvectors and eigenvalues with the usual meaning and the λ_k are all of the eigenvalues.

Definition 11.3.3 *For $A \in \mathcal{L}(V,V)$ where $\dim(V) = n$, the scalars, λ_k in the minimal polynomial,*

$$p(\lambda) = \prod_{k=1}^{p} (\lambda - \lambda_k)$$

are called the eigenvalues of A. The collection of eigenvalues of A is denoted by $\sigma(A)$. For λ an eigenvalue of $A \in \mathcal{L}(V,V)$, the generalized eigenspace is defined as

$$V_\lambda \equiv \{x \in V : (A - \lambda I)^m x = 0 \text{ for some } m \in \mathbb{N}\}$$

and the eigenspace is defined as

$$\{x \in V : (A - \lambda I)\, x = 0\} \equiv \ker(A - \lambda I).$$

Also, for subspaces of V, $V_1, V_2, \cdots, V_r$, the symbol, $V_1 + V_2 + \cdots + V_r$ or the shortened version, $\sum_{i=1}^{r} V_i$ will denote the set of all vectors of the form $\sum_{i=1}^{r} v_i$ where $v_i \in V_i$.

Lemma 11.3.4 *The generalized eigenspace for $\lambda \in \sigma(A)$ where $A \in \mathcal{L}(V,V)$ for V an n dimensional vector space is a subspace, V_λ of V satisfying*

$$A : V_\lambda \to V_\lambda,$$

and there exists a smallest integer, m with the property that

$$\ker(A - \lambda I)^m = \left\{x \in V : (A - \lambda I)^k x = 0 \text{ for some } k \in \mathbb{N}\right\}. \tag{11.1}$$

Proof: The claim that the generalized eigenspace is a subspace is obvious. To establish the second part, note that

$$\left\{\ker(A - \lambda I)^k\right\}$$

yields an increasing sequence of subspaces. Eventually

$$\dim(\ker(A - \lambda I)^m) = \dim\left(\ker(A - \lambda I)^{m+1}\right)$$

and so $\ker(A - \lambda I)^m = \ker(A - \lambda I)^{m+1}$. Now if $\mathbf{x} \in \ker(A - \lambda I)^{m+2}$, then

$$(A - \lambda I)\,\mathbf{x} \in \ker(A - \lambda I)^{m+1} = \ker(A - \lambda I)^m$$

and so there exists $\mathbf{y} \in \ker(A - \lambda I)^m$ such that $(A - \lambda I)\,\mathbf{x} = \mathbf{y}$ and consequently

$$(A - \lambda I)^{m+1}\,\mathbf{x} = (A - \lambda I)^m\,\mathbf{y} = \mathbf{0}$$

showing that $\mathbf{x} \in \ker(A - \lambda I)^{m+1}$. Therefore, continuing this way, it follows that for all $k \in \mathbb{N}$,

$$\ker(A - \lambda I)^m = \ker(A - \lambda I)^{m+k}.$$

Therefore, this shows 11.1.

The following theorem is of major importance and will be the basis for the very important theorems concerning block diagonal matrices.

The following theorem is of major importance and will be the basis for the very important theorems concerning block diagonal matrices.

Theorem 11.3.5 *Let V be a complex vector space of dimension n and suppose $\sigma(A) = \{\lambda_1, \cdots, \lambda_k\}$ where the λ_i are the distinct eigenvalues of A. Denote by V_i the generalized eigenspace for λ_i and let r_i be the multiplicity of λ_i. By this is meant that*

$$V_i = \ker (A - \lambda_i I)^{r_i} \tag{11.2}$$

and r_i is the smallest integer with this property. Then

$$V = \sum_{i=1}^{k} V_i. \tag{11.3}$$

Proof: This is proved by induction on k. First suppose there is only one eigenvalue, λ_1 of multiplicity m. Then by the definition of eigenvalues given in Definition 11.3.3, A satisfies an equation of the form

$$(A - \lambda_1 I)^r = 0$$

where r is as small as possible for this to take place. Thus $\ker (A - \lambda_1 I)^r = V$ and the theorem is proved in the case of one eigenvalue.

Now suppose the theorem is true for any $i \leq k-1$ where $k \geq 2$ and suppose $\sigma(A) = \{\lambda_1, \cdots, \lambda_k\}$.

Claim 1: Let $\mu \neq \lambda_i$, Then $(A - \mu I)^m : V_i \to V_i$ and is one to one and onto for every $m \in \mathbb{N}$.

Proof: It is clear that $(A - \mu I)^m$ maps V_i to V_i because if $v \in V_i$ then $(A - \lambda_i I)^k v = 0$ for some $k \in \mathbb{N}$. Consequently,

$$(A - \lambda_i I)^k (A - \mu I)^m v = (A - \mu I)^m (A - \lambda_i I)^k v = (A - \mu I)^m 0 = 0$$

which shows that $(A - \mu I)^m v \in V_i$.

It remains to verify that $(A - \mu I)^m$ is one to one. This will be done by showing that $(A - \mu I)$ is one to one. Let $w \in V_i$ and suppose $(A - \mu I) w = 0$ so that $Aw = \mu w$. Then for some $m \in \mathbb{N}$, $(A - \lambda_i I)^m w = 0$ and so by the binomial theorem,

$$(\mu - \lambda_i)^m w = \sum_{l=0}^{m} \binom{m}{l} (-\lambda_i)^{m-l} \mu^l w$$

$$\sum_{l=0}^{m} \binom{m}{l} (-\lambda_i)^{m-l} A^l w = (A - \lambda_i I)^m w = 0.$$

Therefore, since $\mu \neq \lambda_i$, it follows $w = 0$ and this verifies $(A - \mu I)$ is one to one. Thus $(A - \mu I)^m$ is also one to one on V_i Letting $\left\{u_1^i, \cdots, u_{r_k}^i\right\}$ be a basis for V_i, it follows $\left\{(A - \mu I)^m u_1^i, \cdots, (A - \mu I)^m u_{r_k}^i\right\}$ is also a basis and so $(A - \mu I)^m$ is also onto.

Let p be the smallest integer such that $\ker (A - \lambda_k I)^p = V_k$ and define

$$W \equiv (A - \lambda_k I)^p (V).$$

Claim 2: $A : W \to W$ and λ_k is not an eigenvalue for A restricted to W.

Proof: Suppose to the contrary that $A (A - \lambda_k I)^p u = \lambda_k (A - \lambda_k I)^p u$ where $(A - \lambda_k I)^p u \neq 0$. Then subtracting $\lambda_k (A - \lambda_k I)^p u$ from both sides yields

$$(A - \lambda_k I)^{p+1} u = 0$$

and so $u \in \ker ((A - \lambda_k I)^p)$ from the definition of p. But this requires $(A - \lambda_k I)^p u = 0$ contrary to $(A - \lambda_k I)^p u \neq 0$. This has verified the claim.

It follows from this claim that the eigenvalues of A restricted to W are a subset of $\{\lambda_1, \cdots, \lambda_{k-1}\}$. Letting

$$V_i' \equiv \left\{ w \in W : (A - \lambda_i)^l w = 0 \text{ for some } l \in \mathbb{N} \right\},$$

it follows from the induction hypothesis that

$$W = \sum_{i=1}^{k-1} V_i' \subseteq \sum_{i=1}^{k-1} V_i.$$

From Claim 1, $(A - \lambda_k I)^p$ maps V_i one to one and onto V_i. Therefore, if $x \in V$, then $(A - \lambda_k I)^p x \in W$. It follows there exist $x_i \in V_i$ such that

$$(A - \lambda_k I)^p x = \sum_{i=1}^{k-1} \overbrace{(A - \lambda_k I)^p x_i}^{\in V_i}.$$

Consequently

$$(A - \lambda_k I)^p \left(\overbrace{x - \sum_{i=1}^{k-1} x_i}^{\in V_k} \right) = 0$$

and so there exists $x_k \in V_k$ such that

$$x - \sum_{i=1}^{k-1} x_i = x_k$$

and this proves the theorem.

Definition 11.3.6 *Let $\{V_i\}_{i=1}^r$ be subspaces of V which have the property that if $v_i \in V_i$ and*

$$\sum_{i=1}^{r} v_i = 0, \tag{11.4}$$

then $v_i = 0$ for each i. Under this condition, a special notation is used to denote $\sum_{i=1}^r V_i$. This notation is

$$V_1 \oplus \cdots \oplus V_r$$

and it is called a direct sum of subspaces.

Theorem 11.3.7 *Let $\{V_i\}_{i=1}^m$ be subspaces of V which have the property 11.4 and let $B_i = \left\{u_1^i, \cdots, u_{r_i}^i\right\}$ be a basis for V_i. Then $\{B_1, \cdots, B_m\}$ is a basis for $V_1 \oplus \cdots \oplus V_m = \sum_{i=1}^m V_i$.*

Proof: It is clear that span $(B_1, \cdots, B_m) = V_1 \oplus \cdots \oplus V_m$. It only remains to verify that $\{B_1, \cdots, B_m\}$ is linearly independent. Arbitrary elements of span $(B_1, \cdots, B_m)$ are of the form

$$\sum_{k=1}^{m} \sum_{i=1}^{r_i} c_i^k u_i^k.$$

Suppose then that

$$\sum_{k=1}^{m} \sum_{i=1}^{r_i} c_i^k u_i^k = 0.$$

Since $\sum_{i=1}^{r_i} c_i^k u_i^k \in V_k$ it follows $\sum_{i=1}^{r_i} c_i^k u_i^k = 0$ for each k. But then $c_i^k = 0$ for each $i = 1, \cdots, r_i$. This proves the theorem.

The following corollary is the main result.

Corollary 11.3.8 *Let V be a complex vector space of dimension, n and let $A \in \mathcal{L}(V,V)$. Also suppose $\sigma(A) = \{\lambda_1, \cdots, \lambda_s\}$ where the λ_i are distinct. Then letting V_{λ_i} denote the generalized eigenspace for λ_i,*

$$V = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_s}$$

and if B_i is a basis for V_{λ_i}, then $\{B_1, B_2, \cdots, B_s\}$ is a basis for V.

Proof: It is necessary to verify that the V_{λ_i} satisfy condition 11.4. Let $V_{\lambda_i} = \ker(A - \lambda_i I)^{r_i}$ and suppose $v_i \in V_{\lambda_i}$ and $\sum_{i=1}^{k} v_i = 0$ where $k \leq s$. It is desired to show this implies each $v_i = 0$. It is clearly true if $k = 1$. Suppose then that the condition holds for $k-1$ and

$$\sum_{i=1}^{k} v_i = 0$$

and not all the $v_i = 0$. By Claim 1 in the proof of Theorem 11.3.5, multiplying by $(A - \lambda_k I)^{r_k}$ yields

$$\sum_{i=1}^{k-1} (A - \lambda_k I)^{r_k} v_i = \sum_{i=1}^{k-1} v_i' = 0$$

where $v_i' \in V_{\lambda_i}$. Now by induction, each $v_i' = 0$ and so each $v_i = 0$ for $i \leq k-1$. Therefore, the sum, $\sum_{i=1}^{k} v_i$ reduces to v_k and so $v_k = 0$ also.

By Theorem 11.3.5, $\sum_{i=1}^{s} V_{\lambda_i} = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_s} = V$ and by Theorem 11.3.7 $\{B_1, B_2, \cdots, B_s\}$ is a basis for V. This proves the corollary.

11.4 Block Diagonal Matrices

In this section the vector space will be $\mathbb{C}^n$ and the linear transformations will be $n \times n$ matrices.

Definition 11.4.1 *Let A and B be two $n \times n$ matrices. Then A is similar to B, written as $A \sim B$ when there exists an invertible matrix, S such that $A = S^{-1}BS$.*

Theorem 11.4.2 *Let A be an $n \times n$ matrix. Letting $\lambda_1, \lambda_2, \cdots, \lambda_r$ be the distinct eigenvalues of A,arranged in any order, there exist square matrices, $P_1, \cdots, P_r$ such that A is similar to the block diagonal matrix,*

$$P = \begin{pmatrix} P_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_r \end{pmatrix}$$

in which P_k has the single eigenvalue λ_k. Denoting by r_k the size of P_k it follows that r_k equals the dimension of the generalized eigenspace for λ_k,

$$r_k = \dim\{\mathbf{x} : (A - \lambda_k I)^m \mathbf{x} = 0 \text{ for some } m\} \equiv \dim(V_{\lambda_k})$$

Furthermore, if S is the matrix satisfying $S^{-1}AS = P$, then S is of the form

$$\begin{pmatrix} B_1 & \cdots & B_r \end{pmatrix}$$

where $B_k = \begin{pmatrix} \mathbf{u}_1^k & \cdots & \mathbf{u}_{r_k}^k \end{pmatrix}$ in which the columns, $\{\mathbf{u}_1^k, \cdots, \mathbf{u}_{r_k}^k\} = D_k$ constitute a basis for V_{λ_k}.

Proof: By Corollary 11.3.8 $\mathbb{C}^n = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_k}$ and a basis for $\mathbb{C}^n$ is $\{D_1, \cdots, D_r\}$ where D_k is a basis for V_{λ_k}.

Let

$$S = \begin{pmatrix} B_1 & \cdots & B_r \end{pmatrix}$$

where the B_i are the matrices described in the statement of the theorem. Then S^{-1} must be of the form

$$S^{-1} = \begin{pmatrix} C_1 \\ \vdots \\ C_r \end{pmatrix}$$

where $C_i B_i = I_{r_i \times r_i}$. Also, if $i \neq j$, then $C_i A B_j = 0$ the last claim holding because $A : V_j \to V_j$ so the columns of AB_j are linear combinations of the columns of B_j and each of these columns is orthogonal to the rows of C_i. Therefore,

$$\begin{aligned} S^{-1}AS &= \begin{pmatrix} C_1 \\ \vdots \\ C_r \end{pmatrix} A \begin{pmatrix} B_1 & \cdots & B_r \end{pmatrix} \\ &= \begin{pmatrix} C_1 \\ \vdots \\ C_r \end{pmatrix} \begin{pmatrix} AB_1 & \cdots & AB_r \end{pmatrix} \\ &= \begin{pmatrix} C_1AB_1 & 0 & \cdots & 0 \\ 0 & C_2AB_2 & \cdots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & C_rAB_r \end{pmatrix} \end{aligned}$$

and $C_{r_k} A B_{r_k}$ is an $r_k \times r_k$ matrix.

What about the eigenvalues of $C_{r_k} A B_{r_k}$? The only eigenvalue of A restricted to V_{λ_k} is λ_k because if $A\mathbf{x} = \mu\mathbf{x}$ for some $\mathbf{x} \in V_{\lambda_k}$ and $\mu \neq \lambda_k$, then as in Claim 1 of Theorem 11.3.5,

$$(A - \lambda_k I)^{r_k} \mathbf{x} \neq \mathbf{0}$$

contrary to the assumption that $\mathbf{x} \in V_{\lambda_k}$. Suppose then that $C_{r_k} A B_{r_k} \mathbf{x} = \lambda \mathbf{x}$ where $\mathbf{x} \neq \mathbf{0}$. Why is $\lambda = \lambda_k$? Let $\mathbf{y} = B_{r_k}\mathbf{x}$ so $\mathbf{y} \in V_{\lambda_k}$. Then

$$S^{-1}A\mathbf{y} = S^{-1}AS \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{x} \\ \vdots \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \vdots \\ C_{r_k} A B_{r_k} \mathbf{x} \\ \vdots \\ \mathbf{0} \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{x} \\ \vdots \\ \mathbf{0} \end{pmatrix}$$

and so

$$A\mathbf{y} = \lambda S \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{x} \\ \vdots \\ \mathbf{0} \end{pmatrix} = \lambda \mathbf{y}.$$

Therefore, $\lambda = \lambda_k$ because, as noted above, λ_k is the only eigenvalue of A restricted to V_{λ_k}. Now letting $P_k = C_{r_k} A B_{r_k}$, this proves the theorem.

The above theorem contains a result which is of sufficient importance to state as a corollary.

Corollary 11.4.3 *Let A be an $n \times n$ matrix and let D_k denote a basis for the generalized eigenspace for λ_k. Then $\{D_1, \cdots, D_r\}$ is a basis for $\mathbb{C}^n$.*

More can be said. Recall Theorem 9.5.3 on Page 166. From this theorem, there exist unitary matrices, U_k such that $U_k^* P_k U_k = T_k$ where T_k is an upper triangular matrix of the form

$$\begin{pmatrix} \lambda_k & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix} \equiv T_k$$

Now let U be the block diagonal matrix defined by

$$U \equiv \begin{pmatrix} U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_r \end{pmatrix}.$$

By Theorem 11.4.2 there exists S such that

$$S^{-1}AS = \begin{pmatrix} P_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_r \end{pmatrix}.$$

Therefore,

$$\begin{aligned} U^*SASU &= \begin{pmatrix} U_1^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_r^* \end{pmatrix} \begin{pmatrix} P_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_r \end{pmatrix} \begin{pmatrix} U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_r \end{pmatrix} \\ &= \begin{pmatrix} U_1^*P_1U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_r^*P_rU_r \end{pmatrix} = \begin{pmatrix} T_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T_r \end{pmatrix}. \end{aligned}$$

This proves most of the following corollary of Theorem 11.4.2.

Corollary 11.4.4 *Let A be an $n \times n$ matrix. Then A is similar to an upper triangular, block diagonal matrix of the form*

$$T \equiv \begin{pmatrix} T_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T_r \end{pmatrix}$$

where T_k is an upper triangular matrix having only λ_k on the main diagonal. The diagonal blocks can be arranged in any order desired. If T_k is an $m_k \times m_k$ matrix, then

$$m_k = \dim \{\mathbf{x} : (A - \lambda_k I)^m \mathbf{x} = 0 \text{ for some } m \in \mathbb{N}\}.$$

Furthermore, m_k is the multiplicity of λ_k as a zero of the characteristic polynomial of A.

Proof: The only thing which remains is the assertion that m_k equals the multiplicity of λ_k as a zero of the characteristic polynomial. However, this is clear from the obserivation that since T is similar to A they have the same characteristic polynomial because

$$\begin{aligned} \det(A - \lambda I) &= \det\left(S(T - \lambda I)S^{-1}\right) \\ &= \det(S)\det\left(S^{-1}\right)\det(T - \lambda I) \\ &= \det\left(SS^{-1}\right)\det(T - \lambda I) \\ &= \det(T - \lambda I) \end{aligned}$$

and the observation that since T is upper triangular, the characteristic polynomial of T is of the form

$$\prod_{k=1}^{r} (\lambda_k - \lambda)^{m_k}.$$

The above corollary has tremendous significance especially if it is pushed even further resulting in the Jordan Canonical form. This form involves still more similarity transformations resulting in an especially revealing and simple form for each of the T_k, but the result of the above corollary is sufficient for most applications.

It is significant because it enables one to obtain great understanding of powers of A by using the matrix T. From Corollary 11.4.4 there exists an $n \times n$ matrix, S[2] such that

$$A = S^{-1}TS.$$

Therefore, $A^2 = S^{-1}TSS^{-1}TS = S^{-1}T^2S$ and continuing this way, it follows

$$A^k = S^{-1}T^kS.$$

where T is given in the above corollary. Consider T^k. By block multiplication,

$$T^k = \begin{pmatrix} T_1^k & & 0 \\ & \ddots & \\ 0 & & T_r^k \end{pmatrix}.$$

The matrix, T_s is an $m_s \times m_s$ matrix which is of the form

$$T_s = \begin{pmatrix} \alpha & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha \end{pmatrix} \tag{11.5}$$

which can be written in the form

$$T_s = D + N$$

for D a multiple of the identity and N an upper triangular matrix with zeros down the main diagonal. Therefore, by the Cayley Hamilton theorem, $N^{m_s} = 0$ because the characteristic equation for N is just $\lambda^{m_s} = 0$. Such a transformation is called nilpotent. You can see $N^{m_s} = 0$ directly also, without having to use the Cayley Hamilton theorem. Now since D is just a multiple of the identity, it follows that $DN = ND$. Therefore, the usual binomial theorem may be applied and this yields the following equations for $k \geq m_s$.

$$\begin{aligned} T_s^k &= (D+N)^k = \sum_{j=0}^{k} \binom{k}{j} D^{k-j}N^j \\ &= \sum_{j=0}^{m_s} \binom{k}{j} D^{k-j}N^j, \end{aligned} \tag{11.6}$$

the third equation holding because $N^{m_s} = 0$. Thus T_s^k is of the form

$$T_s^k = \begin{pmatrix} \alpha^k & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha^k \end{pmatrix}.$$

[2]The S here is written as S^{-1} in the corollary.

Lemma 11.4.5 *Suppose T is of the form T_s described above in 11.5 where the constant, α, on the main diagonal is less than one in absolute value. Then*

$$\lim_{k\to\infty}\left(T^k\right)_{ij}=0.$$

Proof: From 11.6, it follows that for large k, and $j \le m_s$,

$$\binom{k}{j} \le \frac{k\left(k-1\right)\cdots\left(k-m_s+1\right)}{m_s!}.$$

Therefore, letting C be the largest value of $\left|\left(N^j\right)_{pq}\right|$ for $0 \le j \le m_s$,

$$\left|\left(T^k\right)_{pq}\right| \le m_s C\left(\frac{k\left(k-1\right)\cdots\left(k-m_s+1\right)}{m_s!}\right)\left|\alpha\right|^{k-m_s}$$

which converges to zero as $k \to \infty$. This is most easily seen by applying the ratio test to the series

$$\sum_{k=m_s}^{\infty}\left(\frac{k\left(k-1\right)\cdots\left(k-m_s+1\right)}{m_s!}\right)\left|\alpha\right|^{k-m_s}$$

and then noting that if a series converges, then the k^{th} term converges to zero.

11.5 The Matrix Of A Linear Transformation

If V is an n dimensional vector space and $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is a basis for V, there exists a linear map

$$q:\mathbb{F}^n \to V$$

defined as

$$q\left(\mathbf{a}\right) \equiv \sum_{i=1}^{n} a_i\mathbf{v}_i$$

where

$$\mathbf{a}=\sum_{i=1}^{n} a_i\mathbf{e}_i,$$

for $\mathbf{e}_i$ the standard basis vectors for $\mathbb{F}^n$ consisting of

$$\mathbf{e}_i \equiv \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

where the one is in the i^{th} slot. It is clear that q defined in this way, is one to one, onto, and linear. For $\mathbf{v} \in V$, $q^{-1}\left(\mathbf{v}\right)$ is a list of scalars called the components of $\mathbf{v}$ with respect to the basis $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$.

Definition 11.5.1 *Given a linear transformation L, mapping V to W, where $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is a basis of V and $\{\mathbf{w}_1,\cdots,\mathbf{w}_m\}$ is a basis for W, an $m\times n$ matrix $A=(a_{ij})$is called the matrix of the transformation L with respect to the given choice of bases for V and W, if whenever $\mathbf{v}\in V$, then multiplication of the components of $\mathbf{v}$ by (a_{ij}) yields the components of $L\mathbf{v}$.*

The following diagram is descriptive of the definition. Here q_V and q_W are the maps defined above with reference to the bases, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ respectively.

$$\begin{array}{ccccc} & & L & & \\ \{\mathbf{v}_1, \cdots, \mathbf{v}_n\} & V & \rightarrow & W & \{\mathbf{w}_1, \cdots, \mathbf{w}_m\} \\ & q_V \uparrow & \circ & \uparrow q_W & \\ & \mathbb{F}^n & \rightarrow & \mathbb{F}^m & \\ & & A & & \end{array} \tag{11.7}$$

Letting $\mathbf{b} \in \mathbb{F}^n$, this requires

$$\sum_{i,j} a_{ij} b_j \mathbf{w}_i = L \sum_j b_j \mathbf{v}_j = \sum_j b_j L\mathbf{v}_j.$$

Now

$$L\mathbf{v}_j = \sum_i c_{ij} \mathbf{w}_i \tag{11.8}$$

for some choice of scalars c_{ij} because $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ is a basis for W. Hence

$$\sum_{i,j} a_{ij} b_j \mathbf{w}_i = \sum_j b_j \sum_i c_{ij} \mathbf{w}_i = \sum_{i,j} c_{ij} b_j \mathbf{w}_i.$$

It follows from the linear independence of $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ that

$$\sum_j a_{ij} b_j = \sum_j c_{ij} b_j$$

for any choice of $\mathbf{b} \in \mathbb{F}^n$ and consequently

$$a_{ij} = c_{ij}$$

where c_{ij} is defined by 11.8. It may help to write 11.8 in the form

$$\left(\begin{array}{ccc} L\mathbf{v}_1 & \cdots & L\mathbf{v}_n \end{array} \right) = \left(\begin{array}{ccc} \mathbf{w}_1 & \cdots & \mathbf{w}_m \end{array} \right) C = \left(\begin{array}{ccc} \mathbf{w}_1 & \cdots & \mathbf{w}_m \end{array} \right) A \tag{11.9}$$

where $C = (c_{ij}), A = (a_{ij})$.

Example 11.5.2 *Let*

$$V \equiv \{ \textit{ polynomials of degree 3 or less}\},$$

$$W \equiv \{ \textit{ polynomials of degree 2 or less}\},$$

and $L \equiv D$ where D is the differentiation operator. A basis for V is $\{1, x, x^2, x^3\}$ and a basis for W is $\{1, x, x^2\}$.

What is the matrix of this linear transformation with respect to this basis? Using 11.9,

$$\left(\begin{array}{cccc} 0 & 1 & 2x & 3x^2 \end{array} \right) = \left(\begin{array}{ccc} 1 & x & x^2 \end{array} \right) C.$$

It follows from this that

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

Now consider the important case where $V = \mathbb{F}^n$, $W = \mathbb{F}^m$, and the basis chosen is the standard basis of vectors $\mathbf{e}_i$ described above. Let L be a linear transformation from $\mathbb{F}^n$ to

$\mathbb{F}^m$ and let A be the matrix of the transformation with respect to these bases. In this case the coordinate maps q_V and q_W are simply the identity map and the requirement that A is the matrix of the transformation amounts to

$$\pi_i(L\mathbf{b}) = \pi_i(A\mathbf{b})$$

where π_i denotes the map which takes a vector in $\mathbb{F}^m$ and returns the i^{th} entry in the vector, the i^{th} component of the vector with respect to the standard basis vectors. Thus, if the components of the vector in $\mathbb{F}^n$ with respect to the standard basis are $(b_1, \cdots, b_n)$,

$$\mathbf{b} = \begin{pmatrix} b_1 & \cdots & b_n \end{pmatrix}^T = \sum_i b_i \mathbf{e}_i,$$

then

$$\pi_i(L\mathbf{b}) \equiv (L\mathbf{b})_i = \sum_j a_{ij} b_j.$$

What about the situation where different pairs of bases are chosen for V and W? How are the two matrices with respect to these choices related? Consider the following diagram which illustrates the situation.

$$\begin{array}{ccc} \mathbb{F}^n & \xrightarrow{A_2} & \mathbb{F}^m \\ q_2 \downarrow & \circ & p_2 \downarrow \\ V & \xrightarrow{L} & W \\ q_1 \uparrow & \circ & p_1 \uparrow \\ \mathbb{F}^n & \xrightarrow{A_1} & \mathbb{F}^m \end{array}$$

In this diagram q_i and p_i are coordinate maps as described above. From the diagram,

$$p_1^{-1} p_2 A_2 q_2^{-1} q_1 = A_1,$$

where $q_2^{-1} q_1$ and $p_1^{-1} p_2$ are one to one, onto, and linear maps.

Definition 11.5.3 *In the special case where $V = W$ and only one basis is used for $V = W$, this becomes*

$$q_1^{-1} q_2 A_2 q_2^{-1} q_1 = A_1.$$

Letting S be the matrix of the linear transformation $q_2^{-1} q_1$ with respect to the standard basis vectors in $\mathbb{F}^n$,

$$S^{-1} A_2 S = A_1. \tag{11.10}$$

When this occurs, A_1 is said to be similar to A_2 and $A \to S^{-1} A S$ is called a similarity transformation.

Here is some terminology.

Definition 11.5.4 *Let S be a set. The symbol, $\sim$ is called an equivalence relation on S if it satisfies the following axioms.*

1. *$x \sim x$ for all $x \in S$. (Reflexive)*
2. *If $x \sim y$ then $y \sim x$. (Symmetric)*
3. *If $x \sim y$ and $y \sim z$, then $x \sim z$. (Transitive)*

Definition 11.5.5 *$[x]$ denotes the set of all elements of S which are equivalent to x and $[x]$ is called the equivalence class determined by x or just the equivalence class of x.*

With the above definition one can prove the following simple theorem which you should do if you have not seen it.

Theorem 11.5.6 *Let $\sim$ be an equivalence class defined on a set, S and let $\mathcal{H}$ denote the set of equivalence classes. Then if $[x]$ and $[y]$ are two of these equivalence classes, either $x \sim y$ and $[x] = [y]$ or it is not true that $x \sim y$ and $[x] \cap [y] = \emptyset$.*

Theorem 11.5.7 *In the vector space of $n \times n$ matrices, define*

$$A \sim B$$

if there exists an invertible matrix S such that

$$A = S^{-1}BS.$$

Then $\sim$ is an equivalence relation and $A \sim B$ if and only if whenever V is an n dimensional vector space, there exists $L \in \mathcal{L}(V,V)$ and bases $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ such that A is the matrix of L with respect to $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and B is the matrix of L with respect to $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$.

Proof: $A \sim A$ because $S = I$ works in the definition. If $A \sim B$, then $B \sim A$, because

$$A = S^{-1}BS$$

implies

$$B = SAS^{-1}.$$

If $A \sim B$ and $B \sim C$, then

$$A = S^{-1}BS,\ B = T^{-1}CT$$

and so

$$A = S^{-1}T^{-1}CTS = (TS)^{-1}\,CTS$$

which implies $A \sim C$. This verifies the first part of the conclusion.

Now let V be an n dimensional vector space, $A \sim B$ and pick a basis for V,

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}.$$

Define $L \in \mathcal{L}(V,V)$ by

$$L\mathbf{v}_i \equiv \sum_j a_{ji}\mathbf{v}_j$$

where $A = (a_{ij})$. Then if $B = (b_{ij})$, and $S = (s_{ij})$ is the matrix which provides the similarity transformation,

$$A = S^{-1}BS,$$

between A and B, it follows that

$$L\mathbf{v}_i = \sum_{r,s,j} s_{ir}b_{rs}\left(s^{-1}\right)_{sj}\mathbf{v}_j. \tag{11.11}$$

Now define

$$\mathbf{w}_i \equiv \sum_j \left(s^{-1}\right)_{ij}\mathbf{v}_j.$$

Then from 11.11,

$$\sum_i \left(s^{-1}\right)_{ki} L\mathbf{v}_i = \sum_{i,j,r,s} \left(s^{-1}\right)_{ki} s_{ir}b_{rs}\left(s^{-1}\right)_{sj}\mathbf{v}_j$$

and so

$$L\mathbf{w}_k = \sum_s b_{ks}\mathbf{w}_s.$$

This proves the theorem because the if part of the conclusion was established earlier.

Definition 11.5.8 *An $n \times n$ matrix, A, is diagonalizable if there exists an invertible $n \times n$ matrix, S such that $S^{-1}AS = D$, where D is a diagonal matrix. Thus D has zero entries everywhere except on the main diagonal. Write* $\operatorname{diag}(\lambda_1 \cdots, \lambda_n)$ *to denote the diagonal matrix having the λ_i down the main diagonal.*

The following theorem is of great significance.

Theorem 11.5.9 *Let A be an $n \times n$ matrix. Then A is diagonalizable if and only if $\mathbb{F}^n$ has a basis of eigenvectors of A. In this case, S of Definition 11.5.8 consists of the $n \times n$ matrix whose columns are the eigenvectors of A and $D = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$.*

Proof: Suppose first that $\mathbb{F}^n$ has a basis of eigenvectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ where $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$. Then let S denote the matrix $(\mathbf{v}_1 \cdots \mathbf{v}_n)$ and let $S^{-1} \equiv \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix}$ where $\mathbf{u}_i^T\mathbf{v}_j = \delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$. S^{-1} exists because S has rank n. Then from block multiplication,

$$S^{-1}AS = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix} (A\mathbf{v}_1 \cdots A\mathbf{v}_n)$$

$$= \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix} (\lambda_1\mathbf{v}_1 \cdots \lambda_n\mathbf{v}_n)$$

$$= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} = D.$$

Next suppose A is diagonalizable so $S^{-1}AS = D \equiv \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$. Then the columns of S form a basis because S^{-1} is given to exist. It only remains to verify that these columns of A are eigenvectors. But letting $S = (\mathbf{v}_1 \cdots \mathbf{v}_n)$, $AS = SD$ and so $(A\mathbf{v}_1 \cdots A\mathbf{v}_n) = (\lambda_1\mathbf{v}_1 \cdots \lambda_n\mathbf{v}_n)$ which shows that $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$. This proves the theorem.

It makes sense to speak of the determinant of a linear transformation as described in the following corollary.

Corollary 11.5.10 *Let $L \in \mathcal{L}(V, V)$ where V is an n dimensional vector space and let A be the matrix of this linear transformation with respect to a basis on V. Then it is possible to define*

$$\det(L) \equiv \det(A).$$

Proof: Each choice of basis for V determines a matrix for L with respect to the basis. If A and B are two such matrices, it follows from Theorem 11.5.7 that

$$A = S^{-1}BS$$

and so

$$\det(A) = \det\left(S^{-1}\right)\det(B)\det(S).$$

But

$$1 = \det(I) = \det\left(S^{-1}S\right) = \det(S)\det\left(S^{-1}\right)$$

and so

$$\det(A) = \det(B)$$

which proves the corollary.

Definition 11.5.11 *Let $A \in \mathcal{L}(X,Y)$ where X and Y are finite dimensional vector spaces. Define* $\operatorname{rank}(A)$ *to equal the dimension of $A(X)$.*

The following theorem explains how the rank of A is related to the rank of the matrix of A.

Theorem 11.5.12 *Let $A \in \mathcal{L}(X,Y)$. Then* $\operatorname{rank}(A) = \operatorname{rank}(M)$ *where M is the matrix of A taken with respect to a pair of bases for the vector spaces X, and Y.*

Proof: Recall the diagram which describes what is meant by the matrix of A. Here the two bases are as indicated.

$$\begin{array}{ccccc} \{v_1,\cdots,v_n\} & X & \underrightarrow{A} & Y & \{w_1,\cdots,w_m\} \\ & q_X\uparrow & \circ & \uparrow q_Y & \\ & \mathbb{F}^n & \underrightarrow{M} & \mathbb{F}^m & \end{array}$$

Let $\{z_1,\cdots,z_r\}$ be a basis for $A(X)$. Then since the linear transformation, q_Y is one to one and onto, $\left\{q_Y^{-1}z_1,\cdots,q_Y^{-1}z_r\right\}$ is a linearly independent set of vectors in $\mathbb{F}^m$. Let $Au_i = z_i$. Then

$$Mq_X^{-1}u_i = q_Y^{-1}z_i$$

and so the dimension of $M(\mathbb{F}^n) \geq r$. Now if $M(\mathbb{F}^n) < r$ then there exists

$$\mathbf{y} \in M(\mathbb{F}^n) \setminus \operatorname{span}\left\{q_Y^{-1}z_1,\cdots,q_Y^{-1}z_r\right\}.$$

But then there exists $\mathbf{x} \in \mathbb{F}^n$ with $M\mathbf{x} = \mathbf{y}$. Hence

$$\mathbf{y} = M\mathbf{x} = q_Y^{-1}Aq_X\mathbf{x} \in \operatorname{span}\left\{q_Y^{-1}z_1,\cdots,q_Y^{-1}z_r\right\}$$

a contradiction. This proves the theorem.

The following result is a summary of many concepts.

Theorem 11.5.13 *Let $L \in \mathcal{L}(V,V)$ where V is a finite dimensional vector space. Then the following are equivalent.*

1. *L is one to one.*
2. *L maps a basis to a basis.*
3. *L is onto.*

4. $\det(L) \neq 0$

5. *If* $Lv = 0$ *then* $v = 0$.

Proof: Suppose first L is one to one and let $\{v_i\}_{i=1}^n$ be a basis. Then if $\sum_{i=1}^n c_i L v_i = 0$ it follows $L\left(\sum_{i=1}^n c_i v_i\right) = 0$ which means that since $L(0) = 0$, and L is one to one, it must be the case that $\sum_{i=1}^n c_i v_i = 0$. Since $\{v_i\}$ is a basis, each $c_i = 0$ which shows $\{Lv_i\}$ is a linearly independent set. Since there are n of these, it must be that this is a basis.

Now suppose 2.). Then letting $\{v_i\}$ be a basis, and $y \in V$, it follows from part 2.) that there are constants, $\{c_i\}$ such that $y = \sum_{i=1}^n c_i L v_i = L\left(\sum_{i=1}^n c_i v_i\right)$. Thus L is onto. It has been shown that 2.) implies 3.).

Now suppose 3.). Then the operation consisting of multiplication by the matrix of L, M_L, must be onto. However, the vectors in $\mathbb{F}^n$ so obtained, consist of linear combinations of the columns of M_L. Therefore, the column rank of M_L is n. By Theorem 6.3.20 this equals the determinant rank and so $\det(M_L) \equiv \det(L) \neq 0$.

Now assume 4.) If $Lv = 0$ for some $v \neq 0$, it follows that $M_L \mathbf{x} = 0$ for some $\mathbf{x} \neq \mathbf{0}$. Therefore, the columns of M_L are linearly dependent and so by Theorem 6.3.20, $\det(M_L) = \det(L) = 0$ contrary to 4.). Therefore, 4.) implies 5.).

Now suppose 5.) and suppose $Lv = Lw$. Then $L(v - w) = 0$ and so by 5.), $v - w = 0$ showing that L is one to one. This proves the theorem.

Also it is important to note that composition of linear transformation corresponds to multiplication of the matrices. Consider the following diagram.

$$\begin{array}{ccccc} X & \underrightarrow{A} & Y & \underrightarrow{B} & Z \\ q_X \uparrow & \circ & \uparrow q_Y & \circ & \uparrow q_Z \\ \mathbb{F}^n & \underrightarrow{M_A} & \mathbb{F}^m & \underrightarrow{M_B} & \mathbb{F}^p \end{array}$$

where A and B are two linear transformations, $A \in \mathcal{L}(X, Y)$ and $B \in \mathcal{L}(Y, Z)$. Then $B \circ A \in \mathcal{L}(X, Z)$ and so it has a matrix with respect to bases given on X and Z, the coordinate maps for these bases being q_X and q_Z respectively. Then

$$B \circ A = q_Z M_B q_Y q_Y^{-1} M_A q_X^{-1} = q_Z M_B M_A q_X^{-1}.$$

But this shows that $M_B M_A$ plays the role of $M_{B \circ A}$, the matrix of $B \circ A$. Hence the matrix of $B \circ A$ equals the product of the two matrices M_A and M_B. Of course it is interesting to note that although $M_{B \circ A}$ must be unique, the matrices, M_B and M_A are not unique, depending on the basis chosen for Y.

Theorem 11.5.14 *The matrix of the composition of linear transformations equals the product of the the matrices of these linear transformations.*

11.5.1 Some Geometrically Defined Linear Transformations

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

$$A\mathbf{x} = T\mathbf{x} \tag{11.12}$$

for all $\mathbf{x} \in \mathbb{F}^n$. You simply take the matrix of the linear transformation with respect to the standard basis. What is the form of A? 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 11.12. 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 $T(\mathbf{e}_i)$. This proves the following theorem.

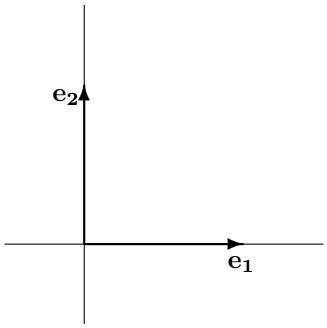
Theorem 11.5.15 *Let T be a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$. Then the matrix, A satisfying 11.12 is given by*

$$\begin{pmatrix} | & & | \\ T(\mathbf{e}_1) & \cdots & T(\mathbf{e}_n) \\ | & & | \end{pmatrix}$$

where $T\mathbf{e}_i$ is the i^{th} column of A.

Example 11.5.16 *Determine the matrix for the transformation mapping $\mathbb{R}^2$ to $\mathbb{R}^2$ which consists of rotating every vector counter clockwise 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 Theorem 11.5.15, 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 11.5.15,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Example 11.5.17 *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

$$\begin{aligned} A_{\theta+\phi} &= \begin{pmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{pmatrix} = A_\theta A_\phi \\ &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}. \end{aligned}$$

Therefore,

$$\begin{pmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{pmatrix} = \begin{pmatrix} \cos\theta\cos\phi - \sin\theta\sin\phi & -\cos\theta\sin\phi - \sin\theta\cos\phi \\ \sin\theta\cos\phi + \cos\theta\sin\phi & \cos\theta\cos\phi - \sin\theta\sin\phi \end{pmatrix}.$$

Don't these look familiar? They are the usual trig. identities for the sum of two angles derived here using linear algebra concepts.

Example 11.5.18 *Find the matrix of the linear transformation which rotates vectors in $\mathbb{R}^3$ counterclockwise about the positive z axis.*

Let T be the name of this linear transformation. In this case, $T\mathbf{e}_3 = \mathbf{e}_3, T\mathbf{e}_1 = (\cos\theta, \sin\theta, 0)^T$, and $T\mathbf{e}_2 = (-\sin\theta, \cos\theta, 0)^T$. Therefore, the matrix of this transformation is just

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{11.13}$$

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 11.5.19 *Let the projection map be defined above and let $\mathbf{u} = (1,2,3)^T$. Find the matrix of this linear transformation with respect to the usual basis.*

You can find this matrix in the same way as in earlier examples. $\mathrm{proj}_{\mathbf{u}}(\mathbf{e}_i)$ 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

$$\frac{1}{14}\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{pmatrix}.$$

Example 11.5.20 *Find the matrix of the linear transformation which reflects all vectors in $\mathbb{R}^3$ through the xz plane.*

As illustrated above, you just need to find $T\mathbf{e}_i$ where T is the name of the transformation. But $T\mathbf{e}_1 = \mathbf{e}_1, T\mathbf{e}_3 = \mathbf{e}_3$, and $T\mathbf{e}_2 = -\mathbf{e}_2$ so the matrix is

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

Example 11.5.21 *Find the matrix of the linear transformation which first rotates counter clockwise about the positive z axis and then reflects through the xz plane.*

This linear transformation is just the composition of two linear transformations having matrices

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

respectively. Thus the matrix desired is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ -\sin\theta & -\cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

11.5.2 Rotations About A Given Vector

As an application, I will consider the problem of rotating counter clockwise about a given unit vector which is possibly not one of the unit vectors in coordinate directions. First consider a pair of perpendicular coordinate vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$ and the problem of rotating the in the counterclockwise direction about $\mathbf{u}_3$ where $\mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2$ so that $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ forms a right handed orthogonal coordinate system. Let T denote the desired rotation. Then

$$T(a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3) = aT\mathbf{u}_1 + bT\mathbf{u}_2 + cT\mathbf{u}_3$$

$$= (a\cos\theta - b\sin\theta)\,\mathbf{u}_1 + (a\sin\theta + b\cos\theta)\,\mathbf{u}_2 + c\mathbf{u}_3.$$

Thus in terms of the basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$, the matrix of this transformation is

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

I want to write this transformation in terms of the usual basis vectors, $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$.

$$\begin{array}{ccc} \mathbb{R}^3 & \overset{M_{\theta_1}}{\rightarrow} & \mathbb{R}^3 \\ \theta_1 \downarrow & \circ & \downarrow \theta_1 \\ \mathbb{R}^3 & \overset{T}{\rightarrow} & \mathbb{R}^3 \\ \theta_2 \uparrow & \circ & \uparrow \theta_2 \\ \mathbb{R}^3 & \overset{M_{\theta_2}}{\rightarrow} & \mathbb{R}^3 \end{array}$$

In the above, θ_2 can be considered as the matrix which has the indicated columns below because $\theta_2\mathbf{x} \equiv \sum_i x_i\mathbf{u}_i$ so that $\theta_2(\mathbf{e}_i) = \mathbf{u}_i$. The matrix M_{θ_1} denotes the matrix of the transformation T taken with respect to the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ while the matrix, M_{θ_2} denotes the matrix of the transormation T taken with respect to the basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$.

$$\theta_2 = \begin{pmatrix} | & | & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \\ | & | & | \end{pmatrix}$$

Since this is an orthogonal matrix, it follows that

$$\theta_2^{-1} = \theta_2^T = \begin{pmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \mathbf{u}_3^T \end{pmatrix}.$$

Also, θ_1 is just the identity matrix. Therefore, from the above diagram,

$$M_{\theta_1} = \theta_2 M_{\theta_2}\theta_2^{-1}$$

$$= \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \mathbf{u}_3^T \end{pmatrix}.$$

Now suppose the unit vector about which the counterclockwise rotation takes place is $\langle a, b, c\rangle$. Then I obtain vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$ such that $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is a right handed orthogonal system with $\mathbf{u}_3 = \langle a, b, c\rangle$ and then use the above result. It is of course somewhat arbitrary how this is accomplished. I will assume, however that $|c| \neq 1$ since if this condition holds, then you are looking at either clockwise or counter clockwise rotation about the positive z axis and this is a problem which has been dealt with earlier. (If $c = -1$, it amounts to clockwise rotation about the positive z axis while if $c = 1$, it is counterclockwise roation about the positive z axis.) Then let $\mathbf{u}_3 = \langle a, b, c\rangle$ and $\mathbf{u}_2 \equiv \frac{1}{\sqrt{a^2+b^2}}\langle b, -a, 0\rangle$. If $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is to be a right hand system it is necessary to have

$$\begin{aligned} \mathbf{u}_1 &= \mathbf{u}_2 \times \mathbf{u}_3 = \frac{1}{\sqrt{(a^2+b^2)(a^2+b^2+c^2)}}\left\langle -ac, -bc, a^2+b^2\right\rangle \\ &= \frac{1}{\sqrt{(a^2+b^2)}}\left\langle -ac, -bc, a^2+b^2\right\rangle \end{aligned}$$

Then from the above, the matrix of the transformation in terms of the standard basis is given by

$$\begin{pmatrix} -\frac{ac}{\sqrt{(a^2+b^2)}} & \frac{b}{\sqrt{(a^2+b^2)}} & a \\ -\frac{bc}{\sqrt{(a^2+b^2)}} & -\frac{a}{\sqrt{(a^2+b^2)}} & b \\ \sqrt{(a^2+b^2)} & 0 & c \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
$$\begin{pmatrix} \frac{-ac}{\sqrt{(a^2+b^2)}} & \frac{-bc}{\sqrt{(a^2+b^2)}} & \sqrt{(a^2+b^2)} \\ \frac{b}{\sqrt{(a^2+b^2)}} & \frac{-a}{\sqrt{(a^2+b^2)}} & 0 \\ a & b & c \end{pmatrix}$$

which after simplification equals

$$= \begin{pmatrix} a^2+\left(1-a^2\right)\cos\theta & ab\left(1-\cos\theta\right)-c\sin\theta & ac\left(1-\cos\theta\right)+b\sin\theta \\ ab\left(1-\cos\theta\right)+c\sin\theta & b^2+\left(1-b^2\right)\cos\theta & bc\left(1-\cos\theta\right)-a\sin\theta \\ ac\left(1-\cos\theta\right)-b\sin\theta & bc\left(1-\cos\theta\right)+a\sin\theta & c^2+\left(1-c^2\right)\cos\theta \end{pmatrix}. \quad (11.14)$$

With this, it is clear how to rotate clockwise about the the unit vector, $\langle a,b,c\rangle$. Just rotate counter clockwise through an angle of $-\theta$. Thus the matrix for this clockwise roation is just

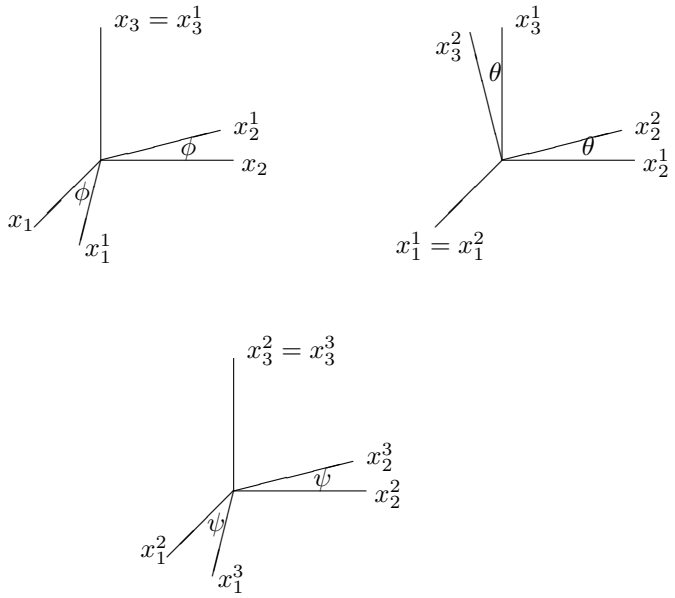
$$= \begin{pmatrix} a^2+\left(1-a^2\right)\cos\theta & ab\left(1-\cos\theta\right)+c\sin\theta & ac\left(1-\cos\theta\right)-b\sin\theta \\ ab\left(1-\cos\theta\right)-c\sin\theta & b^2+\left(1-b^2\right)\cos\theta & bc\left(1-\cos\theta\right)+a\sin\theta \\ ac\left(1-\cos\theta\right)+b\sin\theta & bc\left(1-\cos\theta\right)-a\sin\theta & c^2+\left(1-c^2\right)\cos\theta \end{pmatrix}.$$

In deriving 11.14 it was assumed that $c \neq \pm 1$ but even in this case, it gives the correct answer. Suppose for example that $c = 1$ so you are rotating in the counter clockwise direction about the positive z axis. Then a, b are both equal to zero and 11.14 reduces to 11.13.

11.5.3 The Euler Angles

An important application of the above theory is to the Euler angles, important in the mechanics of rotating bodies. Lagrange studied these things back in the 1700's. To describe the Euler angles consider the following picture in which x_1, x_2 and x_3 are the usual coordinate axes fixed in space and the axes labeled with a superscript denote other coordinate axes.

Here is the picture.



We obtain ϕ by rotating counter clockwise about the fixed x_3 axis. Thus this rotation has the matrix

$$\begin{pmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \equiv M_1(\phi)$$

Next rotate counter clockwise about the x_1^1 axis which results from the first rotation through an angle of θ. Thus it is desired to rotate counter clockwise through an angle θ about the unit vector

$$\begin{pmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos\phi \\ \sin\phi \\ 0 \end{pmatrix}.$$

Therefore, in 11.14, $a = \cos\phi, b = \sin\phi$, and $c = 0$. It follows the matrix of this transformation with respect to the usual basis is

$$\begin{pmatrix} \cos^2\phi + \sin^2\phi\cos\theta & \cos\phi\sin\phi(1-\cos\theta) & \sin\phi\sin\theta \\ \cos\phi\sin\phi(1-\cos\theta) & \sin^2\phi + \cos^2\phi\cos\theta & -\cos\phi\sin\theta \\ -\sin\phi\sin\theta & \cos\phi\sin\theta & \cos\theta \end{pmatrix} \equiv M_2(\phi,\theta)$$

Finally, we rotate counter clockwise about the positive x_3^2 axis by ψ. The vector in the positive x_3^1 axis is the same as the vector in the fixed x_3 axis. Thus the unit vector in the

positive direction of the x_3^2 axis is

$$\begin{pmatrix} \cos^2\phi + \sin^2\phi\cos\theta & \cos\phi\sin\phi\left(1-\cos\theta\right) & \sin\phi\sin\theta \\ \cos\phi\sin\phi\left(1-\cos\theta\right) & \sin^2\phi + \cos^2\phi\cos\theta & -\cos\phi\sin\theta \\ -\sin\phi\sin\theta & \cos\phi\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$
$$= \begin{pmatrix} \cos^2\phi + \sin^2\phi\cos\theta \\ \cos\phi\sin\phi\left(1-\cos\theta\right) \\ -\sin\phi\sin\theta \end{pmatrix} = \begin{pmatrix} \cos^2\phi + \sin^2\phi\cos\theta \\ \cos\phi\sin\phi\left(1-\cos\theta\right) \\ -\sin\phi\sin\theta \end{pmatrix}$$

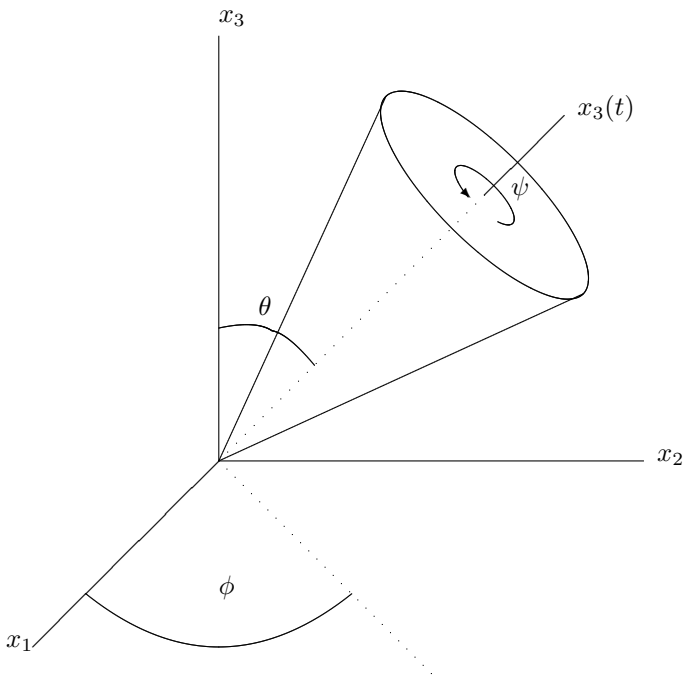
and it is desired to rotate counter clockwise through an angle of ψ about this vector. Thus, in this case,

$$a = \cos^2\phi + \sin^2\phi\cos\theta, b = \cos\phi\sin\phi\left(1-\cos\theta\right), c = -\sin\phi\sin\theta.$$

and you could substitute in to the formula of Theorem 11.14 and obtain a matrix which represents the linear transformation obtained by rotating counter clockwise about the positive x_3^2 axis, $M_3\left(\phi,\theta,\psi\right)$. Then what would be the matrix with respect to the usual basis for the linear transformation which is obtained as a composition of the three just described? By Theorem 11.5.14, this matrix equals the product of these three,

$$M_3\left(\phi,\theta,\psi\right) M_2\left(\phi,\theta\right) M_1\left(\phi\right).$$

I leave the details to you. There are procedures due to Lagrange which will allow you to write differential equations for the Euler angles in a rotating body. To give an idea how these angles apply, consider the following picture.



This is as far as I will go on this topic. The point is, it is possible to give a systematic description in terms of matrix multiplication of a very elaborate geometrical description of a composition of linear transformations. You see from the picture it is possible to describe the motion of the spinning top shown in terms of these Euler angles. I think you can also see that the end result would be pretty horrendous but this is because it involves using the basis corresponding to a fixed in space coordinate system. You wouldn't do this for the application to a spinning top.

Not surprisingly, this also has applications to computer graphics.

11.6 Exercises

1. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/3$.
2. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/4$.
3. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $-\pi/3$.
4. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $2\pi/3$.
5. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/12$. **Hint:** Note that $\pi/12 = \pi/3 - \pi/4$.
6. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $2\pi/3$ and then reflects across the x axis.
7. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/3$ and then reflects across the x axis.
8. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/4$ and then reflects across the x axis.
9. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/6$ and then reflects across the x axis followed by a reflection across the y axis.
10. Find the matrix for the linear transformation which reflects every vector in $\mathbb{R}^2$ across the x axis and then rotates every vector through an angle of $\pi/4$.
11. Find the matrix for the linear transformation which reflects every vector in $\mathbb{R}^2$ across the y axis and then rotates every vector through an angle of $\pi/4$.
12. Find the matrix for the linear transformation which reflects every vector in $\mathbb{R}^2$ across the x axis and then rotates every vector through an angle of $\pi/6$.
13. Find the matrix for the linear transformation which reflects every vector in $\mathbb{R}^2$ across the y axis and then rotates every vector through an angle of $\pi/6$.
14. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $5\pi/12$. **Hint:** Note that $5\pi/12 = 2\pi/3 - \pi/4$.
15. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1, -2, 3)^T$.

16. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1,5,3)^T$.

17. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1,0,3)^T$.

18. Show that the function $T_{\mathbf{u}}$ defined by $T_{\mathbf{u}}(\mathbf{v}) \equiv \mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v})$ is also a linear transformation.

19. If $\mathbf{u} = (1,2,3)^T$, as in Example 11.5.19 and $T_{\mathbf{u}}$ is given in the above problem, find the matrix, $A_{\mathbf{u}}$ which satisfies $A_{\mathbf{u}}\mathbf{x} = T(\mathbf{x})$.

20. If A, B, and C are each $n \times n$ matrices and ABC is invertible, why are each of A, B, and C invertible.

21. Show that $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ by doing the computation $ABC\left(C^{-1}B^{-1}A^{-1}\right)$.

22. If A is invertible, show $\left(A^T\right)^{-1} = \left(A^{-1}\right)^T$.

23. If A is invertible, show $\left(A^2\right)^{-1} = \left(A^{-1}\right)^2$.

24. If A is invertible, show $\left(A^{-1}\right)^{-1} = A$.

25. Give an example of a 3×2 matrix with the property that the linear transformation determined by this matrix is one to one but not onto.

26. Explain why $A\mathbf{x} = \mathbf{0}$ always has a solution.

27. Review problem: Suppose $\det(A - \lambda I) = 0$. Show using Theorem 6.1.17 there exists $\mathbf{x} \neq \mathbf{0}$ such that $(A - \lambda I)\mathbf{x} = \mathbf{0}$.

28. 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 $A\mathbf{x} = \mathbf{0}$.

29. Find $\ker(A)$ for

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

Recall $\ker(A)$ is just the set of solutions to $A\mathbf{x} = \mathbf{0}$.

30. Suppose $A\mathbf{x} = \mathbf{b}$ has a solution. Explain why the solution is unique precisely when $A\mathbf{x} = \mathbf{0}$ has only the trivial (zero) solution.

31. Using Problem 29, find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 11 \\ 7 \\ 18 \\ 7 \end{pmatrix}$$

32. Using Problem 29, find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 6 \\ 7 \\ 13 \\ 7 \end{pmatrix}$$

33. Show that if A is an $m \times n$ matrix, then $\ker(A)$ is a subspace.

34. Verify the linear transformation determined by the matrix of 5.27 maps $\mathbb{R}^3$ onto $\mathbb{R}^2$ but the linear transformation determined by this matrix is not one to one.

11.7 The Jordan Canonical Form

Recall Corollary 11.4.4. For convenience, this corollary is stated below.

Corollary 11.7.1 *Let A be an $n \times n$ matrix. Then A is similar to an upper triangular, block diagonal matrix of the form*

$$T \equiv \begin{pmatrix} T_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T_r \end{pmatrix}$$

where T_k is an upper triangular matrix having only λ_k on the main diagonal. The diagonal blocks can be arranged in any order desired. If T_k is an $m_k \times m_k$ matrix, then

$$m_k = \dim\{\mathbf{x} : (A - \lambda_k I)^m \mathbf{x} = 0 \text{ for some } m \in \mathbb{N}\}.$$

The Jordan Canonical form involves a further reduction in which the upper triangular matrices, T_k assume a particularly revealing and simple form.

Definition 11.7.2 *$J_k(\alpha)$ is a Jordan block if it is a $k \times k$ matrix of the form*

$$J_k(\alpha) = \begin{pmatrix} \alpha & 1 & & 0 \\ 0 & \ddots & \ddots & \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \alpha \end{pmatrix}$$

In words, there is an unbroken string of ones down the super diagonal and the number, α filling every space on the main diagonal with zeros everywhere else. A matrix is strictly upper triangular if it is of the form

$$\begin{pmatrix} 0 & * & * \\ \vdots & \ddots & * \\ 0 & \cdots & 0 \end{pmatrix},$$

where there are zeroes on the main diagonal and below the main diagonal.

The Jordan canonical form involves each of the upper triangular matrices in the conclusion of Corollary 11.4.4 being a block diagonal matrix with the blocks being Jordan blocks in which the size of the blocks decreases from the upper left to the lower right. The idea is to show that every square matrix is similar to a unique such matrix which is in Jordan canonical form.

Note that in the conclusion of Corollary 11.4.4 each of the triangular matrices is of the form $\alpha I + N$ where N is a strictly upper triangular matrix. The existence of the Jordan canonical form follows quickly from the following lemma.

Lemma 11.7.3 *Let N be an $n \times n$ matrix which is strictly upper triangular. Then there exists an invertible matrix, S such that*

$$S^{-1}NS = \begin{pmatrix} J_{r_1}(0) & & & 0 \\ & J_{r_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{r_s}(0) \end{pmatrix}$$

where $r_1 \geq r_2 \geq \cdots \geq r_s \geq 1$ and $\sum_{i=1}^{s} r_i = n$.

Proof: First note the only eigenvalue of N is 0. Let $\mathbf{v}_1$ be an eigenvector. Then $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r\}$ is called a chain based on $\mathbf{v}_1$ if $N\mathbf{v}_{k+1} = \mathbf{v}_k$ for all $k = 1, 2, \cdots, r$. It will be called a maximal chain if there is no solution, $\mathbf{v}$, to the equation, $N\mathbf{v} = \mathbf{v}_r$.

Claim 1: The vectors in any chain are linearly independent and for $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r\}$ a chain based on $\mathbf{v}_1$,

$$N : \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r) \to \operatorname{span}(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r). \tag{11.15}$$

Also if $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r\}$ is a chain, then $r \leq n$.

Proof: First note that 11.15 is obvious because

$$N \sum_{i=1}^{r} c_i \mathbf{v}_i = \sum_{i=2}^{r} c_i \mathbf{v}_{i-1}.$$

It only remains to verify the vectors of a chain are independent. If this is not true, you could consider the set of all dependent chains and pick one, $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r\}$, which is shortest. Thus $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r\}$ is a chain which is dependent and r is as small as possible. Suppose then that $\sum_{i=1}^{r} c_i \mathbf{v}_i = \mathbf{0}$ and not all the $c_i = 0$. It follows from r being the smallest length of any dependent chain that all the $c_i \neq 0$. Now $\mathbf{0} = N^{r-1}\left(\sum_{i=1}^{r} c_i \mathbf{v}_i\right) = c_1 \mathbf{v}_1$ showing that $c_1 = 0$, a contradiction. Therefore, the last claim is obvious. This proves the claim.

Consider the set of all chains based on eigenvectors. Since all have total length no larger than n it follows there exists one which has maximal length, $\left\{\mathbf{v}_1^1, \cdots, \mathbf{v}_{r_1}^1\right\} \equiv B_1$. If $\operatorname{span}(B_1)$ contains all eigenvectors of N, then stop. Otherwise, consider all chains based on eigenvectors not in $\operatorname{span}(B_1)$ and pick one, $B_2 \equiv \left\{\mathbf{v}_1^2, \cdots, \mathbf{v}_{r_2}^2\right\}$ which is as long as possible. Thus $r_2 \leq r_1$. If $\operatorname{span}(B_1, B_2)$ contains all eigenvectors of N, stop. Otherwise, consider all chains based on eigenvectors not in $\operatorname{span}(B_1, B_2)$ and pick one, $B_3 \equiv \left\{\mathbf{v}_1^3, \cdots, \mathbf{v}_{r_3}^3\right\}$ such that r_3 is as large as possible. Continue this way. Thus $r_k \geq r_{k+1}$.

Claim 2: The above process terminates with a finite list of chains, $\{B_1, \cdots, B_s\}$ because for any k, $\{B_1, \cdots, B_k\}$ is linearly independent.

Proof of Claim 2: It suffices to verify that $\{B_1, \cdots, B_k\}$ is linearly independent. This will be accomplished if it can be shown that no vector may be written as a linear combination

of the other vectors. Suppose then that j is such that $\mathbf{v}_i^j$ is a linear combination of the other vectors in $\{B_1, \cdots, B_k\}$ and that $j \leq k$ is as large as possible for this to happen. Also suppose that of the vectors, $\mathbf{v}_i^j \in B_j$ such that this occurs, i is as large as possible. Then

$$\mathbf{v}_i^j = \sum_{q=1}^{p} c_q \mathbf{w}_q$$

where the $\mathbf{w}_q$ are vectors of $\{B_1, \cdots, B_k\}$ not equal to $\mathbf{v}_i^j$. Since j is as large as possible, it follows all the $\mathbf{w}_q$ come from $\{B_1, \cdots, B_j\}$ and that those vectors, $\mathbf{v}_l^j$, which are from B_j have the property that $l < i$. Therefore,

$$\mathbf{v}_1^j = N^{i-1}\mathbf{v}_i^j = \sum_{q=1}^{p} c_q N^{i-1}\mathbf{w}_q$$

and this last sum consists of vectors in span $(B_1, \cdots, B_{j-1})$ contrary to the above construction. Therefore, this proves the claim.

Claim 3: Suppose $N\mathbf{w} = \mathbf{0}$. Then there exists scalars, c_i such that

$$\mathbf{w} = \sum_{i=1}^{s} c_i \mathbf{v}_1^i.$$

Recall that $\mathbf{v}_1^i$ is the eigenvector in the i^{th} chain on which this chain is based.

Proof of Claim 3: From the construction, $\mathbf{w} \in$ span $(B_1, \cdots, B_s)$. Therefore,

$$\mathbf{w} = \sum_{i=1}^{s} \sum_{k=1}^{r_i} c_i^k \mathbf{v}_k^i.$$

Now apply N to both sides to find

$$\mathbf{0} = \sum_{i=1}^{s} \sum_{k=2}^{r_i} c_i^k \mathbf{v}_{k-1}^i$$

and so $c_i^k = 0$ if $k \geq 2$. Therefore,

$$\mathbf{w} = \sum_{i=1}^{s} c_i^1 \mathbf{v}_1^i$$

and this proves the claim.

It remains to verify that span $(B_1, \cdots, B_s) = \mathbb{F}^n$. Suppose $\mathbf{w} \notin$ span $(B_1, \cdots, B_s)$. Since $N^n = 0$, there exists a smallest integer, k such that $N^k\mathbf{w} = \mathbf{0}$ but $N^{k-1}\mathbf{w} \neq \mathbf{0}$. Then $k \leq \min(r_1, \cdots, r_s)$ because there exists a chain of length k based on the eigenvector, $N^{k-1}\mathbf{w}$, namely

$$N^{k-1}\mathbf{w}, N^{k-2}\mathbf{w}, N^{k-3}\mathbf{w}, \cdots, \mathbf{w}$$

and this chain must be no longer than the preceding chains. Since $N^{k-1}\mathbf{w}$ is an eigenvector, it follows from Claim 3 that

$$N^{k-1}\mathbf{w} = \sum_{i=1}^{s} c_i \mathbf{v}_1^i = \sum_{i=1}^{s} c_i N^{k-1}\mathbf{v}_k^i.$$

Therefore,

$$N^{k-1}\left(\mathbf{w} - \sum_{i=1}^{s} c_i \mathbf{v}_k^i\right) = \mathbf{0}$$

and so,

$$NN^{k-2}\left(\mathbf{w}-\sum_{i=1}^{s}c_i\mathbf{v}_k^i\right)=\mathbf{0}$$

which implies by Claim 3 that

$$N^{k-2}\left(\mathbf{w}-\sum_{i=1}^{s}c_i\mathbf{v}_k^i\right)=\sum_{i=1}^{s}d_i\mathbf{v}_1^i=\sum_{i=1}^{s}d_iN^{k-2}\mathbf{v}_{k-1}^i$$

and so

$$N^{k-2}\left(\mathbf{w}-\sum_{i=1}^{s}c_i\mathbf{v}_k^i-\sum_{i=1}^{s}d_i\mathbf{v}_{k-1}^i\right)=\mathbf{0}.$$

Continuing this way it follows that for each $j<k$, there exists a vector, $\mathbf{z}_j\in\operatorname{span}(B_1,\cdots,B_s)$ such that

$$N^{k-j}(\mathbf{w}-\mathbf{z}_j)=\mathbf{0}.$$

In particular, taking $j=(k-1)$ yields

$$N(\mathbf{w}-\mathbf{z}_{k-1})=\mathbf{0}$$

and now using Claim 3 again yields $\mathbf{w}\in\operatorname{span}(B_1,\cdots,B_s)$, a contradiction. Therefore, $\operatorname{span}(B_1,\cdots,B_s)=\mathbb{F}^n$ after all and so $\{B_1,\cdots,B_s\}$ is a basis for $\mathbb{F}^n$.

Now consider the block matrix,

$$S=\begin{pmatrix} B_1 & \cdots & B_s \end{pmatrix}$$

where

$$B_k=\begin{pmatrix} \mathbf{v}_1^k & \cdots & \mathbf{v}_{r_k}^k \end{pmatrix}.$$

Thus

$$S^{-1}=\begin{pmatrix} C_1 \\ \vdots \\ C_s \end{pmatrix}$$

where $C_iB_i=I_{r_i\times r_i}$ and $C_iNB_j=0$ if $i\neq j$. Let

$$C_k=\begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_{r_k}^T \end{pmatrix}.$$

Then

$$\begin{aligned} C_kNB_k &= \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_{r_k}^T \end{pmatrix}\begin{pmatrix} N\mathbf{v}_1^k & \cdots & N\mathbf{v}_{r_k}^k \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_{r_k}^T \end{pmatrix}\begin{pmatrix} \mathbf{0} & \mathbf{v}_1^k & \cdots & \mathbf{v}_{r_k-1}^k \end{pmatrix} \end{aligned}$$

which equals an $r_k\times r_k$ matrix of the form

$$J_{r_k}(0)=\begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 \\ 0 & \cdots & \cdots & 0 \end{pmatrix}$$

That is, it has ones down the super diagonal and zeros everywhere else. It follows

$$\begin{aligned} S^{-1}NS &= \begin{pmatrix} C_1 \\ \vdots \\ C_s \end{pmatrix} N \begin{pmatrix} B_1 & \cdots & B_s \end{pmatrix} \\ &= \begin{pmatrix} J_{r_1}(0) & & & 0 \\ & J_{r_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{r_s}(0) \end{pmatrix} \end{aligned}$$

as claimed. This proves the lemma.

Now let the upper triangular matrices, T_k be given in the conclusion of Corollary 11.4.4. Thus, as noted earlier,

$$T_k = \lambda_k I_{r_k \times r_k} + N_k$$

where N_k is a strictly upper triangular matrix of the sort just discussed in Lemma 11.7.3. Therefore, there exists S_k such that $S_k^{-1} N_k S_k$ is of the form given in Lemma 11.7.3. Now $S_k^{-1}\lambda_k I_{r_k \times r_k} S_k = \lambda_k I_{r_k \times r_k}$ and so $S_k^{-1} T_k S_k$ is of the form

$$\begin{pmatrix} J_{i_1}(\lambda_k) & & & 0 \\ & J_{i_2}(\lambda_k) & & \\ & & \ddots & \\ 0 & & & J_{i_s}(\lambda_k) \end{pmatrix}$$

where $i_1 \geq i_2 \geq \cdots \geq i_s$ and $\sum_{j=1}^{s} i_j = r_k$. This proves the following corollary.

Corollary 11.7.4 *Suppose A is an upper triangular $n \times n$ matrix having α in every position on the main diagonal. Then there exists an invertible matrix, S such that*

$$S^{-1}AS = \begin{pmatrix} J_{k_1}(\alpha) & & & 0 \\ & J_{k_2}(\alpha) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(\alpha) \end{pmatrix}$$

where $k_1 \geq k_2 \geq \cdots \geq k_r \geq 1$ and $\sum_{i=1}^{r} k_i = n$.

The next theorem is the one about the existence of the Jordan canonical form.

Theorem 11.7.5 *Let A be an $n \times n$ matrix having eigenvalues $\lambda_1, \cdots, \lambda_r$ where the multiplicity of λ_i as a zero of the characteristic polynomial equals m_i. Then there exists an invertible matrix, S such that*

$$S^{-1}AS = \begin{pmatrix} J(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & J(\lambda_r) \end{pmatrix} \tag{11.16}$$

where $J(\lambda_k)$ is an $m_k \times m_k$ matrix of the form

$$\begin{pmatrix} J_{k_1}(\lambda_k) & & & 0 \\ & J_{k_2}(\lambda_k) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(\lambda_k) \end{pmatrix} \tag{11.17}$$

where $k_1 \geq k_2 \geq \cdots \geq k_r \geq 1$ and $\sum_{i=1}^{r} k_i = m_k$.

Proof: From Corollary 11.4.4, there exists S such that $S^{-1}AS$ is of the form

$$T \equiv \begin{pmatrix} T_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T_r \end{pmatrix}$$

where T_k is an upper triangular $m_k \times m_k$ matrix having only λ_k on the main diagonal. By Corollary 11.7.4 There exist matrices, S_k such that $S_k^{-1}T_kS_k = J(\lambda_k)$ where $J(\lambda_k)$ is described in 11.17. Now let M be the block diagonal matrix given by

$$M = \begin{pmatrix} S_1 & & 0 \\ & \ddots & \\ 0 & & S_r \end{pmatrix}.$$

It follows that $M^{-1}S^{-1}ASM = M^{-1}TM$ and this is of the desired form. This proves the theorem.

What about the uniqueness of the Jordan canonical form? Obviously if you change the order of the eigenvalues, you get a different Jordan canonical form but it turns out that if the order of the eigenvalues is the same, then the Jordan canonical form is unique. In fact, it is the same for any two similar matrices.

Theorem 11.7.6 *Let A and B be two similar matrices. Let J_A and J_B be Jordan forms of A and B respectively, made up of the blocks $J_A(\lambda_i)$ and $J_B(\lambda_i)$ respectively. Then J_A and J_B are identical except possibly for the order of the $J(\lambda_i)$ where the λ_i are defined above.*

Proof: First note that for λ_i an eigenvalue, the matrices $J_A(\lambda_i)$ and $J_B(\lambda_i)$ are both of size $m_i \times m_i$ because the two matrices A and B, being similar, have exactly the same characteristic equation and the size of a block equals the algebraic multiplicity of the eigenvalue as a zero of the characteristic equation. It is only necessary to worry about the number and size of the Jordan blocks making up $J_A(\lambda_i)$ and $J_B(\lambda_i)$. Let the eigenvalues of A and B be $\{\lambda_1, \cdots, \lambda_r\}$. Consider the two sequences of numbers $\{\operatorname{rank}(A-\lambda I)^m\}$ and $\{\operatorname{rank}(B-\lambda I)^m\}$. Since A and B are similar, these two sequences coincide. (Why?) Also, for the same reason, $\{\operatorname{rank}(J_A-\lambda I)^m\}$ coincides with $\{\operatorname{rank}(J_B-\lambda I)^m\}$. Now pick λ_k an eigenvalue and consider $\{\operatorname{rank}(J_A-\lambda_k I)^m\}$ and $\{\operatorname{rank}(J_B-\lambda_k I)^m\}$. Then

$$J_A - \lambda_k I = \begin{pmatrix} J_A(\lambda_1-\lambda_k) & & & & 0 \\ & \ddots & & & \\ & & J_A(0) & & \\ & & & \ddots & \\ 0 & & & & J_A(\lambda_r-\lambda_k) \end{pmatrix}$$

and a similar formula holds for $J_B - \lambda_k I$. Here

$$J_A(0) = \begin{pmatrix} J_{k_1}(0) & & & 0 \\ & J_{k_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(0) \end{pmatrix}$$

and

$$J_B(0) = \begin{pmatrix} J_{l_1}(0) & & & 0 \\ & J_{l_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{l_p}(0) \end{pmatrix}$$

and it suffices to verify that $l_i = k_i$ for all i. As noted above, $\sum k_i = \sum l_i$. Now from the above formulas,

$$\begin{aligned} \operatorname{rank}(J_A - \lambda_k I)^m &= \sum_{i \neq k} m_i + \operatorname{rank}(J_A(0)^m) \\ &= \sum_{i \neq k} m_i + \operatorname{rank}(J_B(0)^m) \\ &= \operatorname{rank}(J_B - \lambda_k I)^m, \end{aligned}$$

which shows $\operatorname{rank}(J_A(0)^m) = \operatorname{rank}(J_B(0)^m)$ for all m. However,

$$J_B(0)^m = \begin{pmatrix} J_{l_1}(0)^m & & & 0 \\ & J_{l_2}(0)^m & & \\ & & \ddots & \\ 0 & & & J_{l_p}(0)^m \end{pmatrix}$$

with a similar formula holding for $J_A(0)^m$and $\operatorname{rank}(J_B(0)^m) = \sum_{i=1}^p \operatorname{rank}(J_{l_i}(0)^m)$, similar for $\operatorname{rank}(J_A(0)^m)$. In going from m to $m+1$,

$$\operatorname{rank}(J_{l_i}(0)^m) - 1 = \operatorname{rank}\left(J_{l_i}(0)^{m+1}\right)$$

untill $m = l_i$ at which time there is no further change. Therefore, $p = r$ since otherwise, there would exist a discrepancy right away in going from $m = 1$ to $m = 2$. Now suppose the sequence $\{l_i\}$ is not equal to the sequence, $\{k_i\}$. Then $l_{r-b} \neq k_{r-b}$ for some b a nonnegative integer taken to be a small as possible. Say $l_{r-b} > k_{r-b}$. Then, letting $m = k_{r-b}$,

$$\sum_{i=1}^r \operatorname{rank}(J_{l_i}(0)^m) = \sum_{i=1}^r \operatorname{rank}(J_{k_i}(0)^m)$$

and in going to $m+1$ a discrepancy must occur because the sum on the right will contribute less to the decrease in rank than the sum on the left. This proves the theorem.

Markov Chains And Migration Processes

12.1 Regular Markov Matrices

The theorem that any matrix is similar to an appropriate block diagonal matrix is the basis for the proof of limit theorems for certain kinds of matrices called Markov matrices.

Definition 12.1.1 *An $n \times n$ matrix, $A = (a_{ij})$, is a Markov matrix if $a_{ij} \geq 0$ for all i, j and*

$$\sum_i a_{ij} = 1.$$

A Markov matrix is called regular if some power of A has all entries strictly positive. A vector, $\mathbf{v} \in \mathbb{R}^n$, is a steady state if $A\mathbf{v} = \mathbf{v}$.

Lemma 12.1.2 *Suppose $A = (a_{ij})$ is a Markov matrix in which $a_{ij} > 0$ for all i, j. Then if μ is an eigenvalue of A, either $|\mu| < 1$ or $\mu = 1$. In addition to this, if $A\mathbf{v} = \mathbf{v}$ for a nonzero vector, $\mathbf{v} \in \mathbb{R}^n$, then $v_j v_i \geq 0$ for all i, j so the components of $\mathbf{v}$ have the same sign.*

Proof: Let $\sum_j a_{ij} v_j = \mu v_i$ where $\mathbf{v} \equiv (v_1, \cdots, v_n)^T \neq \mathbf{0}$. Then

$$\sum_j a_{ij} v_j \overline{\mu v_i} = |\mu|^2 |v_i|^2$$

and so

$$|\mu|^2 |v_i|^2 = \sum_j a_{ij} \operatorname{Re}(v_j \overline{\mu v_i}) \leq \sum_j a_{ij} |v_j| |\mu| |v_i| \tag{12.1}$$

so

$$|\mu| |v_i| \leq \sum_j a_{ij} |v_j| . \tag{12.2}$$

Summing on i,

$$|\mu| \sum_i |v_i| \leq \sum_i \sum_j a_{ij} |v_j| = \sum_j \sum_i a_{ij} |v_j| = \sum_j |v_j| . \tag{12.3}$$

Therefore, $|\mu| \leq 1$.

If $|\mu| = 1$, then from 12.1,

$$|v_i| \leq \sum_j a_{ij} |v_j|$$

and if inequality holds for any i, then you could sum on i and obtain

$$\sum_i |v_i| < \sum_i \sum_j a_{ij} |v_j| = \sum_j \sum_i a_{ij} |v_j| = \sum_j |v_j| ,$$

a contradiction. Therefore,

$$|v_i|^2 = \sum_j a_{ij} \operatorname{Re}(v_j \overline{\mu v_i}) = \sum_j a_{ij} |v_j| |v_i|$$

equality must hold in 12.1 for each i and so since $a_{ij} > 0$, $v_j \overline{\mu v_i}$ must be real and nonnegative for all j. In particular, for $j = i$, it follows $|v_i|^2 \overline{\mu} \geq 0$ for each i. Hence μ must be real and non negative. Thus $\mu = 1$.

If $A\mathbf{v} = \mathbf{v}$ for nonzero $\mathbf{v} \in \mathbb{R}^n$,

$$v_i = \sum_j a_{ij} v_j$$

and so

$$\begin{aligned} |v_i|^2 &= \sum_j a_{ij} v_j v_i = \sum_j a_{ij} v_j v_i \\ &\leq \sum_j a_{ij} |v_j| |v_i| . \end{aligned}$$

Dividing by $|v_i|$ and summing on i, yields

$$\sum_i |v_i| \leq \sum_i \sum_j a_{ij} |v_j| = \sum_j |v_j|$$

which shows since $a_{ij} > 0$ that for all j,

$$v_j v_i = |v_j| |v_i|$$

and so $v_j v_i$ must be real and nonnegative showing the sign of v_i is constant. This proves the lemma.

Lemma 12.1.3 *If A is any Markov matrix, there exists $\mathbf{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ with $A\mathbf{v} = \mathbf{v}$. Also, if A is a Markov matrix in which $a_{ij} > 0$ for all i, j, and*

$$X_1 \equiv \{\mathbf{x} : (A - I)^m \mathbf{x} = \mathbf{0} \text{ for some } m\} ,$$

then the dimension of $X_1 = 1$.

Proof: Let $\mathbf{u} = (1, \cdots, 1)^T$ be the vector in which there is a one for every entry. Then since A is a Markov matrix,

$$\mathbf{u}^T A = \mathbf{u}^T .$$

Therefore, $A^T \mathbf{u} = \mathbf{u}$ showing that 1 is an eigenvalue for A^T. It follows 1 must also be an eigenvalue for A since A and A^T have the same characteristic equation due to the fact the determinant of a matrix equals the determinant of its transpose. Since A is a real matrix, it follows there exists $\mathbf{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ such that $(A - I)\mathbf{v} = \mathbf{0}$. By Lemma 12.1.2, v_i has the same sign for all i. Without loss of generality assume $\sum_i v_i = 1$ and so $v_i \geq 0$ for all i.

Now suppose A is a Markov matrix in which $a_{ij} > 0$ for all i, j and suppose $\mathbf{w} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ satisfies $A\mathbf{w} = \mathbf{w}$. Then some $w_p \neq 0$ and equality holds in 12.1. Therefore,

$$w_j \overline{w_p} \equiv r_j \geq 0.$$

Then letting $\mathbf{r} \equiv (r_1, \cdots, r_n)^T$,

$$A\mathbf{r} = A\mathbf{w}\overline{w_p} = \mathbf{w}\overline{w_p} = \mathbf{r}.$$

Now defining $||\mathbf{r}||_1 \equiv \sum_i |r_i|$, $\sum_i \left(\frac{r_i}{||\mathbf{r}||_1}\right) = 1$. Also,

$$A\left(\frac{\mathbf{r}}{||\mathbf{r}||_1} - \mathbf{v}\right) = \frac{\mathbf{r}}{||\mathbf{r}||_1} - \mathbf{v}$$

and so, since all eigenvectors for $\lambda = 1$ have all entries the same sign, and

$$\sum_i \left(\frac{r_i}{||\mathbf{r}||_1} - v_i\right) = 1 - 1 = 0,$$

it follows that for all i,

$$\frac{r_i}{||\mathbf{r}||_1} = v_i$$

and so $\frac{\mathbf{r}}{||\mathbf{r}||_1} = \frac{\mathbf{w}\overline{w_p}}{||\mathbf{r}||_1} = \mathbf{v}$ showing that

$$\mathbf{w} = \frac{||\mathbf{r}||_1}{\overline{w_p}}\mathbf{v}.$$

This shows that all eigenvectors for the eigenvalue 1 are multiples of the single eigenvector, $\mathbf{v}$, described above.

Now suppose that

$$(A - I)\,\mathbf{w} = \mathbf{z}$$

where $\mathbf{z}$ is an eigenvector. Then from what was just shown, $\mathbf{z} = \alpha\mathbf{v}$ where $v_j \geq 0$ for all j, and $\sum_j v_j = 1$. It follows that

$$\sum_j a_{ij}w_j - w_i = z_i = \alpha v_i.$$

Then summing on i,

$$\sum_j w_j - \sum_i w_i = 0 = \alpha \sum_i v_i.$$

But $\sum_i v_i = 1$. Therefore, $\alpha = 0$ and so $\mathbf{z}$ is not an eigenvector. Therefore, if $(A - I)^2\,\mathbf{w} = \mathbf{0}$, it follows $(A - I)\,\mathbf{w} = \mathbf{0}$ and so in fact,

$$X_1 = \{\mathbf{w} : (A - I)\,\mathbf{w} = \mathbf{0}\}$$

and this was just shown to be one dimensional. This proves the lemma.

The following lemma is fundamental to what follows.

Lemma 12.1.4 *Let A be a Markov matrix in which $a_{ij} > 0$ for all i, j. Then there exists a basis for $\mathbb{C}^n$ such that with respect to this basis, the matrix for A is the upper triangular, block diagonal matrix,*

$$T = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & T_1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & T_r \end{pmatrix}$$

where T_s is an upper triangular matrix of the form

$$T_s = \begin{pmatrix} \mu_s & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mu_s \end{pmatrix}$$

where $|\mu_s| < 1$.

Proof: This follows from Lemma 12.1.3 and Corollary 11.4.4. The assertion about $|\mu_s|$ follows from Lemma 12.1.2.

Lemma 12.1.5 *Let A be any Markov matrix and let $\mathbf{v}$ be a vector having all its components non negative and having $\sum_i v_i = 1$. Then if $\mathbf{w} = A\mathbf{v}$, it follows $w_i \geq 0$ for all i and $\sum_i w_i = 1$.*

Proof: From the definition of $\mathbf{w}$,

$$w_i \equiv \sum_j a_{ij} v_j \geq 0.$$

Also

$$\sum_i w_i = \sum_i \sum_j a_{ij} v_j = \sum_j \sum_i a_{ij} v_j = \sum_j v_j = 1.$$

The following theorem, a special case of the Perron Frobenius theorem can now be proved.

Theorem 12.1.6 *Suppose A is a Markov matrix in which $a_{ij} > 0$ for all i, j and suppose $\mathbf{w}$ is a vector. Then for each i,*

$$\lim_{k\to\infty} \left(A^k \mathbf{w}\right)_i = v_i$$

where $A\mathbf{v} = \mathbf{v}$. In words, $A^k\mathbf{w}$ always converges to a steady state. In addition to this, if the vector, $\mathbf{w}$ satisfies $w_i \geq 0$ for all i and $\sum_i w_i = 1$, Then the vector, $\mathbf{v}$ will also satisfy the conditions, $v_i \geq 0$, $\sum_i v_i = 1$.

Proof: There exists a matrix, S such that

$$A = S^{-1} T S$$

where T is defined above. Therefore,

$$A^k = S^{-1} T^k S$$

By Lemma 11.4.5, the components of the matrix, A^k converge to the components of the matrix

$$S^{-1} U S$$

where U is an $n \times n$ matrix of the form

$$U = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix},$$

a matrix with a one in the upper left corner and zeros elsewhere. It follows that there exists a vector $\mathbf{v} \in \mathbb{R}^n$ such that

$$\lim_{k\to\infty} \left(A^k \mathbf{w}\right)_i = v_i.$$

and

$$v_i = \lim_{k\to\infty} \left(AA^k \mathbf{w}\right)_i = \lim_{k\to\infty} \sum_j a_{ij} \left(A^k \mathbf{w}\right)_j = \sum_j a_{ij} v_j = (A\mathbf{v})_i$$

Since i is arbitrary, this implies $A\mathbf{v} = \mathbf{v}$ as claimed.

Now if $w_i \geq 0$ and $\sum_i w_i = 1$, then by Lemma 12.1.5, $\left(A^k \mathbf{w}\right)_i \geq 0$ and

$$\sum_i \left(A^k \mathbf{w}\right)_i = 1.$$

Therefore, $\mathbf{v}$ also satisfies these conditions. This proves the theorem.

The following corollary is the fundamental result of this section. This corollary is a simple consequence of the following interesting lemma.

Definition 12.1.7 *Let μ be an eigenvalue of an $n \times n$ matrix, A. The generalized eigenspace equals*

$$X \equiv \{\mathbf{x} : (A - \mu I)^m \mathbf{x} = \mathbf{0} \text{ for some } m \in \mathbb{N}\}$$

Lemma 12.1.8 *Let A be an $n \times n$ matrix having distinct eigenvalues $\{\mu_1, \cdots, \mu_r\}$. Then letting X_i denote the generalized eigenspace corresponding to μ_i,*

$$X_i \subseteq X_i^k$$

where

$$X_i^k \equiv \left\{\mathbf{x} : \left(A^k - \mu_i^k I\right)^m \mathbf{x} = \mathbf{0} \text{ for some } m \in \mathbb{N}\right\}$$

Proof: Let $\mathbf{x} \in X_i$ so that $(A - \mu_i I)^m \mathbf{x} = \mathbf{0}$ for some positive integer, m. Then multiplying both sides by

$$\left(A^{k-1} + \mu_i A \cdots + \mu_i^{k-2} A + \mu_i^{k-1} I\right)^m,$$

it follows $\left(A^k - \mu_i^k I\right)^m \mathbf{x} = \mathbf{0}$ showing that $\mathbf{x} \in X_i^k$ as claimed.

Corollary 12.1.9 *Suppose A is a regular Markov matrix. Then the conclusions of Theorem 12.1.6 holds.*

Proof: In the proof of Theorem 12.1.6 the only thing needed was that A was similar to an upper triangular, block diagonal matrix of the form described in Lemma 12.1.4. This corollary is proved by showing that A is similar to such a matrix. From the assumption that A is regular, some power of A, say A^k is similar to a matrix of this form, having a one in the upper left position and having the diagonal blocks of the form described in Lemma 12.1.4 where the diagonal entries on these blocks have absolute value less than one. Now observe that if A and B are two Markov matrices such that the entries of A are all positive, then AB is also a Markov matrix having all positive entries. Thus A^{k+r} is a Markov matrix having all positive entries for every $r \in \mathbb{N}$. Therefore, each of these Markov matrices has 1 as an eigenvalue and the generalized eigenspace associated with 1 is of dimension 1. By Lemma 12.1.3, 1 is an eigenvalue for A. By Lemma 12.1.8 and Lemma 12.1.3, the generalized eigenspace for 1 is of dimension 1. If μ is an eigenvalue of A, then it is clear that μ^{k+r} is an eigenvalue for A^{k+r} and since these are all Markov matrices having all positive entries, Lemma 12.1.2 implies that for all $r \in \mathbb{N}$, either $\mu^{k+r} = 1$ or $\left|\mu^{k+r}\right| < 1$. Therefore, since r is arbitrary, it follows that either $\mu = 1$ or in the case that $\left|\mu^{k+r}\right| < 1$, $|\mu| < 1$. Therefore, A is similar to an upper triangular matrix described in Lemma 12.1.4 and this proves the corollary.

12.2 Migration Matrices

Definition 12.2.1 *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 $\sum_i a_{ij} = 1$. Thus (a_{ij}) is a Markov matrix referred to as a migration matrix.*

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 $\left(A^k\mathbf{v}\right)_i$. Furthermore, Corollary 12.1.9 can be used to predict in the case where A is regular what the long time population will be for the given locations.

As an example of the above, consider the case where $n = 3$ and the migration matrix is of the form

$$\begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix}.$$

Now

$$\begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix}^2 = \begin{pmatrix} .38 & .02 & .15 \\ .28 & .64 & .02 \\ .34 & .34 & .83 \end{pmatrix}$$

and so the Markov matrix is regular. Therefore, $\left(A^k\mathbf{v}\right)_i$ will converge to the i^{th} component of a steady state. It follows the steady state can be obtained from solving the system

$$\begin{aligned} .6x + .1z &= x \\ .2x + .8y &= y \\ .2x + .2y + .9z &= z \end{aligned}$$

along with the stipulation that the sum of x, y, and z must equal the constant value present at the beginning of the process. The solution to this system is

$$\{y = x, z = 4x, x = x\}.$$

If the total population at the beginning is 100,000, then you solve the following system

$$\begin{aligned} y &= x \\ z &= 4x \\ x + y + z &= 150000 \end{aligned}$$

whose solution is easily seen to be $\{x = 25\,000, z = 100\,000, y = 25\,000\}$. Thus, after a long time there would be about four times as many people in the third location as in either of the other two.

12.3 Markov Chains

A random variable is just a function which can have certain values which have probabilities associated with them. Thus it makes sense to consider the probability the random variable has a certain value or is in some set. The idea of a Markov chain is a sequence of random variables, $\{X_n\}$ which can be in any of a collection of states which can be labeled with nonnegative integers. Thus you can speak of the probability the random variable, X_n is in

state i. The probability that X_{n+1} is in state j given that X_n is in state i is called a one step transition probability. When this probability does not depend on n it is called stationary and this is the case of interest here. Since this probability does not depend on n it can be denoted by p_{ij}. Here is a simple example called a random walk.

Example 12.3.1 *Let there be n points, x_i, and consider a process of something moving randomly from one point to another. Suppose X_n is a sequence of random variables which has values $\{1, 2, \cdots, n\}$ where $X_n = i$ indicates the process has arrived at the i^{th} point. Let p_{ij} be the probability that X_{n+1} has the value j given that X_n has the value i. Since X_{n+1} must have some value, it must be the case that $\sum_j a_{ij} = 1$. Note this says the sum over a row equals 1 and so the situation is a little different than the above in which the sum was over a column.*

As an example, let x_1, x_2, x_3, x_4 be four points taken in order on $\mathbb{R}$ and suppose x_1 and x_4 are absorbing. This means that $p_{4k} = 0$ for all $k \neq 4$ and $p_{1k} = 0$ for all $k \neq 1$. Otherwise, you can move either to the left or to the right with probability $\frac{1}{2}$. The Markov matrix associated with this situation is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ .5 & 0 & .5 & 0 \\ 0 & .5 & 0 & .5 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Definition 12.3.2 *Let the stationary transition probabilities, p_{ij} be defined above. The resulting matrix having p_{ij} as its ij^{th} entry is called the matrix of transition probabilities. The sequence of random variables for which these p_{ij} are the transition probabilities is called a Markov chain. The matrix of transition probabilities is called a Stochastic matrix.*

The next proposition is fundamental and shows the significance of the powers of the matrix of transition probabilities.

Proposition 12.3.3 *Let p_{ij}^n denote the probability that X_n is in state j given that X_1 was in state i. Then p_{ij}^n is the ij^{th} entry of the matrix, P^n where $P = (p_{ij})$.*

Proof: This is clearly true if $n = 1$ and follows from the definition of the p_{ij}. Suppose true for n. Then the probability that X_{n+1} is at j given that X_1 was at i equals $\sum_k p_{ik}^n p_{kj}$ because X_n must have some value, k, and so this represents all possible ways to go from i to j. You can go from i to 1 in n steps with probability p_{i1} and then from 1 to j in one step with probability p_{1j} and so the probability of this is $p_{i1}^n p_{1j}$ but you can also go from i to 2 and then from 2 to j and from i to 3 and then from 3 to j etc. Thus the sum of these is just what is given and represents the probability of X_{n+1} having the value j given X_1 has the value i.

In the above random walk example, lets take a power of the transition probability matrix to determine what happens. Rounding off to two decimal places,

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ .5 & 0 & .5 & 0 \\ 0 & .5 & 0 & .5 \\ 0 & 0 & 0 & 1 \end{pmatrix}^{20} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ .67 & 9.5 \times 10^{-7} & 0 & .33 \\ .33 & 0 & 9.5 \times 10^{-7} & .67 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Thus p_{21} is about 2/3 while p_{32} is about 1/3 and terms like p_{22} are very small. You see this seems to be converging to the matrix,

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

After many iterations of the process, if you start at 2 you will end up at 1 with probability 2/3 and at 4 with probability 1/3. This makes good intuitive sense because it is twice as far from 2 to 4 as it is from 2 to 1.

What theorems can be proved about limits of powers of such matrices? Recall the following definition of terms.

Definition 12.3.4 *Let A be an $n \times n$ matrix. Then*

$$\ker(A) \equiv \{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}.$$

If λ is an eigenvalue, the eigenspace associated with λ is defined as $\ker(A - \lambda I)$ and the generalized eigenspace is given by

$$\{\mathbf{x} : (A - \lambda I)^m \mathbf{x} = \mathbf{0} \text{ for some } m \in \mathbb{N}\}.$$

It is clear that $\ker(A)$ is a subspace of $\mathbb{F}^n$ so has a well defined dimension and also it is a subspace of the generalized eigenspace for λ.

Lemma 12.3.5 *Suppose A is an $n \times n$ matrix and there exists an invertible matrix, S such that $S^{-1}AS = T$. Then if there exist m linearly independent eigenvectors for A associated with the eigenvalue, λ, it follows there exist m linearly independent eigenvectors for T associated with the eigenvalue, λ.*

Proof: Suppose the independent set of eigenvectors for A are $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$. Then consider $\left\{S^{-1}\mathbf{v}_1, \cdots, S^{-1}\mathbf{v}_m\right\}$.

$$A\mathbf{v}_k = STS^{-1}\mathbf{v}_k = \lambda\mathbf{v}_k$$

and so

$$TS^{-1}\mathbf{v}_k = \lambda S^{-1}\mathbf{v}_k.$$

Therefore, $\left\{S^{-1}\mathbf{v}_1, \cdots, S^{-1}\mathbf{v}_m\right\}$ are a set of eigenvectors. Suppose

$$\sum_{k=1}^{m} c_k S^{-1}\mathbf{v}_k = \mathbf{0}.$$

Then

$$S^{-1}\left(\sum_{k=1}^{m} c_k \mathbf{v}_k\right) = \mathbf{0}$$

and since S^{-1} is one to one, it follows $\sum_{k=1}^{m} c_k\mathbf{v}_k = \mathbf{0}$ which requires each $c_k = 0$ due to the linear independence of the $\mathbf{v}_k$.

Lemma 12.3.6 *Suppose λ is an eigenvalue for an $n \times n$ matrix. Then the dimension of the generalized eigenspace for λ equals the algebraic multiplicity of λ_1 as a root of the characteristic equation. If the algebraic multiplicity of the eigenvalue as a root of the characteristic equation equals the dimension of the eigenspace, then the eigenspace equals the generalized eigenspace.*

Proof: By Corollary 11.4.4 on Page 201 there exists an invertible matrix, S such that

$$S^{-1}AS = T = \begin{pmatrix} T_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & T_{r-1} & 0 \\ 0 & \cdots & 0 & T_r \end{pmatrix}$$

where the only eigenvalue of T_k is λ_k and the dimension of the generalized eigenspace for λ_k equals m_k where T_k is an $m_k \times m_k$ matrix. However, the algebraic multiplicity of λ_k is also equal to m_k because both T and A have the same characteristic equation and the characteristic equation for T is of the form $\prod_{i=1}^{r} (\lambda - \lambda_k)^{m_k}$. This proves the first part of the lemma. The second follows immediately because the eigenspace is a subspace of the generalized eigenspace and so if they have the same dimension, they must be equal.[1]

Theorem 12.3.7 *Suppose for all λ an eigenvalue of A either $|\lambda| < 1$ or $\lambda = 1$ and that the dimension of the eigenspace for $\lambda = 1$ equals the algebraic multiplicity of 1 as an eigenvalue of A. Then $\lim_{p\to\infty} A^p$ exists[2]. If for all eigenvalues, λ, it is the case that $|\lambda| < 1$, then $\lim_{p\to\infty} A^p = 0$.*

Proof: From Corollary 11.4.4 on Page 201 there exists an invertible matrix, S such that

$$S^{-1}AS = T = \begin{pmatrix} T_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & T_{r-1} & 0 \\ 0 & \cdots & 0 & T_r \end{pmatrix} \tag{12.4}$$

where T_1 has only 1 on the main diagonal and the matrices, T_k, have only λ_k on the main diagonal where $|\lambda_k| < 1$. Letting m_1 denote the size of T_1, it follows the multiplicity of 1 as an eigenvalue equals m_1 and it is assumed the dimension of $\ker(A - I)$ equals m_1. From the assumption, there exist m_1 linearly independent eigenvectors of A corresponding to $\lambda = 1$. Therefore, there exist m_1 linearly independent eigenvectors for T corresponding to $\lambda = 1$. If $\mathbf{v}$ is one of these eigenvectors, then

$$\mathbf{v} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_r \end{pmatrix}$$

where the $\mathbf{a}_k$ are conformable with the matrices T_k. Therefore,

$$T\mathbf{v} = \begin{pmatrix} \mathbf{a}_1 \\ \lambda_2\mathbf{a}_2 \\ \vdots \\ \lambda_r\mathbf{a}_r \end{pmatrix} = 1 \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_r \end{pmatrix}$$

and so $\mathbf{v}$ must actually be of the form

$$\mathbf{v} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}.$$

It follows there exists a basis of eigenvectors for the matrix T_1 in $\mathbb{C}^{m_1}, \{\mathbf{u}_1, \cdots, \mathbf{u}_{m_1}\}$. Define the $m_1 \times m_1$ matrix, M by

$$M = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_{m_1} \end{pmatrix}$$

[1] Any basis for the eigenspace must be a basis for the generalized eigenspace because if not, you could include a vector from the generalized eigenspace which is not in the eigenspace and the resulting list would be linearly independent, showing the dimension of the generalized eigenspace is larger than the dimension of the eigenspace contrary to the assertion that the two have the same dimension.

[2] The converse of this theorem also is true. You should try to prove the converse.

where the $\mathbf{u}_k$ are the columns of this matrix. Then

$$\begin{aligned} M^{-1}T_1M &= M^{-1}\begin{pmatrix} T_1\mathbf{u}_1 & \cdots & T_1\mathbf{u}_{m_1} \end{pmatrix} \\ &= M^{-1}\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_{m_1} \end{pmatrix} = M^{-1}M = I. \end{aligned}$$

Now let P denote the block matrix given as

$$P = \begin{pmatrix} M & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & I & 0 \\ 0 & \cdots & 0 & I \end{pmatrix}$$

so that

$$P^{-1} = \begin{pmatrix} M^{-1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & I & 0 \\ 0 & \cdots & 0 & I \end{pmatrix}$$

and let S denote the $n \times n$ matrix such that $S^{-1}AS = T$. Then $P^{-1}S^{-1}ASP$ must be of the form

$$\begin{aligned} P^{-1}S^{-1}ASP &= (SP)^{-1}ASP = \\ G &= \begin{pmatrix} I & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & T_{r-1} & 0 \\ 0 & \cdots & 0 & T_r \end{pmatrix}. \end{aligned}$$

Therefore,

$$A^m = \left(SPG\left(SP\right)^{-1}\right)^m = SPG^m\left(SP\right)^{-1}$$

and by Lemma 11.4.5 on Page 203 the entries of the matrix,

$$G^m = \begin{pmatrix} I & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & T_{r-1}^m & 0 \\ 0 & \cdots & 0 & T_r^m \end{pmatrix}$$

converge to the entries of the matrix,

$$L \equiv \begin{pmatrix} I & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix}$$

and so the entries of the matrix, A^m converge to the entries of the matrix

$$SPL\left(SP\right)^{-1}.$$

The last claim also follows since in this case, the matrices, T_k in 12.4 all have diagonal entries whose absolute values are less than 1. This proves the theorem.

Corollary 12.3.8 *Let A be an $n \times n$ matrix with the property that whenever λ is an eigenvalue of A, either $|\lambda| < 1$ or $\lambda = 1$ and the dimension of the eigenspace for $\lambda = 1$ equals the algebraic multiplicity of 1 as an eigenvalue of A. Suppose also that a basis for the eigenspace of $\lambda = 1$ is $D_1 = \{\mathbf{u}_1, \cdots, \mathbf{u}_m\}$ and a basis for the generalized eigenspace for λ_k with $|\lambda_k| < 1$ is D_k where $k = 1, \cdots, r$. Then letting $\mathbf{x}$ be any given vector, there exists a vector, $\mathbf{v} \in \text{span}(D_2, \cdots, D_r)$ and scalars, c_i such that*

$$\mathbf{x} = \sum_{i=1}^{m} c_i \mathbf{u}_i + \mathbf{v} \tag{12.5}$$

and

$$\lim_{k \to \infty} A^k \mathbf{x} = \mathbf{y} \tag{12.6}$$

where

$$\mathbf{y} = \sum_{i=1}^{m} c_i \mathbf{u}_i. \tag{12.7}$$

Proof: The first claim follows from Corollary 11.4.3 on Page 201. By Theorem 12.3.7, the limit in 12.6 exists and letting $\mathbf{y}$ be this limit, it follows

$$\mathbf{y} = \lim_{k \to \infty} A^{k+1} \mathbf{x} = A \lim_{k \to \infty} A^k \mathbf{x} = A\mathbf{y}.$$

Therefore, there exist constants, c_i' such that

$$\mathbf{y} = \sum_{i=1}^{m} c_i' \mathbf{u}_i.$$

Are these constants the same as the c_i? This will be true if $A^k \mathbf{v} \to \mathbf{0}$. But A has only eigenvalues which have absolute value less than 1 on $\text{span}(D_2, \cdots, D_r)$ and so the same is true of a matrix for A relative to the basis $\{D_2, \cdots, D_r\}$. Therefore, the desired result follows from Theorem 12.3.7.

Example 12.3.9 *In the gambler's ruin problem a gambler plays a game with someone, say a casino, until he either wins all the other's money or loses all of his own. A simple version of this is as follows. Let X_k denote the amount of money the gambler has. Each time the game is played he wins with probability $p \in (0,1)$ or loses with probability $(1-p) \equiv q$. In case he wins, his money increases to $X_k + 1$ and if he loses, his money decreases to $X_k - 1$.*

The transition probability matrix, P, describing this situation is as follows.

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ q & 0 & p & 0 & \cdots & 0 & 0 \\ 0 & q & 0 & p & \cdots & 0 & \vdots \\ 0 & 0 & q & 0 & \ddots & \vdots & 0 \\ \vdots & \vdots & 0 & \ddots & \ddots & p & 0 \\ 0 & 0 & \vdots & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{12.8}$$

Here the matrix is $b+1 \times b+1$ because the possible values of X_k are all integers from 0 up to b. The 1 in the upper left corner corresponds to the gampler's ruin. It involves $X_k = 0$

so he has no money left. Once this state has been reached, it is not possible to ever leave it, even if you do have a great positive attitude. This is indicated by the row of zeros to the right of this entry the k^{th} of which gives the probability of going from state 1 corresponding to no money to state k[3].

In this case 1 is a repeated root of the characteristic equation of multiplicity 2 and all the other eigenvalues have absolute value less than 1. To see that this is the case, note the characteristic polynomial is of the form

$$(1-\lambda)^2 \det \begin{pmatrix} -\lambda & p & 0 & \cdots & 0 \\ q & -\lambda & p & \cdots & 0 \\ 0 & q & -\lambda & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & p \\ 0 & \vdots & 0 & q & -\lambda \end{pmatrix}$$

and the factor after $(1-\lambda)^2$ has only zeros which are in absolute value less than 1. (See Problem 4 on Page 240.) It is also obvious that both

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ \vdots \\ 0 \end{pmatrix}, \text{ and } \mathbf{e}_n = \begin{pmatrix} 0 \\ \vdots \\ 1 \end{pmatrix}$$

are eigenvectors which correspond to $\lambda = 1$ and so the dimension of the eigenspace equals the multiplicity of the eigenvalue. Therefore, from Theorem 12.3.7 $\lim_{n\to\infty} p_{ij}^n$ exists for every i, j. The case of $\lim_{n\to\infty} p_{j1}^n$ is particularly interesting because it gives the probability that, starting with an amount j, the gambler is eventually ruined. From Proposition 12.3.3 and 12.8,

$$\begin{aligned} p_{j1}^n &= qp_{(j-1)1}^{n-1} + pp_{(j+1)1}^{n-1} \text{ for } j \in [2,b], \\ p_{11}^n &= 1, \text{ and } p_{(b+1)1}^n = 0. \end{aligned}$$

To simplify the notation, define $P_j \equiv \lim_{n\to\infty} p_{j1}^n$ as the probability of ruin given the initial fortune of the gambler equals j. Then the above simplifies to

$$\begin{aligned} P_j &= qP_{j-1} + pP_{j+1} \text{ for } j \in [2,b], \\ P_1 &= 1, \text{ and } P_{b+1} = 0. \end{aligned} \tag{12.9}$$

Now, knowing that P_j exists, it is not too hard to find it from 12.9. This equation is called a difference equation and there is a standard procedure for finding solutions of these. You try a solution of the form $P_j = x^j$ and then try to find x such that things work out. Therefore, substitute this in to the first equation of 12.9 and obtain

$$x^j = qx^{j-1} + px^{j+1}.$$

Therefore,

$$px^2 - x + q = 0$$

[3]No one will give the gambler money. This is why the only reasonable number for entries in this row to the right of 1is 0.

and so in case $p \neq q$, you can use the fact that $p+q=1$ to obtain

$$\begin{aligned} x &= \frac{1}{2p}\left(1+\sqrt{(1-4pq)}\right) \text{ or } \frac{1}{2p}\left(1-\sqrt{(1-4pq)}\right) \\ &= \frac{1}{2p}\left(1+\sqrt{(1-4p(1-p))}\right) \text{ or } \frac{1}{2p}\left(1-\sqrt{(1-4p(1-p))}\right) \\ &= 1 \text{ or } \frac{q}{p}. \end{aligned}$$

Now it follows that both $P_j = 1$ and $P_j = \left(\frac{q}{p}\right)^j$ satisfy the first equation of 12.9. Therefore, anything of the form

$$\alpha + \beta\left(\frac{q}{p}\right)^j \tag{12.10}$$

will satisfy this equation. Now find a, b such that this also satisfies the second equation of 12.9. Thus it is required that

$$\alpha + \beta\left(\frac{q}{p}\right) = 1,\ \alpha + \beta\left(\frac{q}{p}\right)^{b+1} = 0$$

and so

$$\beta = \frac{p}{-\left(\frac{q}{p}\right)^{b+1}p+q},\ \alpha = -\frac{p}{-\left(\frac{q}{p}\right)^{b+1}p+q}\left(\frac{q}{p}\right)^{b+1}.$$

Substituting this in to 12.10 and simplifying yields the following in the case that $p \neq q$.

$$P_j = \frac{q^{-1+j}p^{b+1-j} - q^b}{p^b - q^b}. \tag{12.11}$$

Next consider the case where $p = q = 1/2$. In this case, you can see that a solution to 12.9 is

$$P_j = \frac{b+1-j}{b}. \tag{12.12}$$

This last case is pretty interesting because it shows, for example that if the gambler starts with a fortune of 1 so that he starts at state $j=2$, then his probability of losing all is $\frac{b-1}{b}$ which might be quite large especially if the other player has a lot of money to begin with. As the gambler starts with more and more money, his probability of losing everything does decrease.

See the book by Karlin and Taylor for more on this sort of thing [9].

12.4 Exercises

1. Suppose $B \in \mathcal{L}(X,X)$ where X is a finite dimensional vector space and $B^m = 0$ for some m a positive integer. Letting $v \in X$, consider the string of vectors, $v, Bv, B^2v, \cdots, B^kv$. Show this string of vectors is a linearly independent set if and only if $B^kv \neq 0$.

2. Suppose the migration matrix for three locations is

$$\begin{pmatrix} .5 & 0 & .3 \\ .3 & .8 & 0 \\ .2 & .2 & .7 \end{pmatrix}.$$

Find a comparison for the populations in the three locations after a long time.

3. For any $n \times n$ matrix, why is the dimension of the eigenspace always less than or equal to the algebraic multiplicity of the eigenvalue as a root of the characteristic equation? **Hint:** See the proof of Theorem 12.3.7. Note the algebraic multiplicity is the size of the appropriate block in the matrix, T and that the eigenvectors of T must have a certain simple form as in the proof of that theorem.

4. Consider the following $m \times m$ matrix in which $p+q=1$ and both p and q are positive numbers.

$$\begin{pmatrix} 0 & p & 0 & \cdots & 0 \\ q & 0 & p & \cdots & 0 \\ 0 & q & 0 & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & p \\ 0 & \vdots & 0 & q & 0 \end{pmatrix}$$

Show if $\mathbf{x} = (x_1, \cdots, x_m)$ is an eigenvector, then the $|x_i|$ cannot be constant. Using this show that if μ is an eigenvalue, it must be the case that $|\mu| < 1$. **Hint:** To verify the first part of this, use Gerschgorin's theorem to observe that if λ is an eigenvalue, $|\lambda| \leq 1$. To verify this last part, there must exist i such that $|x_i| = \max\{|x_j| : j = 1, \cdots, m\}$ and either $|x_{i-1}| < |x_i|$ or $|x_{i+1}| < |x_i|$. Then consider what it means to say that $A\mathbf{x} = \mu\mathbf{x}$.

Inner Product Spaces

The usual example of an inner product space is $\mathbb{C}^n$ or $\mathbb{R}^n$ with the dot product. However, there are many other inner product spaces and the topic is of such importance that it seems appropriate to discuss the general theory of these spaces.

Definition 13.0.1 *A vector space X is said to be a normed linear space if there exists a function, denoted by $|\cdot| : X \to [0,\infty)$ which satisfies the following axioms.*

1. $|x| \geq 0$ *for all* $x \in X$, *and* $|x| = 0$ *if and only if* $x = 0$.
2. $|ax| = |a|\,|x|$ *for all* $a \in \mathbb{F}$.
3. $|x+y| \leq |x| + |y|$.

The notation $||x||$ is also often used. Not all norms are created equal. There are many geometric properties which they may or may not possess. There is also a concept called an inner product which is discussed next. It turns out that the best norms come from an inner product.

Definition 13.0.2 *A mapping $(\cdot,\cdot) : V \times V \to \mathbb{F}$ is called an inner product if it satisfies the following axioms.*

1. $(x,y) = \overline{(y,x)}$.
2. $(x,x) \geq 0$ *for all* $x \in V$ *and equals zero if and only if* $x = 0$.
3. $(ax+by,z) = a\,(x,z) + b\,(y,z)$ *whenever* $a, b \in \mathbb{F}$.

Note that 2 and 3 imply $(x, ay+bz) = \overline{a}(x,y) + \overline{b}(x,z)$.
Then a norm is given by

$$(x,x)^{1/2} \equiv |x|.$$

It remains to verify this really is a norm.

Definition 13.0.3 *A normed linear space in which the norm comes from an inner product as just described is called an inner product space.*

Example 13.0.4 *Let $V = \mathbb{C}^n$ with the inner product given by*

$$(\mathbf{x},\mathbf{y}) \equiv \sum_{k=1}^{n} x_k \overline{y}_k.$$

This is an example of a complex inner product space already discussed.

Example 13.0.5 *Let* $V = \mathbb{R}^n$,

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} \equiv \sum_{j=1}^{n} x_j y_j.$$

This is an example of a real inner product space.

Example 13.0.6 *Let* V *be any finite dimensional vector space and let* $\{v_1, \cdots, v_n\}$ *be a basis. Decree that*

$$(v_i, v_j) \equiv \delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

and define the inner product by

$$(x, y) \equiv \sum_{i=1}^{n} x^i \overline{y^i}$$

where

$$x = \sum_{i=1}^{n} x^i v_i, \; y = \sum_{i=1}^{n} y^i v_i.$$

The above is well defined because $\{v_1, \cdots, v_n\}$ is a basis. Thus the components, x_i associated with any given $x \in V$ are uniquely determined.

This example shows there is no loss of generality when studying finite dimensional vector spaces in assuming the vector space is actually an inner product space. The following theorem was presented earlier with slightly different notation.

Theorem 13.0.7 *(Cauchy Schwarz) In any inner product space*

$$|(x, y)| \leq |x||y|.$$

where $|x| \equiv (x, x)^{1/2}$.

Proof: Let $\omega \in \mathbb{C}$, $|\omega| = 1$, and $\overline{\omega}(x, y) = |(x, y)| = \mathrm{Re}(x, y\omega)$. Let

$$F(t) = (x + ty\omega, x + t\omega y).$$

If $y = 0$ there is nothing to prove because

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

and so $(x, 0) = 0$. Thus, there is no loss of generality in assuming $y \neq 0$. Then from the axioms of the inner product,

$$F(t) = |x|^2 + 2t\,\mathrm{Re}(x, \omega y) + t^2|y|^2 \geq 0.$$

This yields

$$|x|^2 + 2t|(x, y)| + t^2|y|^2 \geq 0.$$

Since this inequality holds for all $t \in \mathbb{R}$, it follows from the quadratic formula that

$$4|(x, y)|^2 - 4|x|^2|y|^2 \leq 0.$$

This yields the conclusion and proves the theorem.

Earlier it was claimed that the inner product defines a norm. In this next proposition this claim is proved.

Proposition 13.0.8 *For an inner product space, $|x| \equiv (x,x)^{1/2}$ does specify a norm.*

Proof: All the axioms are obvious except the triangle inequality. To verify this,

$$\begin{aligned} |x+y|^2 &\equiv (x+y, x+y) \equiv |x|^2 + |y|^2 + 2\,\mathrm{Re}\,(x,y) \\ &\leq |x|^2 + |y|^2 + 2\,|(x,y)| \\ &\leq |x|^2 + |y|^2 + 2\,|x|\,|y| = (|x| + |y|)^2. \end{aligned}$$

The best norms of all are those which come from an inner product because of the following identity which is known as the parallelogram identity.

Proposition 13.0.9 *If $(V,(\cdot,\cdot))$ is an inner product space then for $|x| \equiv (x,x)^{1/2}$, the following identity holds.*

$$|x+y|^2 + |x-y|^2 = 2\,|x|^2 + 2\,|y|^2\,.$$

It turns out that the validity of this identity is equivalent to the existence of an inner product which determines the norm as described above. These sorts of considerations are topics for more advanced courses on functional analysis.

Definition 13.0.10 *A basis for an inner product space, $\{u_1, \cdots, u_n\}$ is an orthonormal basis if*

$$(u_k, u_j) = \delta_{kj} \equiv \begin{cases} 1 \text{ if } k = j \\ 0 \text{ if } k \neq j \end{cases}.$$

Note that if a list of vectors satisfies the above condition for being an orthonormal set, then the list of vectors is automatically linearly independent. To see this, suppose

$$\sum_{j=1}^{n} c^j u_j = 0$$

Then taking the inner product of both sides with u_k,

$$0 = \sum_{j=1}^{n} c^j\,(u_j, u_k) = \sum_{j=1}^{n} c^j \delta_{jk} = c^k.$$

Lemma 13.0.11 *Let X be a finite dimensional inner product space of dimension n whose basis is $\{x_1, \cdots, x_n\}$. Then there exists an orthonormal basis for X, $\{u_1, \cdots, u_n\}$ which has the property that for each $k \leq n$, span$(x_1, \cdots, x_k) = \mathrm{span}\,(u_1, \cdots, u_k)$.*

Proof: Let $\{x_1, \cdots, x_n\}$ be a basis for X. Let $u_1 \equiv x_1/|x_1|$. Thus for $k = 1$, $\mathrm{span}\,(u_1) = \mathrm{span}\,(x_1)$ and $\{u_1\}$ is an orthonormal set. Now suppose for some $k < n$, $u_1, \cdots, u_k$ have been chosen such that $(u_j, u_l) = \delta_{jl}$ and $\mathrm{span}\,(x_1, \cdots, x_k) = \mathrm{span}\,(u_1, \cdots, u_k)$. Then define

$$u_{k+1} \equiv \frac{x_{k+1} - \sum_{j=1}^{k} (x_{k+1}, u_j)\,u_j}{\left| x_{k+1} - \sum_{j=1}^{k} (x_{k+1}, u_j)\,u_j \right|}, \tag{13.1}$$

where the denominator is not equal to zero because the x_j form a basis and so

$$x_{k+1} \notin \mathrm{span}\,(x_1, \cdots, x_k) = \mathrm{span}\,(u_1, \cdots, u_k)$$

Thus by induction,

$$u_{k+1} \in \mathrm{span}\,(u_1, \cdots, u_k, x_{k+1}) = \mathrm{span}\,(x_1, \cdots, x_k, x_{k+1})\,.$$

Also, $x_{k+1} \in \text{span}(u_1, \cdots, u_k, u_{k+1})$ which is seen easily by solving 13.1 for x_{k+1} and it follows

$$\text{span}(x_1, \cdots, x_k, x_{k+1}) = \text{span}(u_1, \cdots, u_k, u_{k+1}).$$

If $l \leq k$,

$$\begin{aligned}(u_{k+1}, u_l) &= C\left((x_{k+1}, u_l) - \sum_{j=1}^{k} (x_{k+1}, u_j)(u_j, u_l)\right) \\ &= C\left((x_{k+1}, u_l) - \sum_{j=1}^{k} (x_{k+1}, u_j)\delta_{lj}\right) \\ &= C((x_{k+1}, u_l) - (x_{k+1}, u_l)) = 0.\end{aligned}$$

The vectors, $\{u_j\}_{j=1}^n$, generated in this way are therefore an orthonormal basis because each vector has unit length.

The process by which these vectors were generated is called the Gram Schmidt process.

Lemma 13.0.12 *Suppose $\{u_j\}_{j=1}^n$ is an orthonormal basis for an inner product space X. Then for all $x \in X$,*

$$x = \sum_{j=1}^{n} (x, u_j) u_j.$$

Proof: By assumption that this is an orthonormal basis,

$$\sum_{j=1}^{n} (x, u_j) \overbrace{(u_j, u_l)}^{\delta_{jl}} = (x, u_l).$$

Letting $y = \sum_{k=1}^n (x, u_k) u_k$, it follows

$$\begin{aligned}(x - y, u_j) &= (x, u_j) - \sum_{k=1}^{n} (x, u_k)(u_k, u_j) \\ &= (x, u_j) - (x, u_j) = 0\end{aligned}$$

for all j. Hence, for any choice of scalars, $c^1, \cdots, c^n$,

$$\left(x - y, \sum_{j=1}^{n} c^j u_j\right) = 0$$

and so $(x - y, z) = 0$ for all $z \in X$. Thus this holds in particular for $z = x - y$. Therefore, $x = y$ and this proves the theorem.

The following theorem is of fundamental importance. First note that a subspace of an inner product space is also an inner product space because you can use the same inner product.

Theorem 13.0.13 *Let M be a subspace of X, a finite dimensional inner product space and let $\{x_i\}_{i=1}^m$ be an orthonormal basis for M. Then if $y \in X$ and $w \in M$,*

$$|y - w|^2 = \inf\left\{|y - z|^2 : z \in M\right\} \tag{13.2}$$

if and only if

$$(y-w,z)=0 \tag{13.3}$$

for all $z \in M$. *Furthermore,*

$$w=\sum_{i=1}^{m}(y,x_i)\,x_i \tag{13.4}$$

is the unique element of M *which has this property.*

Proof: Let $t \in \mathbb{R}$. Then from the properties of the inner product,

$$|y-(w+t(z-w))|^2=|y-w|^2+2t\,\mathrm{Re}\,(y-w,w-z)+t^2|z-w|^2. \tag{13.5}$$

If $(y-w,z)=0$ for all $z \in M$, then letting $t=1$, the middle term in the above expression vanishes and so $|y-z|^2$ is minimized when $z=w$.

Conversely, if 13.2 holds, then the middle term of 13.5 must also vanish since otherwise, you could choose small real t such that

$$|y-w|^2>|y-(w+t(z-w))|^2.$$

Here is why. If $\mathrm{Re}\,(y-w,w-z)<0$, then let t be very small and positive. The middle term in 13.5 will then be more negative than the last term is positive and the right side of this formula will then be less than $|y-w|^2$. If $\mathrm{Re}\,(y-w,w-z)>0$ then choose t small and negative to achieve the same result.

It follows, letting $z_1=w-z$ that

$$\mathrm{Re}\,(y-w,z_1)=0$$

for all $z_1 \in M$. Now letting $\omega \in \mathbb{C}$ be such that $\omega(y-w,z_1)=|(y-w,z_1)|$,

$$|(y-w,z_1)|=(y-w,\overline{\omega}z_1)=\mathrm{Re}\,(y-w,\overline{\omega}z_1)=0,$$

which proves the first part of the theorem since z_1 is arbitrary.

It only remains to verify that w given in 13.4 satisfies 13.3 and is the only point of M which does so. To do this, note that if c_i,d_i are scalars, then the properties of the inner product and the fact the $\{x_i\}$ are orthonormal implies

$$\left(\sum_{i=1}^{m}c_ix_i,\sum_{j=1}^{m}d_jx_j\right)=\sum_i c_i\overline{d_i}.$$

By Lemma 13.0.12,

$$z=\sum_i(z,x_i)\,x_i$$

and so

$$\begin{aligned}\left(y-\sum_{i=1}^{m}(y,x_i)\,x_i,z\right)&=\left(y-\sum_{i=1}^{m}(y,x_i)\,x_i,\sum_{i=1}^{m}(z,x_i)\,x_i\right)\\&=\sum_{i=1}^{m}\overline{(z,x_i)}\,(y,x_i)-\left(\sum_{i=1}^{m}(y,x_i)\,x_i,\sum_{j=1}^{m}(z,x_j)\,x_j\right)\\&=\sum_{i=1}^{m}\overline{(z,x_i)}\,(y,x_i)-\sum_{i=1}^{m}(y,x_i)\,\overline{(z,x_i)}=0.\end{aligned}$$

This shows w given in 13.4 does minimize the function, $z \to |y-z|^2$ for $z \in M$. It only remains to verify uniqueness. Suppose than that $w_i, i = 1,2$ minimizes this function of z for $z \in M$. Then from what was shown above,

$$\begin{aligned} |y-w_1|^2 &= |y-w_2+w_2-w_1|^2 \\ &= |y-w_2|^2 + 2\,\mathrm{Re}\,(y-w_2, w_2-w_1) + |w_2-w_1|^2 \\ &= |y-w_2|^2 + |w_2-w_1|^2 \leq |y-w_2|^2, \end{aligned}$$

the last equal sign holding because w_2 is a minimizer and the last inequality holding because w_1 minimizes.

The next theorem is one of the most important results in the theory of inner product spaces. It is called the Riesz representation theorem.

Theorem 13.0.14 *Let $f \in \mathcal{L}(X, \mathbb{F})$ where X is an inner product space of dimension n. Then there exists a unique $z \in X$ such that for all $x \in X$,*

$$f(x) = (x, z).$$

Proof: First I will verify uniqueness. Suppose z_j works for $j = 1,2$. Then for all $x \in X$,

$$0 = f(x) - f(x) = (x, z_1 - z_2)$$

and so $z_1 = z_2$.

It remains to verify existence. By Lemma 13.0.11, there exists an orthonormal basis, $\{u_j\}_{j=1}^n$. Define

$$z \equiv \sum_{j=1}^{n} \overline{f(u_j)} u_j.$$

Then using Lemma 13.0.12,

$$\begin{aligned} (x,z) &= \left(x, \sum_{j=1}^{n} \overline{f(u_j)} u_j \right) = \sum_{j=1}^{n} f(u_j)(x, u_j) \\ &= f\left(\sum_{j=1}^{n} (x, u_j) u_j \right) = f(x). \end{aligned}$$

This proves the theorem.

Corollary 13.0.15 *Let $A \in \mathcal{L}(X, Y)$ where X and Y are two inner product spaces of finite dimension. Then there exists a unique $A^* \in \mathcal{L}(Y, X)$ such that*

$$(Ax, y)_Y = (x, A^* y)_X \tag{13.6}$$

for all $x \in X$ and $y \in Y$. The following formula holds

$$(\alpha A + \beta B)^* = \overline{\alpha} A^* + \overline{\beta} B^*$$

Proof: Let $f_y \in \mathcal{L}(X, \mathbb{F})$ be defined as

$$f_y(x) \equiv (Ax, y)_Y.$$

Then by the Riesz representation theorem, there exists a unique element of X, $A^*(y)$ such that

$$(Ax, y)_Y = (x, A^*(y))_X.$$

It only remains to verify that A^* is linear. Let a and b be scalars. Then for all $x \in X$,

$$(x, A^*(ay_1 + by_2))_X \equiv \overline{a}(Ax, y_1) + \overline{b}(Ax, y_2)$$

$$\overline{a}(x, A^*(y_1)) + \overline{b}(x, A^*(y_2)) = (x, aA^*(y_1) + bA^*(y_2)).$$

By uniqueness, $A^*(ay_1 + by_2) = aA^*(y_1) + bA^*(y_2)$ which shows A^* is linear as claimed. The last assertion about the map which sends a linear transformation, A to A^*follows from

$$(x, A^*y + A^*y) = (Ax, y) + (Bx, y) = ((A+B)x, y) \equiv \left(x, (A+B)^* y\right)$$

and for α a scalar,

$$\left(x, (\alpha A)^* y\right) = (\alpha Ax, y) = \alpha(x, A^*y) = (x, \overline{\alpha} A^* y).$$

This proves the corollary.

The linear map, A^* is called the adjoint of A. In the case when $A : X \to X$ and $A = A^*$, A is called a self adjoint map.

Theorem 13.0.16 *Let M be an $m \times n$ matrix. Then $M^* = \left(\overline{M}\right)^T$ in words, the transpose of the conjugate of M is equal to the adjoint.*

Proof: Using the definition of the inner product in $\mathbb{C}^n$,

$$(M\mathbf{x}, \mathbf{y}) = (\mathbf{x}, M^*\mathbf{y}) \equiv \sum_i x_i \overline{\sum_j (M^*)_{ij} y_j} = \sum_{i,j} x_i \overline{(M^*)_{ij}} \overline{y_j}.$$

Also

$$(M\mathbf{x}, \mathbf{y}) = \sum_j \sum_i M_{ji} x_i \overline{y_j}.$$

Since $\mathbf{x}, \mathbf{y}$ are arbitrary vectors, it follows that $M_{ji} = \overline{(M^*)_{ij}}$ and so, taking conjugates of both sides,

$$M^*_{ij} = \overline{M_{ji}}$$

which gives the conclusion of the theorem.

The next theorem is interesting.

Theorem 13.0.17 *Suppose V is a subspace of $\mathbb{F}^n$ having dimension $p \leq n$. Then there exists a $Q \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ such that $QV \subseteq \mathbb{F}^p$ and $|Q\mathbf{x}| = |\mathbf{x}|$ for all $\mathbf{x}$. Also*

$$Q^*Q = QQ^* = I.$$

Proof: By Lemma 13.0.11 there exists an orthonormal basis for V, $\{\mathbf{v}_i\}_{i=1}^p$. By using the Gram Schmidt process this may be extended to an orthonormal basis of the whole space, $\mathbb{F}^n$,

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_p, \mathbf{v}_{p+1}, \cdots, \mathbf{v}_n\}.$$

Now define $Q \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ by $Q(\mathbf{v}_i) \equiv \mathbf{e}_i$ and extend linearly. If $\sum_{i=1}^n x_i \mathbf{v}_i$ is an arbitrary element of $\mathbb{F}^n$,

$$\left| Q\left(\sum_{i=1}^n x_i \mathbf{v}_i\right)\right|^2 = \left|\sum_{i=1}^n x_i \mathbf{e}_i\right|^2 = \sum_{i=1}^n |x_i|^2 = \left|\sum_{i=1}^n x_i \mathbf{v}_i\right|^2.$$

It remains to verify that $Q^*Q = QQ^* = I$. To do so, let $\mathbf{x}, \mathbf{y} \in \mathbb{F}^p$. Then

$$(Q(\mathbf{x}+\mathbf{y}), Q(\mathbf{x}+\mathbf{y})) = (\mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}).$$

Thus

$$|Q\mathbf{x}|^2 + |Q\mathbf{y}|^2 + 2\,\mathrm{Re}\,(Q\mathbf{x},Q\mathbf{y}) = |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\,\mathrm{Re}\,(\mathbf{x},\mathbf{y})$$

and since Q preserves norms, it follows that for all $\mathbf{x},\mathbf{y} \in \mathbb{F}^n$,

$$\mathrm{Re}\,(Q\mathbf{x},Q\mathbf{y}) = \mathrm{Re}\,(\mathbf{x},Q^*Q\mathbf{y}) = \mathrm{Re}\,(\mathbf{x},\mathbf{y})\,.$$

Therefore, since this holds for all $\mathbf{x}$, it follows that $Q^*Q\mathbf{y} = \mathbf{y}$ showing $Q^*Q = I$. Now

$$Q = Q\,(Q^*Q) = (QQ^*)\,Q.$$

Since Q is one to one, this implies

$$I = QQ^*$$

and proves the theorem.

Definition 13.0.18 *Let X and Y be inner product spaces and let $x \in X$ and $y \in Y$. Define the tensor product of these two vectors, $y \otimes x$, an element of $\mathcal{L}(X,Y)$ by*

$$y \otimes x\,(u) \equiv y\,(u,x)_X\,.$$

This is also called a rank one transformation because the image of this transformation is contained in the span of the vector, y.

The verification that this is a linear map is left to you. Be sure to verify this! The following lemma has some of the most important properties of this linear transformation.

Lemma 13.0.19 *Let X,Y,Z be inner product spaces. Then for α a scalar,*

$$(\alpha\,(y \otimes x))^* = \overline{\alpha}x \otimes y \tag{13.7}$$

$$(z \otimes y_1)\,(y_2 \otimes x) = (y_2,y_1)\,z \otimes x \tag{13.8}$$

Proof: Let $u \in X$ and $v \in Y$. Then

$$(\alpha\,(y \otimes x)\,u,v) = (\alpha\,(u,x)\,y,v) = \alpha\,(u,x)\,(y,v)$$

and

$$(u,\overline{\alpha}x \otimes y\,(v)) = (u,\overline{\alpha}\,(v,y)\,x) = \alpha\,(y,v)\,(u,x)\,.$$

Therefore, this verifies 13.7.

To verify 13.8, let $u \in X$.

$$(z \otimes y_1)\,(y_2 \otimes x)\,(u) = (u,x)\,(z \otimes y_1)\,(y_2) = (u,x)\,(y_2,y_1)\,z$$

and

$$(y_2,y_1)\,z \otimes x\,(u) = (y_2,y_1)\,(u,x)\,z.$$

Since the two linear transformations on both sides of 13.8 give the same answer for every $u \in X$, it follows the two transformations are the same. This proves the lemma.

Definition 13.0.20 *Let X,Y be two vector spaces. Then define for $A,B \in \mathcal{L}(X,Y)$ and $\alpha \in \mathbb{F}$, new elements of $\mathcal{L}(X,Y)$ denoted by $A+B$ and αA as follows.*

$$(A+B)\,(x) \equiv Ax + Bx,\ (\alpha A)\,x \equiv \alpha\,(Ax)\,.$$

Theorem 13.0.21 *Let X and Y be finite dimensional inner product spaces. Then $\mathcal{L}(X,Y)$ is a vector space with the above definition of what it means to multiply by a scalar and add. Let $\{v_1,\cdots,v_n\}$ be an orthonormal basis for X and $\{w_1,\cdots,w_m\}$ be an orthonormal basis for Y. Then a basis for $\mathcal{L}(X,Y)$ is $\{v_i\otimes w_j : i=1,\cdots,n, j=1,\cdots,m\}$.*

Proof: It is obvious that $\mathcal{L}(X,Y)$ is a vector space. It remains to verify the given set is a basis. Consider the following:

$$\left(\left(A-\sum_{k,l}(Av_k,w_l)\,w_l\otimes v_k\right)v_p,w_r\right)=(Av_p,w_r)-$$

$$\sum_{k,l}(Av_k,w_l)(v_p,v_k)(w_l,w_r)$$

$$=(Av_p,w_r)-\sum_{k,l}(Av_k,w_l)\,\delta_{pk}\delta_{rl}$$

$$=(Av_p,w_r)-(Av_p,w_r)=0.$$

Letting $A-\sum_{k,l}(Av_k,w_l)\,w_l\otimes v_k=B$, this shows that $Bv_p=0$ since w_r is an arbitrary element of the basis for Y. Since v_p is an arbitrary element of the basis for X, it follows $B=0$ as hoped. This has shown $\{v_i\otimes w_j : i=1,\cdots,n, j=1,\cdots,m\}$ spans $\mathcal{L}(X,Y)$.

It only remains to verify the $v_i\otimes w_j$ are linearly independent. Suppose then that

$$\sum_{i,j}c_{ij}v_i\otimes w_j=0$$

Then,

$$0=\left(v_s,\sum_{i,j}c_{ij}v_i\otimes w_j(w_r)\right)=\left(v_s,\sum_{i,j}c_{ij}v_i(w_r,w_j)\right)$$

$$=\sum_{i,j}(v_s,c_{ij}v_i)(w_r,w_j)=\sum_{i,j}\overline{c_{ij}}\delta_{si}\delta_{rj}=\overline{c_{sr}}$$

showing all the coefficients equal zero. This proves independence.

Note this shows the dimension of $\mathcal{L}(X,Y)=nm$. The theorem is also of enormous importance because it shows you can always consider an arbitrary linear transformation as a sum of rank one transformations whose properties are easily understood. The following theorem is also of great interest.

Theorem 13.0.22 *Let $A=\sum_{i,j}c_{ij}w_i\otimes v_j\in\mathcal{L}(X,Y)$ where as before, the vectors, $\{w_i\}$ are an orthonormal basis for Y and the vectors, $\{v_j\}$ are an orthonormal basis for X. Then if the matrix of A has components, M_{ij}, it follows that $M_{ij}=c_{ij}$.*

Proof: Recall the diagram which describes what the matrix of a linear transformation is.

$$\begin{array}{ccccc}\{v_1,\cdots,v_n\} & X & \underrightarrow{A} & Y & \{w_1,\cdots,w_m\}\\ & q_V\uparrow & \circ & \uparrow q_W & \\ & \mathbb{F}^n & \underrightarrow{M} & \mathbb{F}^m & \end{array}$$

Thus, multiplication by the matrix, M followed by the map, q_W is the same as q_V followed by the linear transformation, A. Denoting by M_{ij} the components of the matrix, M, and letting $\mathbf{x}=(x_1,\cdots,x_n)\in\mathbb{F}^n$,

$$\sum_i w_i\sum_j M_{ij}x_j=A\left(\sum_k x_kv_k\right)$$

$$= \sum_{i,j} \sum_{k} c_{ij} x_k \delta_{kj} w_i = \sum_i w_i \sum_j c_{ij} x_j.$$

It follows from the linear independence of the w_i that for any $\mathbf{x} \in \mathbb{F}^n$,

$$\sum_j M_{ij} x_j = \sum_j c_{ij} x_j$$

which establishes the theorem.

13.1 Least squares

A common problem in experimental work is to find a straight line which approximates as well as possible a collection of points in the plane $\{(x_i, y_i)\}_{i=1}^p$. The usual way of dealing with these problems is by the method of least squares and it turns out that all these sorts of approximation problems can be reduced to $A\mathbf{x} = \mathbf{b}$ where the problem is to find the best $\mathbf{x}$ for solving this equation even when there is no solution.

Lemma 13.1.1 *Let V and W be finite dimensional inner product spaces and let $A : V \to W$ be linear. For each $y \in W$ there exists $x \in V$ such that*

$$|Ax - y| \leq |Ax_1 - y|$$

for all $x_1 \in V$. Also, $x \in V$ is a solution to this minimization problem if and only if x is a solution to the equation, $A^*Ax = A^*y$.

Proof: By Theorem 13.0.13 on Page 244 there exists a point, Ax_0, in the finite dimensional subspace, $A(V)$, of W such that for all $x \in V, |Ax - y|^2 \geq |Ax_0 - y|^2$. Also, from this theorem, this happens if and only if $Ax_0 - y$ is perpendicular to every $Ax \in A(V)$. Therefore, the solution is characterized by $(Ax_0 - y, Ax) = 0$ for all $x \in V$ which is the same as saying $(A^*Ax_0 - A^*y, x) = 0$ for all $x \in V$. In other words the solution is obtained by solving $A^*Ax_0 = A^*y$ for x_0.

Consider the problem of finding the least squares regression line in statistics. Suppose you have given points in the plane, $\{(x_i, y_i)\}_{i=1}^n$ and you would like to find constants m and b such that the line $y = mx + b$ goes through all these points. Of course this will be impossible in general. Therefore, try to find m, b such that you do the best you can to solve the system

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix}$$

which is of the form $\mathbf{y} = A\mathbf{x}$. In other words try to make $\left| A \begin{pmatrix} m \\ b \end{pmatrix} - \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right|^2$ as small as possible. According to what was just shown, it is desired to solve the following for m and b.

$$A^*A \begin{pmatrix} m \\ b \end{pmatrix} = A^* \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Since $A^* = A^T$ in this case,

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix}$$

Solving this system of equations for m and b,

$$m = \frac{-\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) + \left(\sum_{i=1}^n x_i y_i\right) n}{\left(\sum_{i=1}^n x_i^2\right) n - \left(\sum_{i=1}^n x_i\right)^2}$$

and

$$b = \frac{-\left(\sum_{i=1}^n x_i\right)\sum_{i=1}^n x_i y_i + \left(\sum_{i=1}^n y_i\right)\sum_{i=1}^n x_i^2}{\left(\sum_{i=1}^n x_i^2\right) n - \left(\sum_{i=1}^n x_i\right)^2}.$$

One could clearly do a least squares fit for curves of the form $y = ax^2 + bx + c$ in the same way. In this case you solve as well as possible for a, b, and c the system

$$\begin{pmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

using the same techniques.

Definition 13.1.2 *Let S be a subset of an inner product space, X. Define*

$$S^\perp \equiv \{x \in X : (x, s) = 0 \text{ for all } s \in S\}.$$

The following theorem also follows from the above lemma. It is sometimes called the Fredholm alternative.

Theorem 13.1.3 *Let $A : V \to W$ where A is linear and V and W are inner product spaces. Then $A(V) = \ker(A^*)^\perp$.*

Proof: Let $y = Ax$ so $y \in A(V)$. Then if $A^*z = 0$,

$$(y, z) = (Ax, z) = (x, A^*z) = 0$$

showing that $y \in \ker(A^*)^\perp$. Thus $A(V) \subseteq \ker(A^*)^\perp$.

Now suppose $y \in \ker(A^*)^\perp$. Does there exists x such that $Ax = y$? Since this might not be imediately clear, take the least squares solution to the problem. Thus let x be a solution to $A^*Ax = A^*y$. It follows $A^*(y - Ax) = 0$ and so $y - Ax \in \ker(A^*)$ which implies from the assumption about y that $(y - Ax, y) = 0$. Also, since Ax is the closest point to y in $A(V)$, Theorem 13.0.13 on Page 244 implies that $(y - Ax, Ax_1) = 0$ for all $x_1 \in V$. In particular this is true for $x_1 = x$ and so $0 = (y - Ax, y) - (y - Ax, Ax) = |y - Ax|^2$, showing that $y = Ax$. Thus $A(V) \supseteq \ker(A^*)^\perp$ and this proves the Theorem.

Corollary 13.1.4 *Let A, V, and W be as described above. If the only solution to $A^*y = 0$ is $y = 0$, then A is onto W.*

Proof: If the only solution to $A^*y = 0$ is $y = 0$, then $\ker(A^*) = \{0\}$ and so every vector from W is contained in $\ker(A^*)^\perp$ and by the above theorem, this shows $A(V) = W$.

13.2 Exercises

1. Find the best solution to the system

$$\begin{aligned} x + 2y &= 6 \\ 2x - y &= 5 \\ 3x + 2y &= 0 \end{aligned}$$

2. Suppose you are given the data, $(1,2),(2,4),(3,8),(0,0)$. Find the linear regression line using your formulas derived above. Then graph your data along with your regression line.

3. Generalize the least squares procedure to the situation in which data is given and you desire to fit it with an expression of the form $y = af(x) + bg(x) + c$ where the problem would be to find a, b and c in order to minimize the error. Could this be generalized to higher dimensions? How about more functions?

4. Let $A \in \mathcal{L}(X,Y)$ where X and Y are finite dimensional vector spaces with the dimension of X equal to n. Define $\text{rank}(A) \equiv \dim(A(X))$ and $\text{nullity}(A) \equiv \dim(\ker(A))$. Show that $\text{nullity}(A) + \text{rank}(A) = \dim(X)$. **Hint:** Let $\{x_i\}_{i=1}^r$ be a basis for $\ker(A)$ and let $\{x_i\}_{i=1}^r \cup \{y_i\}_{i=1}^{n-r}$ be a basis for X. Then show that $\{Ay_i\}_{i=1}^{n-r}$ is linearly independent and spans AX.

5. Let A be an $m \times n$ matrix. Show the column rank of A equals the column rank of A^*A. Next verify column rank of A^*A is no larger than column rank of A^*. Next justify the following inequality to conclude the column rank of A equals the column rank of A^*.

$$\text{rank }(A) = \text{rank }(A^*A) \leq \text{rank }(A^*) \leq$$

$$= \text{rank }(AA^*) \leq \text{rank }(A).$$

Hint: Start with an orthonormal basis, $\{A\mathbf{x}_j\}_{j=1}^r$ of $A(\mathbb{F}^n)$ and verify $\{A^*A\mathbf{x}_j\}_{j=1}^r$ is a basis for $A^*A(\mathbb{F}^n)$.

13.3 The Determinant And Volume

The determinant is the essential algebraic tool which provides a way to give a unified treatment of the concept of volume. With the above preparation the concept of volume is considered in this section. The following lemma is not hard to obtain from earlier topics.

Lemma 13.3.1 *Suppose A is an $m \times n$ matrix where $m > n$. Then A does not map $\mathbb{R}^n$ onto $\mathbb{R}^m$.*

Proof: First note that $A(\mathbb{R}^n)$ has dimension no more than n because a spanning set is $\{A\mathbf{e}_1, \cdots, A\mathbf{e}_n\}$ and so it can't possibly include all of $\mathbb{R}^m$ if $m > n$ because the dimension of $\mathbb{R}^m = m$. This proves the lemma. Here is another proof which uses determinants.

Suppose A did map $\mathbb{R}^n$ onto $\mathbb{R}^m$. Then consider the $m \times m$ matrix,

$$A_1 \equiv \begin{pmatrix} A & 0 \end{pmatrix}$$

where 0 refers to an $n \times (m-n)$ matrix. Thus A_1 cannot be onto $\mathbb{R}^m$ because it has at least one column of zeros and so its determinant equals zero. However, if $\mathbf{y} \in \mathbb{R}^m$ and A is onto, then there exists $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = \mathbf{y}$. Then

$$A_1 \begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix} = A\mathbf{x} + \mathbf{0} = A\mathbf{x} = \mathbf{y}.$$

Since $\mathbf{y}$ was arbitrary, it follows A_1 would have to be onto.

The following proposition is a special case of the exchange theorem but I will give a different proof based on the above lemma.

Proposition 13.3.2 *Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ are vectors in $\mathbb{R}^n$ and $\text{span}\{\mathbf{v}_1, \cdots, \mathbf{v}_p\} = \mathbb{R}^n$. Then $p \geq n$.*

Proof: Define a linear transformation from $\mathbb{R}^p$ to $\mathbb{R}^n$ as follows.

$$A\mathbf{x} \equiv \sum_{i=1}^{p} x_i \mathbf{v}_i.$$

(Why is this a linear transformation?) Thus $A(\mathbb{R}^p) = \text{span}\{\mathbf{v}_1, \cdots, \mathbf{v}_p\} = \mathbb{R}^n$. Then from the above lemma, $p \geq n$ since if this is not so, A could not be onto.

Proposition 13.3.3 *Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ are vectors in $\mathbb{R}^n$ such that $p < n$. Then there exist at least $n-p$ vectors, $\{\mathbf{w}_{p+1}, \cdots, \mathbf{w}_n\}$ such that $\mathbf{w}_i \cdot \mathbf{w}_j = \delta_{ij}$ and $\mathbf{w}_k \cdot \mathbf{v}_j = 0$ for every $j = 1, \cdots, \mathbf{v}_p$.*

Proof: Let $A : \mathbb{R}^p \to \text{span}\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ be defined as in the above proposition so that $A(\mathbb{R}^p) = \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_p)$. Since $p < n$ there exists $\mathbf{z}_{p+1} \notin A(\mathbb{R}^p)$. Then by Theorem 13.0.13 on Page 244 applied to the subspace $A(\mathbb{R}^n)$ there exists $\mathbf{x}_{p+1}$ such that

$$(\mathbf{z}_{p+1} - \mathbf{x}_{p+1}, A\mathbf{y}) = 0$$

for all $\mathbf{y} \in \mathbb{R}^p$. Let $\mathbf{w}_{p+1} \equiv (\mathbf{z}_{p+1} - \mathbf{x}_{p+1}) / |\mathbf{z}_{p+1} - \mathbf{x}_{p+1}|$. Now if $p+1 = n$, stop. $\{\mathbf{w}_{p+1}\}$ is the desired list of vectors. Otherwise, do for $\text{span}\{\mathbf{v}_1, \cdots, \mathbf{v}_p, \mathbf{w}_{p+1}\}$ what was done for $\text{span}\{\mathbf{v}_1, \cdots, \mathbf{v}_p\}$ using $\mathbb{R}^{p+1}$ instead of $\mathbb{R}^p$ and obtain $\mathbf{w}_{p+2}$ in this way such that $\mathbf{w}_{p+2} \cdot \mathbf{w}_{p+1} = 0$ and $\mathbf{w}_{p+2} \cdot \mathbf{v}_k = 0$ for all k. Continue till a list of $n-p$ vectors have been found.

Recall the geometric definition of the cross product of two vectors found on Page 41. As explained there, the magnitude of the cross product of two vectors was the area of the parallelogram determined by the two vectors. There was also a coordinate description of the cross product. In terms of the notation of Proposition 4.5.4 on Page 50 the i^{th} coordinate of the cross product is given by

$$\varepsilon_{ijk} u_j v_k$$

where the two vectors are (u_1, u_2, u_3) and (v_1, v_2, v_3). Therefore, using the reduction identity of Lemma 4.5.3 on Page 50

$$\begin{aligned} |\mathbf{u} \times \mathbf{v}|^2 &= \varepsilon_{ijk} u_j v_k \varepsilon_{irs} u_r v_s \\ &= (\delta_{jr}\delta_{ks} - \delta_{kr}\delta_{js}) u_j v_k u_r v_s \\ &= u_j v_k u_j v_k - u_j v_k u_k v_j \\ &= (\mathbf{u} \cdot \mathbf{u})(\mathbf{v} \cdot \mathbf{v}) - (\mathbf{u} \cdot \mathbf{v})^2 \end{aligned}$$

which equals

$$\det \begin{pmatrix} \mathbf{u} \cdot \mathbf{u} & \mathbf{u} \cdot \mathbf{v} \\ \mathbf{u} \cdot \mathbf{v} & \mathbf{v} \cdot \mathbf{v} \end{pmatrix}.$$

Now recall the box product and how the box product was $\pm$ the volume of the parallelepiped spanned by the three vectors. From the definition of the box product

$$\begin{aligned} \mathbf{u} \times \mathbf{v} \cdot \mathbf{w} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \cdot (w_1\mathbf{i} + w_2\mathbf{j} + w_3\mathbf{k}) \\ &= \det \begin{pmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{pmatrix}. \end{aligned}$$

Therefore,

$$|\mathbf{u}\times\mathbf{v}\cdot\mathbf{w}|^2 = \det\begin{pmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{pmatrix}^2$$

which from the theory of determinants equals

$$\det\begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix}\det\begin{pmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{pmatrix} =$$

$$\det\left(\begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix}\begin{pmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{pmatrix}\right) =$$

$$\det\begin{pmatrix} u_1^2+u_2^2+u_3^2 & u_1v_1+u_2v_2+u_3v_3 & u_1w_1+u_2w_2+u_3w_3 \\ u_1v_1+u_2v_2+u_3v_3 & v_1^2+v_2^2+v_3^2 & v_1w_1+v_2w_2+v_3w_3 \\ u_1w_1+u_2w_2+u_3w_3 & v_1w_1+v_2w_2+v_3w_3 & w_1^2+w_2^2+w_3^2 \end{pmatrix}$$

$$= \det\begin{pmatrix} \mathbf{u}\cdot\mathbf{u} & \mathbf{u}\cdot\mathbf{v} & \mathbf{u}\cdot\mathbf{w} \\ \mathbf{u}\cdot\mathbf{v} & \mathbf{v}\cdot\mathbf{v} & \mathbf{v}\cdot\mathbf{w} \\ \mathbf{u}\cdot\mathbf{w} & \mathbf{v}\cdot\mathbf{w} & \mathbf{w}\cdot\mathbf{w} \end{pmatrix}$$

You see there is a definite pattern emerging here. These earlier cases were for a parallelepiped determined by either two or three vectors in $\mathbb{R}^3$. It makes sense to speak of a parallelepiped in any number of dimensions.

Definition 13.3.4 *Let $\mathbf{u}_1,\cdots,\mathbf{u}_p$ be vectors in $\mathbb{R}^k$. The parallelepiped determined by these vectors will be denoted by $P(\mathbf{u}_1,\cdots,\mathbf{u}_p)$ and it is defined as*

$$P(\mathbf{u}_1,\cdots,\mathbf{u}_p) \equiv \left\{\sum_{j=1}^{p} s_j\mathbf{u}_j : s_j\in[0,1]\right\}.$$

The volume of this parallelepiped is defined as

$$\text{volume of } P(\mathbf{u}_1,\cdots,\mathbf{u}_p) \equiv \left(\det(\mathbf{u}_i\cdot\mathbf{u}_j)\right)^{1/2}.$$

In this definition, $\mathbf{u}_i\cdot\mathbf{u}_j$ is the ij^{th} entry of a $p\times p$ matrix. Note this definition agrees with all earlier notions of area and volume for parallelepipeds and it makes sense in any number of dimensions. However, it is important to verify the above determinant is nonnegative. After all, the above definition requires a square root of this determinant.

Lemma 13.3.5 *Let $\mathbf{u}_1,\cdots,\mathbf{u}_p$ be vectors in $\mathbb{R}^k$ for some k. Then $\det(\mathbf{u}_i\cdot\mathbf{u}_j)\geq 0$.*

Proof: Recall $\mathbf{v}\cdot\mathbf{w} = \mathbf{v}^T\mathbf{w}$. Therefore, in terms of matrix multiplication, the matrix $(\mathbf{u}_i\cdot\mathbf{u}_j)$ is just the following

$$\overbrace{\begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_p^T \end{pmatrix}}^{p\times k}\overbrace{\begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_p \end{pmatrix}}^{k\times p}$$

which is of the form

$$U^TU.$$

First it is necessary to show $\det\left(U^TU\right) \geq 0$. If U were a square matrix, this would be imediate but it isn't. However, by Proposition 13.3.3 there are vectors, $\mathbf{w}_{p+1}, \cdots, \mathbf{w}_k$ such that $\mathbf{w}_i \cdot \mathbf{w}_j = \delta_{ij}$ and for all $i = 1, \cdots, p$, and $l = p+1, \cdots, n$, $\mathbf{w}_l \cdot \mathbf{u}_i = 0$. Then consider

$$U_1 = (\mathbf{u}_1, \cdots, \mathbf{u}_p, \mathbf{w}_{p+1}, \cdots, \mathbf{w}_k) \equiv \begin{pmatrix} U & W \end{pmatrix}$$

where $W^TW = I$. Then

$$U_1^TU_1 = \begin{pmatrix} U^T \\ W^T \end{pmatrix} \begin{pmatrix} U & W \end{pmatrix} = \begin{pmatrix} U^TU & 0 \\ 0 & I \end{pmatrix}. \text{ (Why?)}$$

Now using the cofactor expansion method, this last $k \times k$ matrix has determinant equal to $\det\left(U^TU\right)$ (Why?) On the other hand this equals $\det\left(U_1^TU_1\right) = \det(U_1)\det\left(U_1^T\right) = \det(U_1)^2 \geq 0$.

In the case where $k < p$, U^TU has the form WW^T where $W = U^T$ has more rows than columns. Thus you can define the $p \times p$ matrix,

$$W_1 \equiv \begin{pmatrix} W & 0 \end{pmatrix},$$

and in this case,

$$0 = \det W_1W_1^T = \det \begin{pmatrix} W & 0 \end{pmatrix} \begin{pmatrix} W^T \\ 0 \end{pmatrix} = \det WW^T = \det U^TU.$$

This proves the lemma and shows the definition of volume is well defined.

Note it gives the right answer in the case where all the vectors are perpendicular. Here is why. Suppose $\{\mathbf{u}_1, \cdots, \mathbf{u}_p\}$ are vectors which have the property that $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ if $i \neq j$. Thus $P(\mathbf{u}_1, \cdots, \mathbf{u}_p)$ is a box which has all p sides perpendicular. What should its volume be? Shouldn't it equal the product of the lengths of the sides? What does $\det(\mathbf{u}_i \cdot \mathbf{u}_j)$ give? The matrix $(\mathbf{u}_i \cdot \mathbf{u}_j)$ is a diagonal matrix having the squares of the magnitudes of the sides down the diagonal. Therefore, $\det(\mathbf{u}_i \cdot \mathbf{u}_j)^{1/2}$ equals the product of the lengths of the sides as it should. The matrix, $(\mathbf{u}_i \cdot \mathbf{u}_j)$ whose determinant gives the square of the volume of the parallelepiped spanned by the vectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_p\}$ is called the Gramian matrix and sometimes the metric tensor.

These considerations are of great significance because they allow the computation in a systematic manner of k dimensional volumes of parallelepipeds which happen to be in $\mathbb{R}^n$ for $n \neq k$. Think for example of a plane in $\mathbb{R}^3$ and the problem of finding the area of something on this plane.

Example 13.3.6 *Find the equation of the plane containing the three points,* $(1,2,3), (0,2,1),$ *and* $(3,1,0)$.

These three points determine two vectors, the one from $(0,2,1)$ to $(1,2,3)$, $\mathbf{i} + 0\mathbf{j} + 2\mathbf{k}$, and the one from $(0,2,1)$ to $(3,1,0)$, $3\mathbf{i} + (-1)\mathbf{j} + (-1)\mathbf{k}$. If (x,y,z) denotes a point in the plane, then the volume of the parallelepiped spanned by the vector from $(0,2,1)$ to (x,y,z) and these other two vectors must be zero. Thus

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

Therefore, $-2x - 7y + 13 + z = 0$ is the equation of the plane. You should check it contains all three points.

13.4 Exercises

1. Here are three vectors in $\mathbb{R}^4 : (1,2,0,3)^T, (2,1,-3,2)^T, (0,0,1,2)^T$. Find the volume of the parallelepiped determined by these three vectors.

2. Here are two vectors in $\mathbb{R}^4 : (1,2,0,3)^T, (2,1,-3,2)^T$. Find the volume of the parallelepiped determined by these two vectors.

3. Here are three vectors in $\mathbb{R}^2 : (1,2)^T, (2,1)^T, (0,1)^T$. Find the volume of the parallelepiped determined by these three vectors. Recall that from the above theorem, this should equal 0.

4. If there are $n+1$ or more vectors in $\mathbb{R}^n$, Lemma 13.3.5 implies the parallelepiped determined by these $n+1$ vectors must have zero volume. What is the geometric significance of this assertion?

5. Find the equation of the plane through the three points $(1,2,3), (2,-3,1), (1,1,7)$.

Self Adjoint Operators

14.1 Simultaneous Diagonalization

It is sometimes interesting to consider the problem of finding a single similarity transformation which will diagonalize all the matrices in some set.

Lemma 14.1.1 *Let A be an $n \times n$ matrix and let B be an $m \times m$ matrix. Denote by C the matrix,*

$$C \equiv \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}.$$

Then C is diagonalizable if and only if both A and B are diagonalizable.

Proof: Suppose $S_A^{-1} A S_A = D_A$ and $S_B^{-1} B S_B = D_B$ where D_A and D_B are diagonal matrices. You should use block multiplication to verify that $S \equiv \begin{pmatrix} S_A & 0 \\ 0 & S_B \end{pmatrix}$ is such that $S^{-1} C S = D_C$, a diagonal matrix.

Conversely, suppose C is diagonalized by $S = (\mathbf{s}_1, \cdots, \mathbf{s}_{n+m})$. Thus S has columns $\mathbf{s}_i$. For each of these columns, write in the form

$$\mathbf{s}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix}$$

where $\mathbf{x}_i \in \mathbb{F}^n$ and where $\mathbf{y}_i \in \mathbb{F}^m$. It follows each of the $\mathbf{x}_i$ is an eigenvector of A and that each of the $\mathbf{y}_i$ is an eigenvector of B. If there are n linearly independent $\mathbf{x}_i$, then A is diagonalizable by Theorem 11.5.9 on Page 11.5.9. The row rank of the matrix, $(\mathbf{x}_1, \cdots, \mathbf{x}_{n+m})$ must be n because if this is not so, the rank of S would be less than $n+m$ which would mean S^{-1} does not exist. Therefore, since the column rank equals the row rank, this matrix has column rank equal to n and this means there are n linearly independent eigenvectors of A implying that A is diagonalizable. Similar reasoning applies to B. This proves the lemma.

The following corollary follows from the same type of argument as the above.

Corollary 14.1.2 *Let A_k be an $n_k \times n_k$ matrix and let C denote the block diagonal $(\sum_{k=1}^r n_k) \times (\sum_{k=1}^r n_k)$ matrix given below.*

$$C \equiv \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_r \end{pmatrix}.$$

Then C is diagonalizable if and only if each A_k is diagonalizable.

Definition 14.1.3 *A set, $\mathcal{F}$ of $n \times n$ matrices is simultaneously diagonalizable if and only if there exists a single invertible matrix, S such that $S^{-1}AS = D$, a diagonal matrix for all $A \in \mathcal{F}$.*

Lemma 14.1.4 *If $\mathcal{F}$ is a set of $n \times n$ matrices which is simultaneously diagonalizable, then $\mathcal{F}$ is a commuting family of matrices.*

Proof: Let $A, B \in \mathcal{F}$ and let S be a matrix which has the property that $S^{-1}AS$ is a diagonal matrix for all $A \in \mathcal{F}$. Then $S^{-1}AS = D_A$ and $S^{-1}BS = D_B$ where D_A and D_B are diagonal matrices. Since diagonal matrices commute,

$$\begin{aligned} AB &= SD_AS^{-1}SD_BS^{-1} = SD_AD_BS^{-1} \\ &= SD_BD_AS^{-1} = SD_BS^{-1}SD_AS^{-1} = BA. \end{aligned}$$

Lemma 14.1.5 *Let D be a diagonal matrix of the form*

$$D \equiv \begin{pmatrix} \lambda_1 I_{n_1} & 0 & \cdots & 0 \\ 0 & \lambda_2 I_{n_1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_r I_{n_r} \end{pmatrix}, \tag{14.1}$$

where I_{n_i} denotes the $n_i \times n_i$ identity matrix and suppose B is a matrix which commutes with D. Then B is a block diagonal matrix of the form

$$B = \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & B_r \end{pmatrix} \tag{14.2}$$

where B_i is an $n_i \times n_i$ matrix.

Proof: Suppose $B = (b_{ij})$. Then since it is given to commute with D, $\lambda_i b_{ij} = b_{ij}\lambda_j$. But this shows that if $\lambda_i \neq \lambda_j$, then this could not occur unless $b_{ij} = 0$. Therefore, B must be of the claimed form.

Lemma 14.1.6 *Let $\mathcal{F}$ denote a commuting family of $n \times n$ matrices such that each $A \in \mathcal{F}$ is diagonalizable. Then $\mathcal{F}$ is simultaneously diagonalizable.*

Proof: This is proved by induction on n. If $n = 1$, there is nothing to prove because all the 1×1 matrices are already diagonal matrices. Suppose then that the theorem is true for all $k \leq n-1$ where $n \geq 2$ and let $\mathcal{F}$ be a commuting family of diagonalizable $n \times n$ matrices. Pick $A \in \mathcal{F}$ and let S be an invertible matrix such that $S^{-1}AS = D$ where D is of the form given in 14.1. Now denote by $\widetilde{\mathcal{F}}$ the collection of matrices, $\left\{S^{-1}BS : B \in \mathcal{F}\right\}$. It follows easily that $\widetilde{\mathcal{F}}$ is also a commuting family of diagonalizable matrices. By Lemma 14.1.5 every $B \in \widetilde{\mathcal{F}}$ is of the form given in 14.2 and by block multiplication, the B_i corresponding to different $B \in \widetilde{\mathcal{F}}$ commute. Therefore, by the induction hypothesis, the knowledge that each $B \in \widetilde{\mathcal{F}}$ is diagonalizable, and Corollary 14.1.2, there exist invertible $n_i \times n_i$ matrices, T_i such that $T_i^{-1}B_iT_i$ is a diagonal matrix whenever B_i is one of the matrices making up the block diagonal of any $B \in \mathcal{F}$. It follows that for T defined by

$$T \equiv \begin{pmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & T_r \end{pmatrix},$$

then $T^{-1}BT =$ a diagonal matrix for every $B \in \widetilde{\mathcal{F}}$ including D. Consider ST. It follows that for all $B \in \mathcal{F}$,

$$T^{-1}S^{-1}BST = (ST)^{-1} B (ST) = \text{ a diagonal matrix.}$$

This proves the lemma.

Theorem 14.1.7 *Let $\mathcal{F}$ denote a family of matrices which are diagonalizable. Then $\mathcal{F}$ is simultaneously diagonalizable if and only if $\mathcal{F}$ is a commuting family.*

Proof: If $\mathcal{F}$ is a commuting family, it follows from Lemma 14.1.6 that it is simultaneously diagonalizable. If it is simultaneously diagonalizable, then it follows from Lemma 14.1.4 that it is a commuting family. This proves the theorem.

14.2 Spectral Theory Of Self Adjoint Operators

The following theorem is about the eigenvectors and eigenvalues of a self adjoint operator. The proof given generalizes to the situation of a compact self adjoint operator on a Hilbert space and leads to many very useful results. It is also a very elementary proof because it does not use the fundamental theorem of algebra and it contains a way, very important in applications, of finding the eigenvalues. This proof depends more directly on the methods of analysis than the preceding material. The following is useful notation.

Definition 14.2.1 *Let X be an inner product space and let $S \subseteq X$. Then*

$$S^{\perp} \equiv \{x \in X : (x, s) = 0 \text{ for all } s \in S\}.$$

Note that even if S is not a subspace, $S^{\perp}$ is.

Definition 14.2.2 *A Hilbert space is a complete inner product space. Recall this means that every Cauchy sequence,$\{x_n\}$, one which satisfies*

$$\lim_{n,m\to\infty} |x_n - x_m| = 0,$$

converges. It can be shown, although I will not do so here, that for the field of scalars either $\mathbb{R}$ or $\mathbb{C}$, any finite dimensional inner product space is automatically complete.

Theorem 14.2.3 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Thus $A = A^*$. Then there exists an orthonormal basis of eigenvectors, $\{u_j\}_{j=1}^n$.*

Proof: Consider (Ax, x). This quantity is always a real number because

$$\overline{(Ax, x)} = (x, Ax) = (x, A^*x) = (Ax, x)$$

thanks to the assumption that A is self adjoint. Now define

$$\lambda_1 \equiv \inf \{(Ax, x) : |x| = 1, x \in X_1 \equiv X\}.$$

Claim: λ_1 is finite and there exists $v_1 \in X$ with $|v_1| = 1$ such that $(Av_1, v_1) = \lambda_1$.

Proof of claim: Let $\{u_j\}_{j=1}^n$ be an orthonormal basis for X and for $x \in X$, let $(x_1, \cdots, x_n)$ be defined as the components of the vector x. Thus,

$$x = \sum_{j=1}^{n} x_j u_j.$$

Since this is an orthonormal basis, it follows from the axioms of the inner product that

$$|x|^2 = \sum_{j=1}^{n} |x_j|^2 .$$

Thus

$$(Ax, x) = \left(\sum_{k=1}^{n} x_k A u_k, \sum_{j=1} x_j u_j \right) = \sum_{k,j} x_k \overline{x_j} \, (Au_k, u_j) ,$$

a continuous function of $(x_1, \cdots, x_n)$. Thus this function achieves its minimum on the closed and bounded subset of $\mathbb{F}^n$ given by

$$\{(x_1, \cdots, x_n) \in \mathbb{F}^n : \sum_{j=1}^{n} |x_j|^2 = 1\}.$$

Then $v_1 \equiv \sum_{j=1}^{n} x_j u_j$ where $(x_1, \cdots, x_n)$ is the point of $\mathbb{F}^n$ at which the above function achieves its minimum. This proves the claim.

Continuing with the proof of the theorem, let $X_2 \equiv \{v_1\}^{\perp}$ and let

$$\lambda_2 \equiv \inf \{(Ax, x) : |x| = 1, x \in X_2\}$$

As before, there exists $v_2 \in X_2$ such that $(Av_2, v_2) = \lambda_2$. Now let $X_3 \equiv \{v_1, v_2\}^{\perp}$ and continue in this way. This leads to an increasing sequence of real numbers, $\{\lambda_k\}_{k=1}^{n}$ and an orthonormal set of vectors, $\{v_1, \cdots, v_n\}$. It only remains to show these are eigenvectors and that the λ_j are eigenvalues.

Consider the first of these vectors. Letting $w \in X_1 \equiv X$, the function of the real variable, t, given by

$$f(t) \equiv \frac{(A(v_1 + tw), v_1 + tw)}{|v_1 + tw|^2}$$

$$= \frac{(Av_1, v_1) + 2t \operatorname{Re}(Av_1, w) + t^2 (Aw, w)}{|v_1|^2 + 2t \operatorname{Re}(v_1, w) + t^2 |w|^2}$$

achieves its minimum when $t = 0$. Therefore, the derivative of this function evaluated at $t = 0$ must equal zero. Using the quotient rule, this implies

$$2 \operatorname{Re}(Av_1, w) - 2 \operatorname{Re}(v_1, w) (Av_1, v_1)$$

$$= 2 (\operatorname{Re}(Av_1, w) - \operatorname{Re}(v_1, w) \lambda_1) = 0.$$

Thus $\operatorname{Re}(Av_1 - \lambda_1 v_1, w) = 0$ for all $w \in X$. This implies $Av_1 = \lambda_1 v_1$. To see this, let $w \in X$ be arbitrary and let θ be a complex number with $|\theta| = 1$ and

$$|(Av_1 - \lambda_1 v_1, w)| = \theta (Av_1 - \lambda_1 v_1, w) .$$

Then

$$|(Av_1 - \lambda_1 v_1, w)| = \operatorname{Re}\left(Av_1 - \lambda_1 v_1, \overline{\theta} w\right) = 0.$$

Since this holds for all w, $Av_1 = \lambda_1 v_1$. Now suppose $Av_k = \lambda_k v_k$ for all $k < m$. Observe that $A : X_m \to X_m$ because if $y \in X_m$ and $k < m$,

$$(Ay, v_k) = (y, Av_k) = (y, \lambda_k v_k) = 0,$$

showing that $Ay \in \{v_1, \cdots, v_{m-1}\}^\perp \equiv X_m$. Thus the same argument just given shows that for all $w \in X_m$,

$$(Av_m - \lambda_m v_m, w) = 0. \tag{14.3}$$

For arbitrary $w \in X$.

$$w = \left(w - \sum_{k=1}^{m-1} (w, v_k)\, v_k \right) + \sum_{k=1}^{m-1} (w, v_k)\, v_k \equiv w_\perp + w_m$$

and the term in parenthesis is in $\{v_1, \cdots, v_{m-1}\}^\perp \equiv X_m$ while the other term is contained in the span of the vectors, $\{v_1, \cdots, v_{m-1}\}$. Thus by 14.3,

$$\begin{aligned} (Av_m - \lambda_m v_m, w) &= (Av_m - \lambda_m v_m, w_\perp + w_m) \\ &= (Av_m - \lambda_m v_m, w_m) = 0 \end{aligned}$$

because

$$A : X_m \to X_m \equiv \{v_1, \cdots, v_{m-1}\}^\perp$$

and $w_m \in \operatorname{span}(v_1, \cdots, v_{m-1})$. Therefore, $Av_m = \lambda_m v_m$ for all m. This proves the theorem.

Contained in the proof of this theorem is the following important corollary.

Corollary 14.2.4 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Then all the eigenvalues are real and for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \inf \{(Ax, x) : |x| = 1, x \in X_k\}$$

where

$$X_k \equiv \{u_1, \cdots, u_{k-1}\}^\perp, X_1 \equiv X.$$

Corollary 14.2.5 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Then the largest eigenvalue of A is given by*

$$\max \{(A\mathbf{x}, \mathbf{x}) : |\mathbf{x}| = 1\} \tag{14.4}$$

and the minimum eigenvalue of A is given by

$$\min \{(A\mathbf{x}, \mathbf{x}) : |\mathbf{x}| = 1\}. \tag{14.5}$$

Proof: The proof of this is just like the proof of Theorem 14.2.3. Simply replace inf with sup and obtain a decreasing list of eigenvalues. This establishes 14.4. The claim 14.5 follows from Theorem 14.2.3.

Another important observation is found in the following corollary.

Corollary 14.2.6 *Let $A \in \mathcal{L}(X, X)$ where A is self adjoint. Then $A = \sum_i \lambda_i v_i \otimes v_i$ where $Av_i = \lambda_i v_i$ and $\{v_i\}_{i=1}^n$ is an orthonormal basis.*

Proof : If v_k is one of the orthonormal basis vectors, $Av_k = \lambda_k v_k$. Also,

$$\begin{aligned} \sum_i \lambda_i v_i \otimes v_i (v_k) &= \sum_i \lambda_i v_i (v_k, v_i) \\ &= \sum_i \lambda_i \delta_{ik} v_i = \lambda_k v_k. \end{aligned}$$

Since the two linear transformations agree on a basis, it follows they must coincide. This proves the corollary.

The result of Courant and Fischer which follows resembles Corollary 14.2.4 but is more useful because it does not depend on a knowledge of the eigenvectors.

Theorem 14.2.7 *Let $A \in \mathcal{L}(X,X)$ be self adjoint where X is a finite dimensional Hilbert space. Then for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \max_{w_1,\cdots,w_{k-1}} \left\{ \min \left\{ (Ax,x) : |x| = 1, x \in \{w_1, \cdots, w_{k-1}\}^{\perp} \right\} \right\} \tag{14.6}$$

where if $k = 1$, $\{w_1, \cdots, w_{k-1}\}^{\perp} \equiv X$.

Proof: From Theorem 14.2.3, there exist eigenvalues and eigenvectors with $\{u_1, \cdots, u_n\}$ orthonormal and $\lambda_i \leq \lambda_{i+1}$. Therefore, by Corollary 14.2.6

$$A = \sum_{j=1}^{n} \lambda_j u_j \otimes u_j$$

Fix $\{w_1, \cdots, w_{k-1}\}$.

$$\begin{aligned} (Ax,x) &= \sum_{j=1}^{n} \lambda_j (x,u_j)(u_j,x) \\ &= \sum_{j=1}^{n} \lambda_j |(x,u_j)|^2 \end{aligned}$$

Then let $Y = \{w_1, \cdots, w_{k-1}\}^{\perp}$

$$\inf \{(Ax,x) : |x| = 1, x \in Y\}$$

$$= \inf \left\{ \sum_{j=1}^{n} \lambda_j |(x,u_j)|^2 : |x| = 1, x \in Y \right\}$$

$$\leq \inf \left\{ \sum_{j=1}^{k} \lambda_j |(x,u_j)|^2 : |x| = 1, (x,u_j) = 0 \text{ for } j > k, \text{ and } x \in Y \right\}. \tag{14.7}$$

The reason this is so is that the infimum is taken over a smaller set. Therefore, the infimum gets larger. Now 14.7 is no larger than

$$\inf \left\{ \lambda_k \sum_{j=1}^{k} |(x,u_j)|^2 : |x| = 1, (x,u_j) = 0 \text{ for } j > k, \text{ and } x \in Y \right\} = \lambda_k$$

because since $\{u_1, \cdots, u_n\}$ is an orthonormal basis, $|x|^2 = \sum_{j=1}^{n} |(x,u_j)|^2$. It follows since $\{w_1, \cdots, w_{k-1}\}$ is arbitrary,

$$\sup_{w_1,\cdots,w_{k-1}} \left\{ \inf \left\{ (Ax,x) : |x| = 1, x \in \{w_1, \cdots, w_{k-1}\}^{\perp} \right\} \right\} \leq \lambda_k. \tag{14.8}$$

However, for each $w_1, \cdots, w_{k-1}$, the infimum is achieved so you can replace the inf in the above with min. In addition to this, it follows from Corollary 14.2.4 that there exists a set, $\{w_1, \cdots, w_{k-1}\}$ for which

$$\inf \left\{ (Ax, x) : |x| = 1, x \in \{w_1, \cdots, w_{k-1}\}^{\perp} \right\} = \lambda_k.$$

Pick $\{w_1, \cdots, w_{k-1}\} = \{u_1, \cdots, u_{k-1}\}$. Therefore, the sup in 14.8 is achieved and equals λ_k and 14.6 follows. This proves the theorem.

The following corollary is immediate.

Corollary 14.2.8 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Then for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \max_{w_1, \cdots, w_{k-1}} \left\{ \min \left\{ \frac{(Ax, x)}{|x|^2} : x \neq 0, x \in \{w_1, \cdots, w_{k-1}\}^{\perp} \right\} \right\} \tag{14.9}$$

where if $k = 1, \{w_1, \cdots, w_{k-1}\}^{\perp} \equiv X$.

Here is a version of this for which the roles of max and min are reversed.

Corollary 14.2.9 *Let $A \in \mathcal{L}(X, X)$ be self adjoint where X is a finite dimensional Hilbert space. Then for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ the eigenvalues of A, there exist orthonormal vectors $\{u_1, \cdots, u_n\}$ for which*

$$Au_k = \lambda_k u_k.$$

Furthermore,

$$\lambda_k \equiv \min_{w_1, \cdots, w_{n-k}} \left\{ \max \left\{ \frac{(Ax, x)}{|x|^2} : x \neq 0, x \in \{w_1, \cdots, w_{n-k}\}^{\perp} \right\} \right\} \tag{14.10}$$

where if $k = n, \{w_1, \cdots, w_{n-k}\}^{\perp} \equiv X$.

14.3 Positive And Negative Linear Transformations

The notion of a positive definite or negative definite linear transformation is very important in many applications. In particular it is used in versions of the second derivative test for functions of many variables. Here the main interest is the case of a linear transformation which is an $n \times n$ matrix but the theorem is stated and proved using a more general notation because all these issues discussed here have interesting generalizations to functional analysis.

Lemma 14.3.1 *Let X be a finite dimensional Hilbert space and let $A \in \mathcal{L}(X, X)$. Then if $\{v_1, \cdots, v_n\}$ is an orthonormal basis for X and $M(A)$ denotes the matrix of the linear transformation, A then $M(A^*) = M(A)^*$. In particular, A is self adjoint, if and only if $M(A)$ is.*

Proof: Consider the following picture

$$\begin{array}{ccc} & A & \\ X & \rightarrow & X \\ q\uparrow & \circ & \uparrow q \\ \mathbb{F}^n & \rightarrow & \mathbb{F}^n \\ & M(A) & \end{array}$$

where q is the coordinate map which satisfies $q(\mathbf{x}) \equiv \sum_i x_i v_i$. Therefore, since $\{v_1, \cdots, v_n\}$ is orthonormal, it is clear that $|\mathbf{x}| = |q(\mathbf{x})|$. Therefore,

$$\begin{aligned} |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\operatorname{Re}(\mathbf{x},\mathbf{y}) &= |\mathbf{x}+\mathbf{y}|^2 = |q(\mathbf{x}+\mathbf{y})|^2 \\ &= |q(\mathbf{x})|^2 + |q(\mathbf{y})|^2 + 2\operatorname{Re}(q(\mathbf{x}), q(\mathbf{y})) \end{aligned} \tag{14.11}$$

Now in any inner product space,

$$(x, iy) = \operatorname{Re}(x, iy) + i\operatorname{Im}(x, iy).$$

Also

$$(x, iy) = (-i)(x, y) = (-i)\operatorname{Re}(x, y) + \operatorname{Im}(x, y).$$

Therefore, equating the real parts, $\operatorname{Im}(x, y) = \operatorname{Re}(x, iy)$ and so

$$(x, y) = \operatorname{Re}(x, y) + i\operatorname{Re}(x, iy) \tag{14.12}$$

Now from 14.11, since q preserves distances, $.\operatorname{Re}(q(\mathbf{x}), q(\mathbf{y})) = \operatorname{Re}(\mathbf{x},\mathbf{y})$ which implies from 14.12 that

$$(\mathbf{x},\mathbf{y}) = (q(\mathbf{x}), q(\mathbf{y})). \tag{14.13}$$

Now consulting the diagram which gives the meaning for the matrix of a linear transformation, observe that $q \circ M(A) = A \circ q$ and $q \circ M(A^*) = A^* \circ q$. Therefore, from 14.13

$$(A(q(\mathbf{x})), q(\mathbf{y})) = (q(\mathbf{x}), A^* q(\mathbf{y})) = (q(\mathbf{x}), q(M(A^*)(\mathbf{y}))) = (\mathbf{x}, M(A^*)(\mathbf{y}))$$

but also

$$(A(q(\mathbf{x})), q(\mathbf{y})) = (q(M(A)(\mathbf{x})), q(\mathbf{y})) = (M(A)(\mathbf{x}), \mathbf{y}) = \left(\mathbf{x}, M(A)^*(\mathbf{y})\right).$$

Since $\mathbf{x}, \mathbf{y}$ are arbitrary, this shows that $M(A^*) = M(A)^*$ as claimed. Therefore, if A is self adjoint, $M(A) = M(A^*) = M(A)^*$ and so $M(A)$ is also self adjoint. If $M(A) = M(A)^*$ then $M(A) = M(A^*)$ and so $A = A^*$. This proves the lemma.

The following corollary is one of the items in the above proof.

Corollary 14.3.2 *Let X be a finite dimensional Hilbert space and let $\{v_1, \cdots, v_n\}$ be an orthonormal basis for X. Also, let q be the coordinate map associated with this basis satisfying $q(\mathbf{x}) \equiv \sum_i x_i v_i$. Then $(\mathbf{x},\mathbf{y})_{\mathbb{F}^n} = (q(\mathbf{x}), q(\mathbf{y}))_X$. Also, if $A \in \mathcal{L}(X, X)$, and $M(A)$ is the matrix of A with respect to this basis,*

$$(Aq(\mathbf{x}), q(\mathbf{y}))_X = (M(A)\mathbf{x}, \mathbf{y})_{\mathbb{F}^n}.$$

Definition 14.3.3 *A self adjoint $A \in \mathcal{L}(X, X)$, is positive definite if whenever $\mathbf{x} \neq \mathbf{0}$, $(A\mathbf{x}, \mathbf{x}) > 0$ and A is negative definite if for all $\mathbf{x} \neq \mathbf{0}$, $(A\mathbf{x}, \mathbf{x}) < 0$. A is positive semidefinite or just nonnegative for short if for all $\mathbf{x}$, $(A\mathbf{x}, \mathbf{x}) \geq 0$. A is negative semidefinite or nonpositive for short if for all $\mathbf{x}$, $(A\mathbf{x}, \mathbf{x}) \leq 0$.*

The following lemma is of fundamental importance in determining which linear transformations are positive or negative definite.

Lemma 14.3.4 *Let X be a finite dimensional Hilbert space. A self adjoint $A \in \mathcal{L}(X,X)$ is positive definite if and only if all its eigenvalues are positive and negative definite if and only if all its eigenvalues are negative. It is positive semidefinite if all the eigenvalues are nonnegative and it is negative semidefinite if all the eigenvalues are nonpositive.*

Proof: Suppose first that A is positive definite and let λ be an eigenvalue. Then for $\mathbf{x}$ an eigenvector corresponding to λ, $\lambda(\mathbf{x},\mathbf{x}) = (\lambda\mathbf{x},\mathbf{x}) = (A\mathbf{x},\mathbf{x}) > 0$. Therefore, $\lambda > 0$ as claimed.

Now suppose all the eigenvalues of A are positive. From Theorem 14.2.3 and Corollary 14.2.6, $A = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \otimes \mathbf{u}_i$ where the λ_i are the positive eigenvalues and $\{\mathbf{u}_i\}$ are an orthonormal set of eigenvectors. Therefore, letting $\mathbf{x} \neq \mathbf{0}$, $(A\mathbf{x},\mathbf{x}) = ((\sum_{i=1}^{n} \lambda_i \mathbf{u}_i \otimes \mathbf{u}_i)\mathbf{x},\mathbf{x}) = (\sum_{i=1}^{n} \lambda_i (\mathbf{x},\mathbf{u}_i)(\mathbf{u}_i,\mathbf{x})) = \sum_{i=1}^{n} \lambda_i \left|(\mathbf{u}_i,\mathbf{x})^2\right| > 0$ because, since $\{\mathbf{u}_i\}$ is an orthonormal basis, $|\mathbf{x}|^2 = \sum_{i=1}^{n} \left|(\mathbf{u}_i,\mathbf{x})^2\right|$.

To establish the claim about negative definite, it suffices to note that A is negative definite if and only if $-A$ is positive definite and the eigenvalues of A are (-1) times the eigenvalues of $-A$. The claims about positive semidefinite and negative semidefinite are obtained similarly. This proves the lemma.

The next theorem is about a way to recognize whether a self adjoint $A \in \mathcal{L}(X,X)$ is positive or negative definite without having to find the eigenvalues. In order to state this theorem, here is some notation.

Definition 14.3.5 *Let A be an $n \times n$ matrix. Denote by A_k the $k \times k$ matrix obtained by deleting the $k+1, \cdots, n$ columns and the $k+1, \cdots, n$ rows from A. Thus $A_n = A$ and A_k is the $k \times k$ submatrix of A which occupies the upper left corner of A.*

The following theorem is proved in [5]

Theorem 14.3.6 *Let X be a finite dimensional Hilbert space and let $A \in \mathcal{L}(X,X)$ be self adjoint. Then A is positive definite if and only if $\det(M(A)_k) > 0$ for every $k = 1, \cdots, n$. Here $M(A)$ denotes the matrix of A with respect to some fixed orthonormal basis of X.*

Proof: This theorem is proved by induction on n. It is clearly true if $n = 1$. Suppose then that it is true for $n-1$ where $n \geq 2$. Since $\det(M(A)) > 0$, it follows that all the eigenvalues are nonzero. Are they all positive? Suppose not. Then there is some even number of them which are negative, even because the product of all the eigenvalues is known to be positive, equaling $\det(M(A))$. Pick two, λ_1 and λ_2 and let $M(A)\mathbf{u}_i = \lambda_i \mathbf{u}_i$ where $\mathbf{u}_i \neq \mathbf{0}$ for $i = 1, 2$ and $(\mathbf{u}_1,\mathbf{u}_2) = 0$. Now if $\mathbf{y} \equiv \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2$ is an element of $\operatorname{span}(\mathbf{u}_1,\mathbf{u}_2)$, then since these are eigenvalues and $(\mathbf{u}_1,\mathbf{u}_2) = 0$, a short computation shows

$$\begin{aligned} &(M(A)(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2), \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2) \\ &= |\alpha_1|^2 \lambda_1 |\mathbf{u}_1|^2 + |\alpha_2|^2 \lambda_2 |\mathbf{u}_2|^2 < 0. \end{aligned}$$

Now letting $\mathbf{x} \in \mathbb{C}^{n-1}$, the induction hypothesis implies

$$(\mathbf{x}^*, 0) M(A) \begin{pmatrix} \mathbf{x} \\ 0 \end{pmatrix} = \mathbf{x}^* M(A)_{n-1} \mathbf{x} = (M(A)\mathbf{x},\mathbf{x}) > 0.$$

Now the dimension of $\{\mathbf{z} \in \mathbb{C}^n : z_n = 0\}$ is $n-1$ and the dimension of $\operatorname{span}(\mathbf{u}_1,\mathbf{u}_2) = 2$ and so there must be some nonzero $\mathbf{x} \in \mathbb{C}^n$ which is in both of these subspaces of $\mathbb{C}^n$. However,

the first computation would require that $(M(A)\mathbf{x},\mathbf{x}) < 0$ while the second would require that $(M(A)\mathbf{x},\mathbf{x}) > 0$. This contradiction shows that all the eigenvalues must be positive. This proves the if part of the theorem. The only if part is left to the reader.

Corollary 14.3.7 *Let X be a finite dimensional Hilbert space and let $A \in \mathcal{L}(X,X)$ be self adjoint. Then A is negative definite if and only if $\det(M(A)_k)(-1)^k > 0$ for every $k = 1,\cdots,n$. Here $M(A)$ denotes the matrix of A with respect to some fixed orthonormal basis of X.*

Proof: This is immediate from the above theorem by noting that, as in the proof of Lemma 14.3.4, A is negative definite if and only if $-A$ is positive definite. Therefore, if $\det(-M(A)_k) > 0$ for all $k = 1,\cdots,n$, it follows that A is negative definite. However, $\det(-M(A)_k) = (-1)^k \det(M(A)_k)$. This proves the corollary.

14.4 Fractional Powers

With the above theory, it is possible to take fractional powers of certain elements of $\mathcal{L}(X,X)$ where X is a finite dimensional Hilbert space. The main result is the following theorem.

Theorem 14.4.1 *Let $A \in \mathcal{L}(X,X)$ be self adjoint and nonnegative and let k be a positive integer. Then there exists a unique self adjoint nonnegative $B \in \mathcal{L}(X,X)$ such that $B^k = A$.*

Proof: By Theorem 14.2.3, there exists an orthonormal basis of eigenvectors of A, say $\{v_i\}_{i=1}^n$ such that $Av_i = \lambda_i v_i$. Therefore, by Corollary 14.2.6, $A = \sum_i \lambda_i v_i \otimes v_i$.

Now by Lemma 14.3.4, each $\lambda_i \geq 0$. Therefore, it makes sense to define

$$B \equiv \sum_i \lambda_i^{1/k} v_i \otimes v_i.$$

It is easy to verify that

$$(v_i \otimes v_i)(v_j \otimes v_j) = \begin{cases} 0 \text{ if } i \neq j \\ v_i \otimes v_i \text{ if } i = j \end{cases}.$$

Therefore, a short computation verifies that $B^k = \sum_i \lambda_i v_i \otimes v_i = A$. This proves existence.

In order to prove uniqueness, let $p(t)$ be a polynomial which has the property that $p(\lambda_i) = \lambda_i^{1/k}$. Then a similar short computation shows

$$p(A) = \sum_i p(\lambda_i) v_i \otimes v_i = \sum_i \lambda_i^{1/k} v_i \otimes v_i = B.$$

Now suppose $C^k = A$ where $C \in \mathcal{L}(X,X)$ is self adjoint and nonnegative. Then

$$CB = Cp(A) = Cp\left(C^k\right) = p\left(C^k\right)C = BC.$$

Therefore, $\{B,C\}$ is a commuting family of linear transformations which are both self adjoint. Letting $M(B)$ and $M(C)$ denote matrices of these linear transformations taken with respect to some fixed orthonormal basis, $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$, it follows that $M(B)$ and $M(C)$ commute and that both can be diagonalized (Lemma 14.3.1). See the diagram for a short verification of the claim the two matrices commute..

$$\begin{array}{ccccc} & B & & C & \\ X & \rightarrow & X & \rightarrow & X \\ q\uparrow & \circ & \uparrow q & \circ & \uparrow q \\ \mathbb{F}^n & \rightarrow & \mathbb{F}^n & \rightarrow & \mathbb{F}^n \\ & M(B) & & M(C) & \end{array}$$

Therefore, by Theorem 14.1.7, these two matrices can be simultaneously diagonalized. Thus

$$U^{-1}M(B)U = D_1,\ U^{-1}M(C)U = D_2$$

where the D_i is a diagonal matrix consisting of the eigenvalues of B. Then raising these to powers,

$$U^{-1}M(A)U = U^{-1}M(B)^k U = D_1^k$$

and

$$U^{-1}M(A)U = U^{-1}M(C)^k U = D_2^k.$$

Therefore, $D_1^k = D_2^k$ and since the diagonal entries of D_i are nonnegative, this requires that $D_1 = D_2$. Therefore, $M(B) = M(C)$ and so $B = C$. This proves the theorem.

14.5 Polar Decompositions

An application of Theorem 14.2.3, is the following fundamental result, important in geometric measure theory and continuum mechanics. It is sometimes called the right polar decomposition. The notation used is that which is seen in continuum mechanics, see for example Gurtin [6]. Don't confuse the U in this theorem with a unitary transformation. It is not so. When the following theorem is applied in continuum mechanics, F is normally the deformation gradient, the derivative of a nonlinear map from some subset of three dimensional space to three dimensional space. In this context, U is called the right Cauchy Green strain tensor. It is a measure of how a body is stretched independent of rigid motions.

Theorem 14.5.1 *Let X be a Hilbert space of dimension n and let Y be a Hilbert space of dimension $m \geq n$ and let $F \in \mathcal{L}(X,Y)$. Then there exists $R \in \mathcal{L}(X,Y)$ and $U \in \mathcal{L}(X,X)$ such that*

$$F = RU,\ U = U^*,$$

all eigenvalues of U are non negative,

$$U^2 = F^*F, R^*R = I,$$

and $|R\mathbf{x}| = |\mathbf{x}|$.

Proof: $(F^*F)^* = F^*F$ and so by Theorem 14.2.3, there is an orthonormal basis of eigenvectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ such that

$$F^*F\mathbf{v}_i = \lambda_i \mathbf{v}_i.$$

It is also clear that $\lambda_i \geq 0$ because

$$\lambda_i(\mathbf{v}_i, \mathbf{v}_i) = (F^*F\mathbf{v}_i, \mathbf{v}_i) = (F\mathbf{v}_i, F\mathbf{v}_i) \geq 0.$$

Let

$$U \equiv \sum_{i=1}^{n} \lambda_i^{1/2} \mathbf{v}_i \otimes \mathbf{v}_i.$$

Then $U^2 = F^*F$, $U = U^*$, and the eigenvalues of U, $\left\{\lambda_i^{1/2}\right\}_{i=1}^n$ are all non negative.

Now R is defined on $U(X)$ by

$$RU\mathbf{x} \equiv F\mathbf{x}.$$

This is well defined because if $U\mathbf{x}_1 = U\mathbf{x}_2$, then $U^2(\mathbf{x}_1 - \mathbf{x}_2) = 0$ and so

$$0 = (F^*F(\mathbf{x}_1 - \mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2) = |F(\mathbf{x}_1 - \mathbf{x}_2)|^2.$$

Now $|RU\mathbf{x}|^2 = |U\mathbf{x}|^2$ because

$$|RU\mathbf{x}|^2 = |F\mathbf{x}|^2 = (F\mathbf{x},F\mathbf{x})$$

$$= (F^*F\mathbf{x},\mathbf{x}) = \left(U^2\mathbf{x},\mathbf{x}\right) = (U\mathbf{x},U\mathbf{x}) = |U\mathbf{x}|^2.$$

Let $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$ be an orthonormal basis for

$$U(X)^\perp \equiv \{\mathbf{x}\in X : (\mathbf{x},\mathbf{z}) = 0 \text{ for all } \mathbf{z}\in U(X)\}$$

and let $\{\mathbf{y}_1,\cdots,\mathbf{y}_p\}$ be an orthonormal basis for $F(X)^\perp$. Then $p\geq r$ because if $\{F(\mathbf{z}_i)\}_{i=1}^s$ is an orthonormal basis for $F(X)$, it follows that $\{U(\mathbf{z}_i)\}_{i=1}^s$ is orthonormal in $U(X)$ because

$$(U\mathbf{z}_i,U\mathbf{z}_j) = \left(U^2\mathbf{z}_i,\mathbf{z}_j\right) = (F^*F\mathbf{z}_i,\mathbf{z}_j) = (F\mathbf{z}_i,F\mathbf{z}_j).$$

Therefore,

$$p+s = m \geq n = r + \dim U(X) \geq r+s.$$

Now define $R\in\mathcal{L}(X,Y)$ by $R\mathbf{x}_i \equiv \mathbf{y}_i$, $i=1,\cdots,r$. Note that R is already defined on $U(X)$. It has been extended by telling what R does to a basis for $U(X)^\perp$. Thus

$$\left|R\left(\sum_{i=1}^r c_i\mathbf{x}_i + U\mathbf{v}\right)\right|^2 = \left|\sum_{i=1}^r c_i\mathbf{y}_i + F\mathbf{v}\right|^2 = \sum_{i=1}^r |c_i|^2 + |F\mathbf{v}|^2$$

$$= \sum_{i=1}^r |c_i|^2 + |U\mathbf{v}|^2 = \left|\sum_{i=1}^r c_i\mathbf{x}_i + U\mathbf{v}\right|^2,$$

and so $|R\mathbf{z}| = |\mathbf{z}|$ which implies that for all $\mathbf{x},\mathbf{y}$,

$$|\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\,\mathrm{Re}\,(\mathbf{x},\mathbf{y}) = |\mathbf{x}+\mathbf{y}|^2$$

$$= |R(\mathbf{x}+\mathbf{y})|^2 = |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\,\mathrm{Re}\,(R\mathbf{x},R\mathbf{y}).$$

Therefore, as in Lemma 14.3.1,

$$(\mathbf{x},\mathbf{y}) = (R\mathbf{x},R\mathbf{y}) = (R^*R\mathbf{x},\mathbf{y})$$

for all $\mathbf{x},\mathbf{y}$ and so $R^*R = I$ as claimed. This proves the theorem.

The following corollary follows as a simple consequence of this theorem. It is called the left polar decomposition.

Corollary 14.5.2 *Let $F\in\mathcal{L}(X,Y)$ and suppose $n\geq m$ where X is a Hilbert space of dimension n and Y is a Hilbert space of dimension m. Then there exists a symmetric nonnegative element of $\mathcal{L}(X,X)$, U, and an element of $\mathcal{L}(X,Y)$, R, such that*

$$F = UR,\ RR^* = I.$$

Proof: Recall that $L^{**} = L$ and $(ML)^* = L^*M^*$. Now apply Theorem 14.5.1 to $F^*\in\mathcal{L}(X,Y)$. Thus,

$$F^* = R^*U$$

where R^* and U satisfy the conditions of that theorem. Then

$$F = UR$$

and $RR^* = R^{**}R^* = I$. This proves the corollary.

The following existence theorem for the polar decomposition of an element of $\mathcal{L}(X,X)$ is a corollary.

Corollary 14.5.3 *Let $F \in \mathcal{L}(X,X)$. Then there exists a symmetric nonnegative element of $\mathcal{L}(X,X)$, W, and a unitary matrix, Q such that $F = WQ$, and there exists a symmetric nonnegative element of $\mathcal{L}(X,X)$, U, and a unitary R, such that $F = RU$.*

This corollary has a fascinating relation to the question whether a given linear transformation is normal. Recall that an $n \times n$ matrix, A, is normal if $AA^* = A^*A$. Retain the same definition for an element of $\mathcal{L}(X,X)$.

Theorem 14.5.4 *Let $F \in \mathcal{L}(X,X)$. Then F is normal if and only if in Corollary 14.5.3 $RU = UR$ and $QW = WQ$.*

Proof: I will prove the statement about $RU = UR$ and leave the other part as an exercise. First suppose that $RU = UR$ and show F is normal. To begin with,

$$UR^* = (RU)^* = (UR)^* = R^*U.$$

Therefore,

$$\begin{aligned} F^*F &= UR^*RU = U^2 \\ FF^* &= RUUR^* = URR^*U = U^2 \end{aligned}$$

which shows F is normal.

Now suppose F is normal. Is $RU = UR$? Since F is normal,

$$FF^* = RUUR^* = RU^2R^*$$

and

$$F^*F = UR^*RU = U^2.$$

Therefore, $RU^2R^* = U^2$, and both are nonnegative and self adjoint. Therefore, the square roots of both sides must be equal by the uniqueness part of the theorem on fractional powers. It follows that the square root of the first, RUR^* must equal the square root of the second, U. Therefore, $RUR^* = U$ and so $RU = UR$. This proves the theorem in one case. The other case in which W and Q commute is left as an exercise.

14.6 The Singular Value Decomposition

In this section, A will be an $m \times n$ matrix. To begin with, here is a simple lemma.

Lemma 14.6.1 *Let A be an $m \times n$ matrix. Then A^*A is self adjoint and all its eigenvalues are nonnegative.*

Proof: It is obvious that A^*A is self adjoint. Suppose $A^*A\mathbf{x} = \lambda\mathbf{x}$. Then $\lambda|\mathbf{x}|^2 = (\lambda\mathbf{x},\mathbf{x}) = (A^*A\mathbf{x},\mathbf{x}) = (A\mathbf{x},A\mathbf{x}) \geq 0$.

Definition 14.6.2 *Let A be an $m \times n$ matrix. The singular values of A are the square roots of the positive eigenvalues of A^*A.*

With this definition and lemma here is the main theorem on the singular value decomposition.

Theorem 14.6.3 *Let A be an $m \times n$ matrix. Then there exist unitary matrices, U and V of the appropriate size such that*

$$U^* A V = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

where σ is of the form

$$\sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{pmatrix}$$

for the σ_i the singular values of A.

Proof: By the above lemma and Theorem 14.2.3 there exists an orthonormal basis, $\{\mathbf{v}_i\}_{i=1}^n$ such that $A^* A \mathbf{v}_i = \sigma_i^2 \mathbf{v}_i$ where $\sigma_i^2 > 0$ for $i = 1, \cdots, k, (\sigma_i > 0)$, and equals zero if $i > k$. Thus for $i > k$, $A\mathbf{v}_i = \mathbf{0}$ because

$$(A\mathbf{v}_i, A\mathbf{v}_i) = (A^* A \mathbf{v}_i, \mathbf{v}_i) = (\mathbf{0}, \mathbf{v}_i) = 0.$$

For $i = 1, \cdots, k$, define $\mathbf{u}_i \in \mathbb{F}^m$ by

$$\mathbf{u}_i \equiv \sigma_i^{-1} A \mathbf{v}_i.$$

Thus $A\mathbf{v}_i = \sigma_i \mathbf{u}_i$. Now

$$\begin{aligned} (\mathbf{u}_i, \mathbf{u}_j) &= \left(\sigma_i^{-1} A \mathbf{v}_i, \sigma_j^{-1} A \mathbf{v}_j\right) = \left(\sigma_i^{-1} \mathbf{v}_i, \sigma_j^{-1} A^* A \mathbf{v}_j\right) \\ &= \left(\sigma_i^{-1} \mathbf{v}_i, \sigma_j^{-1} \sigma_j^2 \mathbf{v}_j\right) = \frac{\sigma_j}{\sigma_i} (\mathbf{v}_i, \mathbf{v}_j) = \delta_{ij}. \end{aligned}$$

Thus $\{\mathbf{u}_i\}_{i=1}^k$ is an orthonormal set of vectors in $\mathbb{F}^m$. Also,

$$A A^* \mathbf{u}_i = A A^* \sigma_i^{-1} A \mathbf{v}_i = \sigma_i^{-1} A A^* A \mathbf{v}_i = \sigma_i^{-1} A \sigma_i^2 \mathbf{v}_i = \sigma_i^2 \mathbf{u}_i.$$

Now extend $\{\mathbf{u}_i\}_{i=1}^k$ to an orthonormal basis for all of $\mathbb{F}^m$, $\{\mathbf{u}_i\}_{i=1}^m$ and let $U \equiv (\mathbf{u}_1 \cdots \mathbf{u}_m)$ while $V \equiv (\mathbf{v}_1 \cdots \mathbf{v}_n)$. Thus U is the matrix which has the $\mathbf{u}_i$ as columns and V is defined as the matrix which has the $\mathbf{v}_i$ as columns. Then

$$\begin{aligned} U^* A V &= \begin{pmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_k^* \\ \vdots \\ \mathbf{u}_m^* \end{pmatrix} A (\mathbf{v}_1 \cdots \mathbf{v}_n) \\ &= \begin{pmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_k^* \\ \vdots \\ \mathbf{u}_m^* \end{pmatrix} (\sigma_1 \mathbf{u}_1 \cdots \sigma_k \mathbf{u}_k \mathbf{0} \cdots \mathbf{0}) \\ &= \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

where σ is given in the statement of the theorem.

The singular value decomposition has as an immediate corollary the following interesting result.

Corollary 14.6.4 *Let A be an $m \times n$ matrix. Then the rank of A and A^* equals the number of singular values.*

Proof: Since V and U are unitary, it follows that

$$\begin{aligned} \operatorname{rank}(A) &= \operatorname{rank}(U^*AV) \\ &= \operatorname{rank}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \\ &= \text{number of singular values.} \end{aligned}$$

Also since U, V are unitary,

$$\begin{aligned} \operatorname{rank}(A^*) &= \operatorname{rank}(V^*A^*U) \\ &= \operatorname{rank}\left((U^*AV)^*\right) \\ &= \operatorname{rank}\left(\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}^*\right) \\ &= \text{number of singular values.} \end{aligned}$$

This proves the corollary.

The singular value decomposition also has a very interesting connection to the problem of least squares solutions. Recall that it was desired to find $\mathbf{x}$ such that $|A\mathbf{x} - \mathbf{y}|$ is as small as possible. Lemma 13.1.1 shows that there is a solution to this problem which can be found by solving the system $A^*A\mathbf{x} = A^*\mathbf{y}$. Each $\mathbf{x}$ which solves this system solves the minimization problem as was shown in the lemma just mentioned. Now consider this equation for the solutions of the minimization problem in terms of the singular value decomposition.

$$\overbrace{V\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}U^*}^{A^*}\overbrace{U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}V^*}^{A}\mathbf{x} = \overbrace{V\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}U^*}^{A^*}\mathbf{y}.$$

Therefore, this yields the following upon using block multiplication and multiplying on the left by V^*.

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix}V^*\mathbf{x} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}U^*\mathbf{y}. \tag{14.14}$$

One solution to this equation which is very easy to spot is

$$\mathbf{x} = V\begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix}U^*\mathbf{y}. \tag{14.15}$$

14.7 The Moore Penrose Inverse

This particular solution is important enough that it motivates the following definintion.

Definition 14.7.1 *Let A be an $m \times n$ matrix. Then the Moore Penrose inverse of A, denoted by A^+ is defined as*

$$A^+ \equiv V\begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix}U^*.$$

Thus $A^+\mathbf{y}$ is a solution to the minimization problem to find $\mathbf{x}$ which minimizes $|A\mathbf{x} - \mathbf{y}|$. In fact, one can say more about this.

Proposition 14.7.2 *$A^+\mathbf{y}$ is the solution to the problem of minimizing $|A\mathbf{x}-\mathbf{y}|$ for all $\mathbf{x}$ which has smallest norm. Thus*

$$\left|AA^+\mathbf{y}-\mathbf{y}\right| \leq |A\mathbf{x}-\mathbf{y}| \text{ for all } \mathbf{x}$$

and if $\mathbf{x}_1$ satisfies $|A\mathbf{x}_1-\mathbf{y}| \leq |A\mathbf{x}-\mathbf{y}|$ for all $\mathbf{x}$, then $|A^+\mathbf{y}| \leq |\mathbf{x}_1|$.

Proof: Consider $\mathbf{x}$ satisfying 14.14 which has smallest norm. This is equivalent to making $|V^*\mathbf{x}|$ as small as possible because V^* is unitary and so it preserves norms. For $\mathbf{z}$ a vector, denote by $(\mathbf{z})_k$ the vector in $\mathbb{F}^k$ which consists of the first k entries of $\mathbf{z}$. Then if $\mathbf{x}$ is a solution to 14.14

$$\begin{pmatrix} \sigma^2 (V^*\mathbf{x})_k \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \sigma (U^*\mathbf{y})_k \\ \mathbf{0} \end{pmatrix}$$

and so $(V^*\mathbf{x})_k = \sigma^{-1}(U^*\mathbf{y})_k$. Thus the first k entries of $V^*\mathbf{x}$ are determined. In order to make $|V^*\mathbf{x}|$ as small as possible, the remaining $n-k$ entries should equal zero. Therefore,

$$V^*\mathbf{x} = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y}$$

which shows that $A^+\mathbf{y} = \mathbf{x}$. This proves the proposition.

Lemma 14.7.3 *The matrix, A^+ satisfies the following conditions.*

$$AA^+A = A,\ A^+AA^+ = A^+,\ A^+A \text{ and } AA^+ \text{ are Hermitian.} \tag{14.16}$$

Proof: The proof is completely routine and is left to the reader.

A much more interesting observation is that A^+ is characterized as being the unique matrix which satisfies 14.16. This is the content of the following Theorem.

Theorem 14.7.4 *Let A be an $m \times n$ matrix. Then a matrix, A_0, is the Moore Penrose inverse of A if and only if A_0 satisfies*

$$AA_0A = A,\ A_0AA_0 = A_0,\ A_0A \text{ and } AA_0 \text{ are Hermitian.} \tag{14.17}$$

Proof: From the above lemma, the Moore Penrose inverse satisfies 14.17. Suppose then that A_0 satisfies 14.17. It is necessary to verify $A_0 = A^+$. Recall that from the singular value decomposition, there exist unitary matrices, U and V such that

$$U^*AV = \Sigma \equiv \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix},\ A = U\Sigma V^*.$$

Let

$$V^*A_0U = \begin{pmatrix} P & Q \\ R & S \end{pmatrix} \tag{14.18}$$

where P is $k \times k$.

Next use the first equation of 14.17 to write

$$\overbrace{U\Sigma V^*}^{A} \overbrace{V \begin{pmatrix} P & Q \\ R & S \end{pmatrix} U^*}^{A_0} \overbrace{U\Sigma V^*}^{A} = \overbrace{U\Sigma V^*}^{A}.$$

Then multiplying both sides on the left by V^* and on the right by U,

$$\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

Now this requires

$$\begin{pmatrix} \sigma P \sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}. \tag{14.19}$$

Therefore, $P = \sigma^{-1}$. Now from the requirement that AA_0 is Hermitian,

$$\overbrace{U\Sigma V^*}^{A}\overbrace{V\begin{pmatrix} P & Q \\ R & S \end{pmatrix}U^*}^{A_0} = U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} P & Q \\ R & S \end{pmatrix}U^*$$

must be Hermitian. Therefore, it is necessary that

$$\begin{aligned}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} P & Q \\ R & S \end{pmatrix} &= \begin{pmatrix} \sigma P & \sigma Q \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} I & \sigma Q \\ 0 & 0 \end{pmatrix}\end{aligned}$$

is Hermitian. Then

$$\begin{pmatrix} I & \sigma Q \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ \left(\overline{Q}\right)^T \sigma & 0 \end{pmatrix}$$

which requires that $Q = 0$. From the requirement that A_0A is Hermitian, it is necessary that

$$\begin{aligned}\overbrace{V\begin{pmatrix} P & Q \\ R & S \end{pmatrix}U^*}^{A_0}\overbrace{U\Sigma V^*}^{A} &= V\begin{pmatrix} P\sigma & 0 \\ R\sigma & 0 \end{pmatrix}V^* \\ &= V\begin{pmatrix} I & 0 \\ R\sigma & 0 \end{pmatrix}V^*\end{aligned}$$

is Hermitian. Therefore, also

$$\begin{pmatrix} I & 0 \\ R\sigma & 0 \end{pmatrix}$$

is Hermitian. Thus $R = 0$ by reasoning similar to that used to show $Q = 0$.

Use 14.18 and the second equation of 14.17 to write

$$\overbrace{V\begin{pmatrix} P & Q \\ R & S \end{pmatrix}U^*}^{A_0}\overbrace{U\Sigma V^*}^{A}\overbrace{V\begin{pmatrix} P & Q \\ R & S \end{pmatrix}U^*}^{A_0} = \overbrace{V\begin{pmatrix} P & Q \\ R & S \end{pmatrix}U^*}^{A_0}.$$

which implies

$$\begin{pmatrix} P & Q \\ R & S \end{pmatrix}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} P & Q \\ R & S \end{pmatrix} = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}.$$

This yields

$$\begin{aligned}\begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix}\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix} &= \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} && (14.20) \\ &= \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & S \end{pmatrix}. && (14.21)\end{aligned}$$

Therefore, $S = 0$ also and so

$$V^*A_0U = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

which says

$$A_0 = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^* \equiv A^+.$$

This proves the theorem.

The theorem is significant because there is no mention of eigenvalues or eigenvectors in the characterization of the Moore Penrose inverse given in 14.17. It also shows immediately that the Moore Penrose inverse is a generalization of the usual inverse. See Problem 4.

14.8 Exercises

1. Show $(A^*)^* = A$ and $(AB)^* = B^*A^*$.

2. Suppose $A : X \to X$, an inner product space, and $A \geq 0$. This means $(Ax, x) \geq 0$ for all $x \in X$ and $A = A^*$. Show that A has a square root, U, such that $U^2 = A$. **Hint:** Let $\{u_k\}_{k=1}^n$ be an orthonormal basis of eigenvectors with $Au_k = \lambda_k u_k$. Show each $\lambda_k \geq 0$ and consider
$$U \equiv \sum_{k=1}^{n} \lambda_k^{1/2} u_k \otimes u_k$$

3. Prove Corollary 14.2.9.

4. Show that if A is an $n \times n$ matrix which has an inverse then $A^+ = A^{-1}$.

5. Using the singular value decomposition, show that for any square matrix, A, it follows that A^*A is unitarily similar to AA^*.

6. Let A, B be a $m \times n$ matrices. Define an inner product on the set of $m \times n$ matrices by
$$(A, B)_F \equiv \operatorname{trace}(AB^*).$$
Show this is an inner product satisfying all the inner product axioms. Recall for M an $n \times n$ matrix, $\operatorname{trace}(M) \equiv \sum_{i=1}^n M_{ii}$. The resulting norm, $||\cdot||_F$ is called the Frobenius norm and it can be used to measure the distance between two matrices.

7. Let A be an $m \times n$ matrix. Show
$$||A||_F^2 \equiv (A, A)_F = \sum_j \sigma_j^2$$
where the σ_j are the singular values of A.

8. Prove that Theorem 14.3.6 and Corollary 14.3.7 can be strengthened so that the condition on the A_k is necessary as well as sufficient. **Hint:** Consider vectors of the form $\begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix}$ where $\mathbf{x} \in \mathbb{F}^k$.

9. Show directly that if A is an $n \times n$ matrix and $A = A^*$ (A is Hermitian) then all the eigenvalues and eigenvectors are real and that eigenvectors associated with distinct eigenvalues are orthogonal, (their inner product is zero).

10. Let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ be an orthonormal basis for $\mathbb{F}^n$. Let Q be a matrix whose i^{th} column is $\mathbf{v}_i$. Show
$$Q^*Q = QQ^* = I.$$

11. Show that a matrix, Q is unitary if and only if it preserves distances. This means $|Q\mathbf{v}| = |\mathbf{v}|$.

12. Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \cdots, \mathbf{w}_n\}$ are two orthonormal bases for $\mathbb{F}^n$ and suppose Q is an $n \times n$ matrix satisfying $Q\mathbf{v}_i = \mathbf{w}_i$. Then show Q is unitary. If $|\mathbf{v}| = 1$, show there is a unitary transformation which maps $\mathbf{v}$ to $\mathbf{e}_1$.

13. Finish the proof of Theorem 14.5.4.

14. Let A be a Hermitian matrix so $A = A^*$ and suppose all eigenvalues of A are larger than δ^2. Show
$$(A\mathbf{v}, \mathbf{v}) \geq \delta^2 |\mathbf{v}|^2$$
Where here, the inner product is
$$(\mathbf{v}, \mathbf{u}) \equiv \sum_{j=1}^{n} v_j \overline{u_j}.$$

15. Let X be an inner product space. Show $|x+y|^2 + |x-y|^2 = 2|x|^2 + 2|y|^2$. This is called the parallelogram identity.

Norms For Finite Dimensional Vector Spaces

In this chapter, X and Y are finite dimensional vector spaces which have a norm. The following is a definition.

Definition 15.0.1 *A linear space X is a normed linear space if there is a norm defined on X, $||\cdot||$ satisfying*

$$||\mathbf{x}|| \geq 0, \;\; ||\mathbf{x}|| = 0 \text{ if and only if } \mathbf{x} = 0,$$

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

$$||c\mathbf{x}|| = |c| \, ||\mathbf{x}||$$

whenever c is a scalar. A set, $U \subseteq X$, a normed linear space is open if for every $p \in U$, there exists $\delta > 0$ such that

$$B(p, \delta) \equiv \{x : ||x - p|| < \delta\} \subseteq U.$$

Thus, a set is open if every point of the set is an interior point.

To begin with recall the Cauchy Schwarz inequality which is stated here for convenience in terms of the inner product space, $\mathbb{C}^n$.

Theorem 15.0.2 *The following inequality holds for a_i and $b_i \in \mathbb{C}$.*

$$\left| \sum_{i=1}^{n} a_i \overline{b}_i \right| \leq \left(\sum_{i=1}^{n} |a_i|^2 \right)^{1/2} \left(\sum_{i=1}^{n} |b_i|^2 \right)^{1/2}. \tag{15.1}$$

Definition 15.0.3 *Let $(X, ||\cdot||)$ be a normed linear space and let $\{x_n\}_{n=1}^{\infty}$ be a sequence of vectors. Then this is called a Cauchy sequence if for all $\varepsilon > 0$ there exists N such that if $m, n \geq N$, then*

$$||x_n - x_m|| < \varepsilon.$$

This is written more briefly as

$$\lim_{m,n \to \infty} ||x_n - x_m|| = 0.$$

Definition 15.0.4 *A normed linear space, $(X, ||\cdot||)$ is called a Banach space if it is complete. This means that, whenever, $\{\mathbf{x}_n\}$ is a Cauchy sequence there exists a unique $\mathbf{x} \in X$ such that $\lim_{n \to \infty} ||\mathbf{x} - \mathbf{x}_n|| = 0$.*

Let X be a finite dimensional normed linear space with norm $||\cdot||$ where the field of scalars is denoted by $\mathbb{F}$ and is understood to be either $\mathbb{R}$ or $\mathbb{C}$. Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis for X. If $\mathbf{x} \in X$, denote by x_i the i^{th} component of $\mathbf{x}$ with respect to this basis. Thus

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{v}_i.$$

Definition 15.0.5 *For $\mathbf{x} \in X$ and $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ a basis, define a new norm by*

$$|\mathbf{x}| \equiv \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}.$$

Similarly, for $\mathbf{y} \in Y$ with basis $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$, and y_i its components with respect to this basis,

$$|\mathbf{y}| \equiv \left(\sum_{i=1}^{m} |y_i|^2 \right)^{1/2}$$

For $A \in \mathcal{L}(X, Y)$, the space of linear mappings from X to Y,

$$||A|| \equiv \sup\{|A\mathbf{x}| : |\mathbf{x}| \leq 1\}. \tag{15.2}$$

The first thing to show is that the two norms, $||\cdot||$ and $|\cdot|$, are equivalent. This means the conclusion of the following theorem holds.

Theorem 15.0.6 *Let $(X, ||\cdot||)$ be a finite dimensional normed linear space and let $|\cdot|$ be described above relative to a given basis, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$. Then $|\cdot|$ is a norm and there exist constants $\delta, \Delta > 0$ independent of $\mathbf{x}$ such that*

$$\delta ||\mathbf{x}|| \leq |\mathbf{x}| \leq \Delta ||\mathbf{x}||. \tag{15.3}$$

Proof: All of the above properties of a norm are obvious except the second, the triangle inequality. To establish this inequality, use the Cauchy Schwartz inequality to write

$$\begin{aligned} |\mathbf{x}+\mathbf{y}|^2 &\equiv \sum_{i=1}^{n} |x_i + y_i|^2 \leq \sum_{i=1}^{n} |x_i|^2 + \sum_{i=1}^{n} |y_i|^2 + 2 \operatorname{Re} \sum_{i=1}^{n} x_i \overline{y}_i \\ &\leq |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2 \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} \left(\sum_{i=1}^{n} |y_i|^2 \right)^{1/2} \\ &= |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2 |\mathbf{x}| |\mathbf{y}| = (|\mathbf{x}| + |\mathbf{y}|)^2 \end{aligned}$$

and this proves the second property above.

It remains to show the equivalence of the two norms. By the Cauchy Schwartz inequality again,

$$\begin{aligned} ||\mathbf{x}|| &\equiv \left| \left| \sum_{i=1}^{n} x_i \mathbf{v}_i \right| \right| \leq \sum_{i=1}^{n} |x_i| \, ||\mathbf{v}_i|| \leq |\mathbf{x}| \left(\sum_{i=1}^{n} ||\mathbf{v}_i||^2 \right)^{1/2} \\ &\equiv \delta^{-1} |\mathbf{x}|. \end{aligned}$$

This proves the first half of the inequality.

Suppose the second half of the inequality is not valid. Then there exists a sequence $\mathbf{x}^k \in X$ such that

$$\left|\mathbf{x}^k\right| > k \left|\left|\mathbf{x}^k\right|\right|, \; k = 1, 2, \cdots.$$

Then define

$$\mathbf{y}^k \equiv \frac{\mathbf{x}^k}{|\mathbf{x}^k|}.$$

It follows

$$\left|\mathbf{y}^k\right| = 1, \quad \left|\mathbf{y}^k\right| > k \left|\left|\mathbf{y}^k\right|\right|. \tag{15.4}$$

Letting y_i^k be the components of $\mathbf{y}^k$ with respect to the given basis, it follows the vector

$$\left(y_1^k, \cdots, y_n^k\right)$$

is a unit vector in $\mathbb{F}^n$. By the Heine Borel theorem, there exists a subsequence, still denoted by k such that

$$\left(y_1^k, \cdots, y_n^k\right) \to \left(y_1, \cdots, y_n\right).$$

It follows from 15.4 and this that for

$$\mathbf{y} = \sum_{i=1}^n y_i \mathbf{v}_i,$$

$$0 = \lim_{k\to\infty} \left|\left|\mathbf{y}^k\right|\right| = \lim_{k\to\infty} \left|\left|\sum_{i=1}^n y_i^k \mathbf{v}_i\right|\right| = \left|\left|\sum_{i=1}^n y_i \mathbf{v}_i\right|\right|$$

but not all the y_i equal zero. This contradicts the assumption that $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis and proves the second half of the inequality.

Corollary 15.0.7 *If $(X, ||\cdot||)$ is a finite dimensional normed linear space with the field of scalars $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, then X is complete.*

Proof: Let $\{\mathbf{x}^k\}$ be a Cauchy sequence. Then letting the components of $\mathbf{x}^k$ with respect to the given basis be

$$x_1^k, \cdots, x_n^k,$$

it follows from Theorem 15.0.6, that

$$\left(x_1^k, \cdots, x_n^k\right)$$

is a Cauchy sequence in $\mathbb{F}^n$ and so

$$\left(x_1^k, \cdots, x_n^k\right) \to \left(x_1, \cdots, x_n\right) \in \mathbb{F}^n.$$

Thus,

$$\mathbf{x}^k = \sum_{i=1}^n x_i^k \mathbf{v}_i \to \sum_{i=1}^n x_i \mathbf{v}_i \in X.$$

This proves the corollary.

Corollary 15.0.8 *Suppose X is a finite dimensional linear space with the field of scalars either $\mathbb{C}$ or $\mathbb{R}$ and $||\cdot||$ and $|||\cdot|||$ are two norms on X. Then there exist positive constants, δ and Δ, independent of $\mathbf{x} \in X$ such that*

$$\delta \left|||\mathbf{x}|||\right| \leq ||\mathbf{x}|| \leq \Delta \left|||\mathbf{x}|||\right|.$$

Thus any two norms are equivalent.

Proof: Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis for X and let $|\cdot|$ be the norm taken with respect to this basis which was described earlier. Then by Theorem 15.0.6, there are positive constants $\delta_1, \Delta_1, \delta_2, \Delta_2$, all independent of $\mathbf{x} \in X$ such that

$$\delta_2 |||\mathbf{x}||| \leq |\mathbf{x}| \leq \Delta_2 |||\mathbf{x}|||,$$

$$\delta_1 ||\mathbf{x}|| \leq |\mathbf{x}| \leq \Delta_1 ||\mathbf{x}||.$$

Then

$$\delta_2 |||\mathbf{x}||| \leq |\mathbf{x}| \leq \Delta_1 ||\mathbf{x}|| \leq \frac{\Delta_1}{\delta_1} |\mathbf{x}| \leq \frac{\Delta_1 \Delta_2}{\delta_1} |||\mathbf{x}|||$$

and so

$$\frac{\delta_2}{\Delta_1} |||\mathbf{x}||| \leq ||\mathbf{x}|| \leq \frac{\Delta_2}{\delta_1} |||\mathbf{x}|||$$

which proves the corollary.

Definition 15.0.9 *Let X and Y be normed linear spaces with norms $||\cdot||_X$ and $||\cdot||_Y$ respectively. Then $\mathcal{L}(X,Y)$ denotes the space of linear transformations, called bounded linear transformations, mapping X to Y which have the property that*

$$||A|| \equiv \sup\{||Ax||_Y : ||x||_X \leq 1\} < \infty.$$

Then $||A||$ is referred to as the operator norm of the bounded linear transformation, A.

It is an easy exercise to verify that $||\cdot||$ is a norm on $\mathcal{L}(X,Y)$ and it is always the case that

$$||Ax||_Y \leq ||A|| \, ||x||_X.$$

Theorem 15.0.10 *Let X and Y be finite dimensional normed linear spaces of dimension n and m respectively and denote by $||\cdot||$ the norm on either X or Y. Then if A is any linear function mapping X to Y, then $A \in \mathcal{L}(X,Y)$ and $(\mathcal{L}(X,Y), ||\cdot||)$ is a complete normed linear space of dimension nm with*

$$||A\mathbf{x}|| \leq ||A|| \, ||\mathbf{x}||.$$

Proof: It is necessary to show the norm defined on linear transformations really is a norm. Again the first and third properties listed above for norms are obvious. It remains to show the second and verify $||A|| < \infty$. Letting $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis and $|\cdot|$ defined with respect to this basis as above, there exist constants $\delta, \Delta > 0$ such that

$$\delta ||\mathbf{x}|| \leq |\mathbf{x}| \leq \Delta ||\mathbf{x}||.$$

Then,

$$\begin{aligned} ||A+B|| &\equiv \sup\{||(A+B)(\mathbf{x})|| : ||\mathbf{x}|| \leq 1\} \\ &\leq \sup\{||A\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} + \sup\{||B\mathbf{x}|| : ||\mathbf{x}|| \leq 1\} \\ &\equiv ||A|| + ||B||. \end{aligned}$$

Next consider the claim that $||A|| < \infty$. This follows from

$$||A(\mathbf{x})|| = \left|\left| A\left(\sum_{i=1}^n x_i \mathbf{v}_i\right)\right|\right| \leq \sum_{i=1}^n |x_i| \, ||A(\mathbf{v}_i)||$$

$$\leq |\mathbf{x}| \left(\sum_{i=1}^{n} ||A(\mathbf{v}_i)||^2 \right)^{1/2} \leq \Delta ||\mathbf{x}|| \left(\sum_{i=1}^{n} ||A(\mathbf{v}_i)||^2 \right)^{1/2} < \infty.$$

Thus $||A|| \leq \Delta \left(\sum_{i=1}^{n} ||A(\mathbf{v}_i)||^2 \right)^{1/2}$.

Next consider the assertion about the dimension of $\mathcal{L}(X,Y)$. Let the two sets of bases be

$$\{\mathbf{v}_1, \cdots, \mathbf{v}_n\} \text{ and } \{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$$

for X and Y respectively. Let $\mathbf{w}_i \otimes \mathbf{v}_k \in \mathcal{L}(X,Y)$ be defined by

$$\mathbf{w}_i \otimes \mathbf{v}_k \mathbf{v}_l \equiv \begin{cases} \mathbf{0} \text{ if } l \neq k \\ \mathbf{w}_i \text{ if } l = k \end{cases}$$

and let $L \in \mathcal{L}(X,Y)$. Then

$$L\mathbf{v}_r = \sum_{j=1}^{m} d_{jr} \mathbf{w}_j$$

for some d_{jk}. Also

$$\sum_{j=1}^{m} \sum_{k=1}^{n} d_{jk} \mathbf{w}_j \otimes \mathbf{v}_k (\mathbf{v}_r) = \sum_{j=1}^{m} d_{jr} \mathbf{w}_j.$$

It follows that

$$L = \sum_{j=1}^{m} \sum_{k=1}^{n} d_{jk} \mathbf{w}_j \otimes \mathbf{v}_k$$

because the two linear transformations agree on a basis. Since L is arbitrary this shows

$$\{\mathbf{w}_i \otimes \mathbf{v}_k : i = 1, \cdots, m, \ k = 1, \cdots, n\}$$

spans $\mathcal{L}(X,Y)$. If

$$\sum_{i,k} d_{ik} \mathbf{w}_i \otimes \mathbf{v}_k = \mathbf{0},$$

then

$$\mathbf{0} = \sum_{i,k} d_{ik} \mathbf{w}_i \otimes \mathbf{v}_k (\mathbf{v}_l) = \sum_{i=1}^{m} d_{il} \mathbf{w}_i$$

and so, since $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ is a basis, $d_{il} = 0$ for each $i = 1, \cdots, m$. Since l is arbitrary, this shows $d_{il} = 0$ for all i and l. Thus these linear transformations form a basis and this shows the dimension of $\mathcal{L}(X,Y)$ is mn as claimed. By Corollary 15.0.7 $(\mathcal{L}(X,Y), ||\cdot||)$ is complete. If $\mathbf{x} \neq \mathbf{0}$,

$$||A\mathbf{x}|| \frac{1}{||\mathbf{x}||} = \left|\left| A \frac{\mathbf{x}}{||\mathbf{x}||} \right|\right| \leq ||A||$$

This proves the theorem.

Note by Corollary 15.0.8 you can define a norm any way desired on any finite dimensional linear space which has the field of scalars $\mathbb{R}$ or $\mathbb{C}$ and any other way of defining a norm on this space yields an equivalent norm. Thus, it doesn't much matter as far as notions of convergence are concerned which norm is used for a finite dimensional space. In particular in the space of $m \times n$ matrices, you can use the operator norm defined above, or some other way of giving this space a norm. A popular choice for a norm is the Frobenius norm defined below.

Definition 15.0.11 *Make the space of $m \times n$ matrices into a Hilbert space by defining*

$$(A, B) \equiv tr\left(AB^*\right).$$

Another way of describing a norm for an $n \times n$ matrix is as follows.

Definition 15.0.12 *Let A be an $m \times n$ matrix. Define the spectral norm of A, written as $||A||_2$ to be*

$$\max\left\{|\lambda|^{1/2} : \lambda \text{ is an eigenvalue of } A^*A\right\}.$$

Actually, this is nothing new. It turns out that $||\cdot||_2$ is nothing more than the operator norm for A taken with respect to the usual Euclidean norm,

$$|\mathbf{x}| = \left(\sum_{k=1}^{n} |x_k|^2\right)^{1/2}.$$

Proposition 15.0.13 *The following holds.*

$$||A||_2 = \sup\{|A\mathbf{x}| : |\mathbf{x}| = 1\} \equiv ||A||.$$

Proof: Note that A^*A is Hermitian and so by Corollary 14.2.5,

$$\begin{aligned} ||A||_2 &= \max\left\{(A^*A\mathbf{x}, \mathbf{x})^{1/2} : |\mathbf{x}| = 1\right\} \\ &= \max\left\{(A\mathbf{x}, A\mathbf{x})^{1/2} : |\mathbf{x}| = 1\right\} \\ &\leq ||A||. \end{aligned}$$

Now to go the other direction, let $|\mathbf{x}| \leq 1$. Then

$$|A\mathbf{x}| = \left|(A\mathbf{x}, A\mathbf{x})^{1/2}\right| = (A^*A\mathbf{x}, \mathbf{x})^{1/2} \leq ||A||_2,$$

and so, taking the sup over all $|\mathbf{x}| \leq 1$, it follows $||A|| \leq ||A||_2$.

An interesting application of the notion of equivalent norms on $\mathbb{R}^n$ is the process of giving a norm on a finite Cartesian product of normed linear spaces.

Definition 15.0.14 *Let X_i, $i = 1, \cdots, n$ be normed linear spaces with norms, $||\cdot||_i$. For*

$$\mathbf{x} \equiv (x_1, \cdots, x_n) \in \prod_{i=1}^{n} X_i$$

define $\theta : \prod_{i=1}^n X_i \to \mathbb{R}^n$ by

$$\theta(\mathbf{x}) \equiv (||x_1||_1, \cdots, ||x_n||_n)$$

Then if $||\cdot||$ is any norm on $\mathbb{R}^n$, define a norm on $\prod_{i=1}^n X_i$, also denoted by $||\cdot||$ by

$$||\mathbf{x}|| \equiv ||\theta\mathbf{x}||.$$

The following theorem follows immediately from Corollary 15.0.8.

Theorem 15.0.15 *Let X_i and $||\cdot||_i$ be given in the above definition and consider the norms on $\prod_{i=1}^n X_i$ described there in terms of norms on $\mathbb{R}^n$. Then any two of these norms on $\prod_{i=1}^n X_i$ obtained in this way are equivalent.*

For example, define

$$||\mathbf{x}||_1 \equiv \sum_{i=1}^{n} |x_i|,$$

$$||\mathbf{x}||_\infty \equiv \max\{|x_i|, i = 1, \cdots, n\},$$

or

$$||\mathbf{x}||_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$$

and all three are equivalent norms on $\prod_{i=1}^{n} X_i$.

In addition to $||\cdot||_1$ and $||\cdot||_\infty$ mentioned above, it is common to consider the so called p norms for $\mathbf{x} \in \mathbb{C}^n$.

Definition 15.0.16 *Let $\mathbf{x} \in \mathbb{C}^n$. Then define for $p \geq 1$,*

$$||\mathbf{x}||_p \equiv \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

The following inequality is called Holder's inequality.

Proposition 15.0.17 *For $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$,*

$$\sum_{i=1}^{n} |x_i|\,|y_i| \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^{p'}\right)^{1/p'}$$

The proof will depend on the following lemma.

Lemma 15.0.18 *If $a, b \geq 0$ and p' is defined by $\frac{1}{p} + \frac{1}{p'} = 1$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^{p'}}{p'}.$$

Proof of the Proposition: If $\mathbf{x}$ or $\mathbf{y}$ equals the zero vector there is nothing to prove. Therefore, assume they are both nonzero. Let $A = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$ and $B = \left(\sum_{i=1}^{n} |y_i|^{p'}\right)^{1/p'}$. Then using Lemma 15.0.18,

$$\begin{aligned} \sum_{i=1}^{n} \frac{|x_i|}{A}\frac{|y_i|}{B} &\leq \sum_{i=1}^{n} \left[\frac{1}{p}\left(\frac{|x_i|}{A}\right)^p + \frac{1}{p'}\left(\frac{|y_i|}{B}\right)^{p'}\right] \\ &= 1 \end{aligned}$$

and so

$$\sum_{i=1}^{n} |x_i|\,|y_i| \leq AB = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^{p'}\right)^{1/p'}.$$

This proves the proposition.

Theorem 15.0.19 *The p norms do indeed satisfy the axioms of a norm.*

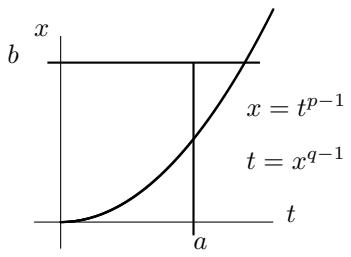
Proof: It is obvious that $||\cdot||_p$ does indeed satisfy most of the norm axioms. The only one that is not clear is the triangle inequality. To save notation write $||\cdot||$ in place of $||\cdot||_p$ in what follows. Note also that $\frac{p}{p'} = p-1$. Then using the Holder inequality,

$$\begin{aligned}
||\mathbf{x}+\mathbf{y}||^p &= \sum_{i=1}^n |x_i+y_i|^p \\
&\leq \sum_{i=1}^n |x_i+y_i|^{p-1}|x_i| + \sum_{i=1}^n |x_i+y_i|^{p-1}|y_i| \\
&= \sum_{i=1}^n |x_i+y_i|^{\frac{p}{p'}}|x_i| + \sum_{i=1}^n |x_i+y_i|^{\frac{p}{p'}}|y_i| \\
&\leq \left(\sum_{i=1}^n |x_i+y_i|^p\right)^{1/p'} \left[\left(\sum_{i=1}^n |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^n |y_i|^p\right)^{1/p}\right] \\
&= ||\mathbf{x}+\mathbf{y}||^{p/p'} \left(||\mathbf{x}||_p + ||\mathbf{y}||_p\right)
\end{aligned}$$

so $||\mathbf{x}+\mathbf{y}|| \leq ||\mathbf{x}||_p + ||\mathbf{y}||_p$. This proves the theorem.

It only remains to prove Lemma 15.0.18.

Proof of the lemma: Let $p' = q$ to save on notation and consider the following picture:



$$ab \leq \int_0^a t^{p-1}dt + \int_0^b x^{q-1}dx = \frac{a^p}{p} + \frac{b^q}{q}.$$

Note equality occurs when $a^p = b^q$.

Now $||A||_p$ may be considered as the operator norm of A taken with respect to $||\cdot||_p$. In the case when $p = 2$, this is just the spectral norm. There is an easy estimate for $||A||_p$ in terms of the entries of A.

Theorem 15.0.20 *The following holds.*

$$||A||_p \leq \left(\sum_k \left(\sum_j |A_{jk}|^p\right)^{q/p}\right)^{1/q}$$

Proof: Let $||\mathbf{x}||_p \leq 1$ and let $A = (\mathbf{a}_1, \cdots, \mathbf{a}_n)$ where the $\mathbf{a}_k$ are the columns of A. Then

$$A\mathbf{x} = \left(\sum_k x_k \mathbf{a}_k\right)$$

and so by Holder's inequality,

$$\begin{aligned} ||A\mathbf{x}||_p &\equiv \left|\left|\sum_k x_k \mathbf{a}_k\right|\right|_p \leq \sum_k |x_k| \, ||\mathbf{a}_k||_p \\ &\leq \left(\sum_k |x_k|^p\right)^{1/p} \left(\sum_k ||\mathbf{a}_k||_p^q\right)^{1/q} \\ &\leq \left(\sum_k \left(\sum_j |A_{jk}|^p\right)^{q/p}\right)^{1/q} \end{aligned}$$

and this shows $||A||_p \leq \left(\sum_k \left(\sum_j |A_{jk}|^p\right)^{q/p}\right)^{1/q}$ and proves the theorem.

15.1 The Condition Number

Let $A \in \mathcal{L}(X,X)$ be a linear transformation where X is a finite dimensional vector space and consider the problem $Ax = b$ where it is assumed there is a unique solution to this problem. How does the solution change if A is changed a little bit and if b is changed a little bit? This is clearly an interesting question because you often do not know A and b exactly. If a small change in these quantities results in a large change in the solution, x, then it seems clear this would be undesirable. In what follows $||\cdot||$ when applied to a linear transformation will always refer to the operator norm.

Lemma 15.1.1 *Let $A, B \in \mathcal{L}(X,X), A^{-1} \in \mathcal{L}(X,X)$, and suppose $||B|| < ||A||$. Then $(A+B)^{-1}$ exists and*

$$\left|\left|(A+B)^{-1}\right|\right| \leq ||A^{-1}|| \left|\left|\frac{1}{1-||A^{-1}B||}\right|\right|.$$

The above formula makes sense because $\left|\left|A^{-1}B\right|\right| < 1$. Also, if L is any invertible linear transformation,

$$||L|| \leq \frac{1}{||L^{-1}||}. \tag{15.5}$$

Proof: For $||x|| \leq 1, x \neq 0$, and L an invertible linear transformation, $x = L^{-1}Lx$ and so $||x|| \leq \left|\left|L^{-1}\right|\right| ||Lx||$ which implies

$$||Lx|| \leq \frac{1}{||L^{-1}||} ||x||.$$

Therefore,

$$||L|| \leq \frac{1}{||L^{-1}||}. \tag{15.6}$$

Similarly

$$\left|\left|L^{-1}\right|\right| \leq \frac{1}{||L||}. \tag{15.7}$$

This establishes 15.5.

Suppose $(A+B)x = 0$. Then $0 = A\left(I + A^{-1}B\right)x$ and so since A is one to one, $\left(I + A^{-1}B\right)x = 0$. Therefore,

$$0 = \left|\left|\left(I + A^{-1}B\right)x\right|\right| \geq ||x|| - \left|\left|A^{-1}Bx\right|\right|$$

and so from 15.7

$$\begin{aligned} ||x|| &\leq \left|\left|A^{-1}Bx\right|\right| \leq \left|\left|A^{-1}\right|\right| ||B|| \, ||x|| \quad (15.8) \\ &\leq \frac{||B||}{||A||} ||x|| \quad (15.9) \end{aligned}$$

which is a contradiction unless $||x|| = 0$. Therefore, $(A+B)^{-1}$ exists.

Now

$$(A+B)^{-1} = \left(A\left(I + A^{-1}B\right)\right)^{-1}.$$

$$\left|\left|A^{-1}B\right|\right| \leq \left|\left|A^{-1}\right|\right| ||B|| \leq \frac{||B||}{||A||} < 1 = ||I||.$$

Letting $A^{-1}B$ play the role of B in the above and I play the role of A, it follows $\left(I + A^{-1}B\right)^{-1}$ exists. Hence

$$(A+B)^{-1} = \left(I + A^{-1}B\right)^{-1} A^{-1}$$

and so by 15.7 applied to $I + A^{-1}B$,

$$\begin{aligned} \left|\left|(A+B)^{-1}\right|\right| &\leq \left|\left|A^{-1}\right|\right| \left|\left|\left(I + A^{-1}B\right)^{-1}\right|\right| \\ &\leq \left|\left|A^{-1}\right|\right| \frac{1}{||I + A^{-1}B||} \\ &\leq \left|\left|A^{-1}\right|\right| \frac{1}{(1 - ||A^{-1}B||)}. \end{aligned}$$

This proves the lemma.

Proposition 15.1.2 *Suppose A is invertible, $b \neq 0$, $Ax = b$, and $A_1 x_1 = b_1$ where $||A - A_1|| < ||A||$. Then*

$$\frac{||x_1 - x||}{||x||} \leq \frac{1}{(1 - ||A^{-1}(A_1 - A)||)} ||A|| \left|\left|A^{-1}\right|\right| \left(\frac{||A_1 - A||}{||A||} + \frac{||b - b_1||}{||b||}\right). \quad (15.10)$$

Proof: It follows from the assumptions that

$$Ax - A_1 x + A_1 x - A_1 x_1 = b - b_1.$$

Hence

$$A_1(x - x_1) = (A_1 - A)x + b - b_1.$$

Now $A_1 = (A + (A_1 - A))$ and so by the above lemma, A_1^{-1} exists and so

$$(x - x_1) = A_1^{-1}(A_1 - A)x + A_1^{-1}(b - b_1).$$

By the estimate in the lemma,

$$||x - x_1|| \leq \frac{\left|\left|A^{-1}\right|\right|}{1 - ||A^{-1}(A_1 - A)||} (||A_1 - A|| \, ||x|| + ||b - b_1||).$$

Dividing by $||x||$,

$$\frac{||x-x_1||}{||x||} \leq \frac{\left|\left|A^{-1}\right|\right|}{1-||A^{-1}(A_1-A)||}\left(||A_1-A|| + \frac{||b-b_1||}{||x||}\right) \tag{15.11}$$

Now $b = A\left(A^{-1}b\right)$ and so $||b|| \leq ||A||\left|\left|A^{-1}b\right|\right|$ and so

$$||x|| = \left|\left|A^{-1}b\right|\right| \geq ||b||/||A||.$$

Therefore, from 15.11,

$$\begin{aligned}\frac{||x-x_1||}{||x||} &\leq \frac{\left|\left|A^{-1}\right|\right|}{1-||A^{-1}(A_1-A)||}\left(\frac{||A||\,||A_1-A||}{||A||} + \frac{||A||\,||b-b_1||}{||b||}\right)\\ &\leq \frac{\left|\left|A^{-1}\right|\right|||A||}{1-||A^{-1}(A_1-A)||}\left(\frac{||A_1-A||}{||A||} + \frac{||b-b_1||}{||b||}\right)\end{aligned}$$

which proves the proposition.

This shows that the number, $\left|\left|A^{-1}\right|\right|||A||$, controls how sensitive the relative change in the solution of $Ax = b$ is to small changes in A and b. This number is called the condition number. It is a bad thing when it is large because a small relative change in b, for example could yield a large relative change in x.

15.2 The Spectral Radius

Even though it is in general impractical to compute the Jordan form, its existence is all that is needed in order to prove an important theorem about something which is relatively easy to compute. This is the spectral radius of a matrix.

Definition 15.2.1 *Define $\sigma(A)$ to be the eigenvalues of A. Also,*

$$\rho(A) \equiv \max(|\lambda| : \lambda \in \sigma(A))$$

The number, $\rho(A)$ is known as the spectral radius of A.

Before beginning this discussion, it is necessary to define what is meant by convergence in $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$.

Definition 15.2.2 *Let $\{A_k\}_{k=1}^{\infty}$ be a sequence in $\mathcal{L}(X,Y)$ where X, Y are finite dimensional normed linear spaces. Then $\lim_{n\to\infty} A_k = A$ if for every $\varepsilon > 0$ there exists N such that if $n > N$, then*

$$||A - A_n|| < \varepsilon.$$

Here the norm refers to any of the norms defined on $\mathcal{L}(X,Y)$. By Corollary 15.0.8 and Theorem 11.2.2 it doesn't matter which one is used. Define the symbol for an infinite sum in the usual way. Thus

$$\sum_{k=1}^{\infty} A_k \equiv \lim_{n\to\infty} \sum_{k=1}^{n} A_k$$

Lemma 15.2.3 *Suppose $\{A_k\}_{k=1}^{\infty}$ is a sequence in $\mathcal{L}(X,Y)$ where X, Y are finite dimensional normed linear spaces. Then if*

$$\sum_{k=1}^{\infty} ||A_k|| < \infty,$$

It follows that

$$\sum_{k=1}^{\infty} A_k \tag{15.12}$$

exists. In words, absolute convergence implies convergence.

Proof: For $p \leq m \leq n$,

$$\left|\left|\sum_{k=1}^{n} A_k - \sum_{k=1}^{m} A_k\right|\right| \leq \sum_{k=p}^{\infty} ||A_k||$$

and so for p large enough, this term on the right in the above inequality is less than ε. Since ε is arbitrary, this shows the partial sums of 15.12 are a Cauchy sequence. Therefore by Corollary 15.0.7 it follows that these partial sums converge.

The next lemma is normally discussed in advanced calculus courses but is proved here for the convenience of the reader. It is known as the root test.

Lemma 15.2.4 *Let $\{a_p\}$ be a sequence of nonnegative terms and let*

$$r = \limsup_{p\to\infty} a_p^{1/p}.$$

Then if $r < 1$, it follows the series, $\sum_{k=1}^{\infty} a_k$ converges and if $r > 1$, then a_p fails to converge to 0 so the series diverges. If A is an $n \times n$ matrix and

$$1 < \limsup_{p\to\infty} ||A^p||^{1/p}, \tag{15.13}$$

then $\sum_{k=0}^{\infty} A^k$ fails to converge.

Proof: Suppose $r < 1$. Then there exists N such that if $p > N$,

$$a_p^{1/p} < R$$

where $r < R < 1$. Therefore, for all such p, $a_p < R^p$ and so by comparison with the geometric series, $\sum R^p$, it follows $\sum_{p=1}^{\infty} a_p$ converges.

Next suppose $r > 1$. Then letting $1 < R < r$, it follows there are infinitely many values of p at which

$$R < a_p^{1/p}$$

which implies $R^p < a_p$, showing that a_p cannot converge to 0.

To see the last claim, if 15.13 holds, then from the first part of this lemma, $||A^p||$ fails to converge to 0 and so $\left\{\sum_{k=0}^{m} A^k\right\}_{m=0}^{\infty}$ is not a Cauchy sequence. Hence $\sum_{k=0}^{\infty} A^k \equiv \lim_{m\to\infty} \sum_{k=0}^{m} A^k$ cannot exist.

In this section a significant way to estimate $\rho(A)$ is presented. It is based on the following lemma.

Lemma 15.2.5 *If $|\lambda| > \rho(A)$, for A an $n \times n$ matrix, then the series,*

$$\frac{1}{\lambda} \sum_{k=0}^{\infty} \frac{A^k}{\lambda^k}$$

converges.

Proof: Let J denote the Jordan canonical form of A. Also, let $||A|| \equiv \max\{|a_{ij}|, i,j = 1,2,\cdots,n\}$. Then for some invertible matrix, S, $A = S^{-1}JS$. Therefore,

$$\frac{1}{\lambda}\sum_{k=0}^{p}\frac{A^k}{\lambda^k} = S^{-1}\left(\frac{1}{\lambda}\sum_{k=0}^{p}\frac{J^k}{\lambda^k}\right)S.$$

Now from the structure of the Jordan form, $J = D + N$ where D is the diagonal matrix consisting of the eigenvalues of A listed according to algebraic multiplicity and N is a nilpotent matrix which commutes with D. Say $N^m = 0$. Therefore, for k much larger than m, say $k > 2m$,

$$J^k = (D+N)^k = \sum_{l=0}^{m}\binom{k}{l}D^{k-l}N^l.$$

It follows that

$$\left|\left|J^k\right|\right| \leq C(m,N)\,k(k-1)\cdots(k-m+1)\,||D||^k$$

and so

$$\limsup_{k\to\infty}\left|\left|\frac{J^k}{\lambda^k}\right|\right|^{1/k} \leq \lim_{k\to\infty}\left(\frac{C(m,N)\,k(k-1)\cdots(k-m+1)\,||D||^k}{|\lambda|^k}\right)^{1/k} = \frac{||D||}{|\lambda|} < 1.$$

Therefore, this shows by the root test that $\sum_{k=0}^{\infty}\left|\left|\frac{J^k}{\lambda^k}\right|\right|$ converges. Therefore, by Lemma 15.2.3 it follows that

$$\lim_{k\to\infty}\frac{1}{\lambda}\sum_{l=0}^{k}\frac{J^l}{\lambda^l}$$

exists. In particular this limit exists in every norm placed on $\mathcal{L}(\mathbb{F}^n,\mathbb{F}^n)$, and in particular for every operator norm. Now for any operator norm, $||AB|| \leq ||A||\,||B||$. Therefore,

$$\left|\left|S^{-1}\left(\frac{1}{\lambda}\sum_{k=0}^{p}\frac{J^k}{\lambda^k} - \frac{1}{\lambda}\sum_{k=0}^{\infty}\frac{J^k}{\lambda^k}\right)S\right|\right| \leq \left|\left|S^{-1}\right|\right|\,||S||\left|\left|\left(\frac{1}{\lambda}\sum_{k=0}^{p}\frac{J^k}{\lambda^k} - \frac{1}{\lambda}\sum_{k=0}^{\infty}\frac{J^k}{\lambda^k}\right)\right|\right|$$

and this converges to 0 as $p \to \infty$. Therefore,

$$\frac{1}{\lambda}\sum_{k=0}^{p}\frac{A^k}{\lambda^k} \to S^{-1}\left(\frac{1}{\lambda}\sum_{k=0}^{\infty}\frac{J^k}{\lambda^k}\right)S$$

and this proves the lemma.

Actually this lemma is usually accomplished using the theory of functions of a complex variable but the theory involving the Laurent series is not assumed here. In infinite dimensional spaces you have to use complex variable techniques however.

Lemma 15.2.6 *Let A be an $n\times n$ matrix. Then for any $||\cdot||$, $\rho(A) \geq \limsup_{p\to\infty}||A^p||^{1/p}$.*

Proof: By Lemma 15.2.5 and Lemma 15.2.4, if $|\lambda| > \rho(A)$,

$$\limsup\left|\left|\frac{A^k}{\lambda^k}\right|\right|^{1/k} \leq 1,$$

and it doesn't matter which norm is used because they are all equivalent. Therefore, $\limsup_{k\to\infty}\left|\left|A^k\right|\right|^{1/k} \leq |\lambda|$. Therefore, since this holds for all $|\lambda| > \rho(A)$, this proves the lemma.

Now denote by $\sigma(A)^p$ the collection of all numbers of the form λ^p where $\lambda \in \sigma(A)$.

Lemma 15.2.7 $\sigma(A^p) = \sigma(A)^p$

Proof: In dealing with $\sigma(A^p)$, is suffices to deal with $\sigma(J^p)$ where J is the Jordan form of A because J^p and A^p are similar. Thus if $\lambda \in \sigma(A^p)$, then $\lambda \in \sigma(J^p)$ and so $\lambda = \alpha$ where α is one of the entries on the main diagonal of J^p. Thus $\lambda \in \alpha(A)^p$ and this shows $\sigma(A^p) \subseteq \sigma(A)^p$.

Now take $\alpha \in \sigma(A)$ and consider α^p.

$$\alpha^p I - A^p = \left(\alpha^{p-1} I + \cdots + \alpha A^{p-2} + A^{p-1}\right)(\alpha I - A)$$

and so $\alpha^p I - A^p$ fails to be one to one which shows that $\alpha^p \in \sigma(A^p)$ which shows that $\sigma(A)^p \subseteq \sigma(A^p)$. This proves the lemma.

Lemma 15.2.8 *Let A be an $n \times n$ matrix and suppose $|\lambda| > ||A||_2$. Then $(\lambda I - A)^{-1}$ exists.*

Proof: Suppose $(\lambda I - A)\mathbf{x} = \mathbf{0}$ where $\mathbf{x} \neq \mathbf{0}$. Then

$$|\lambda| \, ||\mathbf{x}||_2 = ||A\mathbf{x}||_2 \leq ||A|| \, ||\mathbf{x}||_2 < |\lambda| \, ||\mathbf{x}||_2,$$

a contradiction. Therefore, $(\lambda I - A)$ is one to one and this proves the lemma.

The main result is the following theorem due to Gelfand in 1941.

Theorem 15.2.9 *Let A be an $n \times n$ matrix. Then for any $||\cdot||$ defined on $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$*

$$\rho(A) = \lim_{p \to \infty} ||A^p||^{1/p}.$$

Proof: If $\lambda \in \sigma(A)$, then by Lemma 15.2.7 $\lambda^p \in \sigma(A^p)$ and so by Lemma 15.2.8, it follows that

$$|\lambda|^p \leq ||A^p||$$

and so $|\lambda| \leq ||A^p||^{1/p}$. Since this holds for every $\lambda \in \sigma(A)$, it follows that for each p,

$$\rho(A) \leq ||A^p||^{1/p}.$$

Now using Lemma 15.2.6,

$$\rho(A) \geq \limsup_{p \to \infty} ||A^p||^{1/p} \geq \liminf_{p \to \infty} ||A^p||^{1/p} \geq \rho(A)$$

which proves the theorem.

Example 15.2.10 *Consider* $\begin{pmatrix} 9 & -1 & 2 \\ -2 & 8 & 4 \\ 1 & 1 & 8 \end{pmatrix}$. *Estimate the absolute value of the largest eigenvalue.*

A laborious computation reveals the eigenvalues are 5, and 10. Therefore, the right answer in this case is 10. Consider $\left|\left|A^7\right|\right|^{1/7}$ where the norm is obtained by taking the maximum of all the absolute values of the entries. Thus

$$\begin{pmatrix} 9 & -1 & 2 \\ -2 & 8 & 4 \\ 1 & 1 & 8 \end{pmatrix}^7 = \begin{pmatrix} 8\,015\,625 & -1\,984\,375 & 3\,968\,750 \\ -3\,968\,750 & 6\,031\,250 & 7\,937\,500 \\ 1\,984\,375 & 1\,984\,375 & 6\,031\,250 \end{pmatrix}$$

and taking the seventh root of the largest entry gives

$$\rho(A) \approx 8\,015\,625^{1/7} = 9.\,688\,951\,236\,71.$$

Of course the interest lies primarily in matrices for which the exact roots to the characteristic equation are not known.

15.3 Iterative Methods For Linear Systems

Consider the problem of solving the equation

$$A\mathbf{x} = \mathbf{b} \tag{15.14}$$

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 [10].

Definition 15.3.1 *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 15.14. The vectors in this list are called iterates and they are obtained according to the following procedure. Letting* $A = (a_{ij})$,

$$a_{ii}x_i^{r+1} = -\sum_{j \neq i} a_{ij}x_j^r + b_i. \tag{15.15}$$

In terms of matrices, letting

$$A = \begin{pmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix}$$

The iterates are defined as

$$\begin{aligned} &\begin{pmatrix} * & 0 & \cdots & 0 \\ 0 & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & * \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ \vdots \\ x_n^{r+1} \end{pmatrix} \\ = \quad &- \begin{pmatrix} 0 & * & \cdots & * \\ * & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ \vdots \\ x_n^r \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \end{aligned} \tag{15.16}$$

The matrix on the left in 15.16 is obtained by retaining the main diagonal of A and setting every other entry equal to zero. The matrix on the right in 15.16 is obtained from A by setting every diagonal entry equal to zero and retaining all the other entries unchanged.

Example 15.3.2 *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

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Multiplying by the invese of the matrix on the left, [1]this iteration reduces to

$$\begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix}. \tag{15.17}$$

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

$$\mathbf{x}^3 = - \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix}}^{\mathbf{x}_2} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .166 \\ .26 \\ .2 \\ .7 \end{pmatrix}$$

[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.

$$\mathbf{x}^4 = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} .166 \\ .26 \\ .2 \\ .7 \end{pmatrix}}^{\mathbf{x}_3} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .24 \\ .408\,5 \\ .356 \\ .9 \end{pmatrix}$$

$$\mathbf{x}^5 = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} .24 \\ .408\,5 \\ .356 \\ .9 \end{pmatrix}}^{\mathbf{x}_4} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .197 \\ .351 \\ .256\,6 \\ .822 \end{pmatrix}$$

$$\mathbf{x}^6 = -\begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & \frac{2}{5} & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \overbrace{\begin{pmatrix} .197 \\ .351 \\ .256\,6 \\ .822 \end{pmatrix}}^{\mathbf{x}^5} + \begin{pmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{3}{5} \\ 1 \end{pmatrix} = \begin{pmatrix} .216 \\ .386 \\ .295 \\ .871 \end{pmatrix}.$$

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 15.15 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.

Definition 15.3.3 *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*

$$\sum_{j=1}^{i} a_{ij} x_j^{r+1} = -\sum_{j=i+1}^{n} a_{ij} x_j^r + b_i. \tag{15.18}$$

In terms of matrices, letting

$$A = \begin{pmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{pmatrix}$$

The iterates are defined as

$$\begin{aligned} &\begin{pmatrix} * & 0 & \cdots & 0 \\ * & * & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & * \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ \vdots \\ x_n^{r+1} \end{pmatrix} \\ = \quad &-\begin{pmatrix} 0 & * & \cdots & * \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ \vdots \\ x_n^r \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \end{aligned} \tag{15.19}$$

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 15.18 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 15.3.4 *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 15.3.5 *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

$$A\mathbf{x} = \mathbf{b} \tag{15.20}$$

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

$$\mathbf{x}^{r+1} = -B^{-1}C\mathbf{x}^r + B^{-1}\mathbf{b}. \tag{15.21}$$

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. Recall the definition of the spectral radius of $M, \rho(M)$, in Definition 15.2.1 on Page 287.

Theorem 15.3.6 *Suppose $\rho\left(B^{-1}C\right) < 1$. Then the iterates in 15.21 converge to the unique solution of 15.20.*

I will prove this theorem in the next section. The proof depends on analysis which should not be surprising because it involves a statement about convergence of sequences.

15.4 Theory Of Convergence

Definition 15.4.1 *A normed vector space, E with norm $||\cdot||$ is called a Banach space if it is also complete. This means that every Cauchy sequence converges. Recall that a sequence $\{x_n\}_{n=1}^{\infty}$ is a Cauchy sequence if for every $\varepsilon > 0$ there exists N such that whenever $m, n > N$,*

$$||x_n - x_m|| < \varepsilon.$$

Thus whenever $\{x_n\}$ is a Cauchy sequence, there exists x such that

$$\lim_{n\to\infty} ||x - x_n|| = 0.$$

Example 15.4.2 *Let Ω be a nonempty subset of a normed linear space, F. Denote by $BC(\Omega; E)$ the set of bounded continuous functions having values in E where E is a Banach space. Then define the norm on $BC(\Omega; E)$ by*

$$||f|| \equiv \sup\{||f(x)||_E : x \in \Omega\}.$$

Lemma 15.4.3 *The space $BC(\Omega; E)$ with the given norm is a Banach space.*

Proof: It is obvious $||\cdot||$ is a norm. It only remains to verify $BC(\Omega; E)$ is complete. Let $\{f_n\}$ be a Cauchy sequence. Then pick $x \in \Omega$.

$$||f_n(x) - f_m(x)||_E \leq ||f_n - f_m|| < \varepsilon$$

whenever m, n are large enough. Thus, for each $x, \{f_n(x)\}$ is a Cauchy sequence in E. Since E is complete, it follows there exists a function, f defined on Ω such that $f(x) = \lim_{n\to\infty} f_n(x)$.

It remains to verify that $f \in BC(\Omega; E)$ and that $||f - f_n|| \to 0$. I will first show that

$$\lim_{n\to\infty} \left(\sup_{x\in\Omega} \{||f(x) - f_n(x)||_E\} \right) = 0. \tag{15.22}$$

From this it will follow that f is bounded. Then I will show that f is continuous and $||f - f_n|| \to 0$. Let $\varepsilon > 0$ be given and let N be such that for $m, n > N$

$$||f_n - f_m|| < \varepsilon/3.$$

Then it follows that for all x,

$$||f(x) - f_m(x)||_E = \lim_{n\to\infty} ||f_n(x) - f_m(x)||_E \leq \varepsilon/3$$

Therefore, for $m > N$,

$$\sup_{x\in\Omega} \{||f(x) - f_m(x)||_E\} \leq \frac{\varepsilon}{3} < \varepsilon.$$

This proves 15.22. Then by the triangle inequality and letting N be as just described, pick $m > N$. Then for any $x \in \Omega$

$$||f(x)||_E \leq ||f_m(x)||_E + \varepsilon \leq ||f_m|| + \varepsilon.$$

Hence f is bounded. Now pick $x \in \Omega$ and let $\varepsilon > 0$ be given and N be as above. Then

$$\begin{aligned} ||f(x) - f(y)||_E &\leq ||f(x) - f_m(x)||_E + ||f_m(x) - f_m(y)||_E + ||f_m(y) - f(y)||_E \\ &\leq \frac{\varepsilon}{3} + ||f_m(x) - f_m(y)||_E + \frac{\varepsilon}{3}. \end{aligned}$$

Now by continuity of f_m, the middle term is less than $\varepsilon/3$ whenever $||x-y||$ is sufficiently small. Therefore, f is also continuous. Finally, from the above,

$$||f - f_n|| \leq \frac{\varepsilon}{3}$$

whenever $n > N$ and so $\lim_{n\to\infty} ||f - f_n|| = 0$ as claimed. This proves the lemma.

The most familiar example of a Banach space is $\mathbb{F}^n$. The following lemma is of great importance so it is stated in general.

Lemma 15.4.4 *Suppose $T : E \to E$ where E is a Banach space with norm $|\cdot|$. Also suppose*

$$|T\mathbf{x} - T\mathbf{y}| \leq r\,|\mathbf{x} - \mathbf{y}| \tag{15.23}$$

for some $r \in (0,1)$. Then there exists a unique fixed point, $\mathbf{x} \in E$ such that

$$T\mathbf{x} = \mathbf{x}. \tag{15.24}$$

Letting $\mathbf{x}^1 \in E$, this fixed point, $\mathbf{x}$, is the limit of the sequence of iterates,

$$\mathbf{x}^1, T\mathbf{x}^1, T^2\mathbf{x}^1, \cdots. \tag{15.25}$$

In addition to this, there is a nice estimate which tells how close $\mathbf{x}^1$ is to $\mathbf{x}$ in terms of things which can be computed.

$$\left|\mathbf{x}^1 - \mathbf{x}\right| \leq \frac{1}{1-r}\left|\mathbf{x}^1 - T\mathbf{x}^1\right|. \tag{15.26}$$

Proof: This follows easily when it is shown that the above sequence, $\left\{T^k\mathbf{x}^1\right\}_{k=1}^{\infty}$ is a Cauchy sequence. Note that

$$\left|T^2\mathbf{x}^1 - T\mathbf{x}^1\right| \leq r\left|T\mathbf{x}^1 - \mathbf{x}^1\right|.$$

Suppose

$$\left|T^k\mathbf{x}^1 - T^{k-1}\mathbf{x}^1\right| \leq r^{k-1}\left|T\mathbf{x}^1 - \mathbf{x}^1\right|. \tag{15.27}$$

Then

$$\begin{aligned}\left|T^{k+1}\mathbf{x}^1 - T^k\mathbf{x}^1\right| &\leq r\left|T^k\mathbf{x}^1 - T^{k-1}\mathbf{x}^1\right| \\ &\leq rr^{k-1}\left|T\mathbf{x}^1 - \mathbf{x}^1\right| = r^k\left|T\mathbf{x}^1 - \mathbf{x}^1\right|.\end{aligned}$$

By induction, this shows that for all $k \geq 2$, 15.27 is valid. Now let $k > l \geq N$.

$$\begin{aligned}\left|T^k\mathbf{x}^1 - T^l\mathbf{x}^1\right| &= \left|\sum_{j=l}^{k-1}\left(T^{j+1}\mathbf{x}^1 - T^j\mathbf{x}^1\right)\right| \\ &\leq \sum_{j=l}^{k-1}\left|T^{j+1}\mathbf{x}^1 - T^j\mathbf{x}^1\right| \\ &\leq \sum_{j=N}^{k-1} r^j\left|T\mathbf{x}^1 - \mathbf{x}^1\right| \leq \left|T\mathbf{x}^1 - \mathbf{x}^1\right|\frac{r^N}{1-r}\end{aligned}$$

which converges to 0 as $N \to \infty$. Therefore, this is a Cauchy sequence so it must converge to $\mathbf{x} \in E$. Then

$$\mathbf{x} = \lim_{k\to\infty} T^k\mathbf{x}^1 = \lim_{k\to\infty} T^{k+1}\mathbf{x}^1 = T\lim_{k\to\infty} T^k\mathbf{x}^1 = T\mathbf{x}.$$

This shows the existence of the fixed point. To show it is unique, suppose there were another one, $\mathbf{y}$. Then

$$|\mathbf{x}-\mathbf{y}| = |T\mathbf{x}-T\mathbf{y}| \leq r\,|\mathbf{x}-\mathbf{y}|$$

and so $\mathbf{x}=\mathbf{y}$.

It remains to verify the estimate.

$$\begin{aligned}\left|\mathbf{x}^1-\mathbf{x}\right| &\leq \left|\mathbf{x}^1-T\mathbf{x}^1\right|+\left|T\mathbf{x}^1-\mathbf{x}\right| \\ &= \left|\mathbf{x}^1-T\mathbf{x}^1\right|+\left|T\mathbf{x}^1-T\mathbf{x}\right| \\ &\leq \left|\mathbf{x}^1-T\mathbf{x}^1\right|+r\left|\mathbf{x}^1-\mathbf{x}\right|\end{aligned}$$

and solving the inequality for $\left|\mathbf{x}^1-\mathbf{x}\right|$ gives the estimate desired. This proves the lemma.

The following corollary is what will be used to prove the convergence condition for the various iterative procedures.

Corollary 15.4.5 *Suppose $T: E \to E$, for some constant C*

$$|T\mathbf{x}-T\mathbf{y}| \leq C\,|\mathbf{x}-\mathbf{y}|,$$

for all $\mathbf{x},\mathbf{y} \in E$, and for some $N \in \mathbb{N}$,

$$\left|T^N\mathbf{x}-T^N\mathbf{y}\right| \leq r\,|\mathbf{x}-\mathbf{y}|,$$

for all $\mathbf{x},\mathbf{y} \in E$ where $r \in (0,1)$. Then there exists a unique fixed point for T and it is still the limit of the sequence, $\left\{T^k\mathbf{x}^1\right\}$ for any choice of $\mathbf{x}^1$.

Proof: From Lemma 15.4.4 there exists a unique fixed point for T^N denoted here as $\mathbf{x}$. Therefore, $T^N\mathbf{x}=\mathbf{x}$. Now doing T to both sides,

$$T^NT\mathbf{x} = T\mathbf{x}.$$

By uniqueness, $T\mathbf{x}=\mathbf{x}$ because the above equation shows $T\mathbf{x}$ is a fixed point of T^N and there is only one.

It remains to consider the convergence of the sequence. Without loss of generality, it can be assumed $C \geq 1$. Then if $r \leq N-1$,

$$|T^r\mathbf{x}-T^r\mathbf{x}| \leq C^N\,|\mathbf{x}-\mathbf{y}| \tag{15.28}$$

for all $\mathbf{x},\mathbf{y} \in E$. By Lemma 15.4.4 there exists K such that if $k,l \geq K$, then

$$\left|T^{kN}\mathbf{x}^1-T^{lN}\mathbf{x}^1\right| < \eta = \frac{\varepsilon}{2C^N} \tag{15.29}$$

and also K is large enough that

$$2r^KC^N\frac{\left|T^N\mathbf{x}^1-\mathbf{x}^1\right|}{1-r} < \frac{\varepsilon}{2} \tag{15.30}$$

Now let $p,q > KN$ and define k_p, k_q, r_p, and r_q by

$$p = k_pN+r_p,\ q = k_qN+r_q,\ 0 \leq r_q, r_p < N.$$

Then both k_p and k_q are larger than K. Therefore, from 15.28 and 15.30,

$$\begin{aligned}
\left|T^p\mathbf{x}^1 - T^q\mathbf{x}^1\right| &= \left|T^{r_p}T^{k_pN}\mathbf{x}^1 - T^{r_q}T^{k_qN}\mathbf{x}^1\right| \\
&\leq \left|T^{k_pN}T^{r_p}\mathbf{x}^1 - T^{k_pN}T^{r_q}\mathbf{x}^1\right| + \left|T^{r_q}T^{k_pN}\mathbf{x}^1 - T^{r_q}T^{k_qN}\mathbf{x}^1\right| \\
&\leq r^{k_p}\left|T^{r_p}\mathbf{x}^1 - T^{r_q}\mathbf{x}^1\right| + C^N\left|T^{k_pN}\mathbf{x}^1 - T^{k_qN}\mathbf{x}^1\right| \\
&\leq r^K\left(\left|T^{r_p}\mathbf{x}^1 - \overbrace{T^{r_p}\mathbf{x}}^{=\mathbf{x}}\right| + \left|\overbrace{T^{r_q}\mathbf{x}}^{=\mathbf{x}} - T^{r_q}\mathbf{x}^1\right|\right) + C^N\eta \\
&\leq r^K\left(C^N\left|\mathbf{x}^1 - \mathbf{x}\right| + C^N\left|\mathbf{x}^1 - \mathbf{x}\right|\right) + C^N\eta \\
&\leq 2r^KC^N\frac{\left|T^N\mathbf{x}^1 - \mathbf{x}^1\right|}{1-r} + C^N\eta < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

This shows $\left\{T^k\mathbf{x}^1\right\}$ is a Cauchy sequence and since a subsequence converges to $\mathbf{x}$, it follows this sequence also must converge to $\mathbf{x}$. Here is why. Let $\varepsilon > 0$ be given. There exists M such that if $k, l > M$, then

$$\left|T^k\mathbf{x}^1 - T^l\mathbf{x}^1\right| < \frac{\varepsilon}{2}.$$

Now let $k > M$. Then let $l > M$ and also be large enough that

$$\left|T^{lN}\mathbf{x}^1 - \mathbf{x}\right| < \frac{\varepsilon}{2}.$$

Then

$$\begin{aligned}
\left|T^k\mathbf{x}^1 - \mathbf{x}\right| &\leq \left|T^k\mathbf{x}^1 - T^{lN}\mathbf{x}^1\right| + \left|T^{lN}\mathbf{x}^1 - \mathbf{x}\right| \\
&< \left|T^k\mathbf{x}^1 - T^{lN}\mathbf{x}^1\right| + \frac{\varepsilon}{2} \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

This proves the corollary.

Theorem 15.4.6 *Suppose $\rho\left(B^{-1}C\right) < 1$. Then the iterates in 15.21 converge to the unique solution of 15.20.*

Proof: Consider the iterates in 15.21. Let $T\mathbf{x} = B^{-1}C\mathbf{x} + \mathbf{b}$. Then

$$\begin{aligned}
\left|T^k\mathbf{x} - T^k\mathbf{y}\right| &= \left|\left(B^{-1}C\right)^k\mathbf{x} - \left(B^{-1}C\right)^k\mathbf{y}\right| \\
&\leq \left|\left|\left(B^{-1}C\right)^k\right|\right|\left|\mathbf{x} - \mathbf{y}\right|.
\end{aligned}$$

Here $||\cdot||$ refers to any of the operator norms. It doesn't matter which one you pick because they are all equivalent. I am writing the proof to indicate the operator norm taken with respect to the usual norm on E. Since $\rho\left(B^{-1}C\right) < 1$, it follows from Gelfand's theorem, Theorem 15.2.9 on Page 290, there exists N such that if $k \geq N$, then for some $r^{1/k} < 1$,

$$\left|\left|\left(B^{-1}C\right)^k\right|\right|^{1/k} < r^{1/k} < 1.$$

Consequently,

$$\left|T^N\mathbf{x} - T^N\mathbf{y}\right| \leq r\left|\mathbf{x} - \mathbf{y}\right|.$$

Also $|T\mathbf{x} - T\mathbf{y}| \leq \left|\left|B^{-1}C\right|\right|\left|\mathbf{x} - \mathbf{y}\right|$ and so Corollary 15.4.5 applies and gives the conclusion of this theorem.

15.5 Exercises

1. Solve the system
$$\begin{pmatrix} 4 & 1 & 1 \\ 1 & 5 & 2 \\ 0 & 2 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

2. Solve the system
$$\begin{pmatrix} 4 & 1 & 1 \\ 1 & 7 & 2 \\ 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

3. Solve the system
$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 7 & 2 \\ 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

4. If you are considering a system of the form $A\mathbf{x} = \mathbf{b}$ and A^{-1} does not exist, will either the Gauss Seidel or Jacobi methods work? Explain. What does this indicate about finding eigenvectors for a given eigenvalue?

15.6 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

$$\mathbf{u}_m = A^m \mathbf{u}_1 = \sum_{k=1}^{n-1} c_k \lambda_k^m \mathbf{x}_k + \lambda_n^m c_n \mathbf{x}_n. \tag{15.31}$$

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 15.31. 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 entry 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 15.6.1 *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

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.499\,26 \\ -1.466\,2 \times 10^{-3} \end{pmatrix} = \begin{pmatrix} 11.974 \\ -5.991\,2 \\ 8.838\,6 \times 10^{-3} \end{pmatrix}$$

and

$$12.053 \begin{pmatrix} 1.0 \\ -.499\,26 \\ -1.466\,2 \times 10^{-3} \end{pmatrix} = \begin{pmatrix} 12.053 \\ -6.017\,6 \\ -1.767\,2 \times 10^{-2} \end{pmatrix}.$$

15.6.1 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, I 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.

Lemma 15.6.2 *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 $\frac{1}{\lambda_k-\alpha}$. Conversely, if*

$$(A-\alpha I)^{-1}\mathbf{y} = \frac{1}{\lambda-\alpha}\mathbf{y} \tag{15.32}$$

and $\mathbf{y}\neq\mathbf{0}$, then $A\mathbf{y}=\lambda\mathbf{y}$. Furthermore, each generalized eigenspace is invariant with respect to $(A-\alpha I)^{-1}$. That is, $(A-\alpha I)^{-1}$ maps each generalized eigenspace to itself.

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

$$\frac{1}{\lambda_k-\alpha}\mathbf{x}_k = (A-\alpha I)^{-1}\mathbf{x}_k.$$

Suppose 15.32. Then $\mathbf{y} = \frac{1}{\lambda-\alpha}\left[A\mathbf{y}-\alpha\mathbf{y}\right]$. Solving for $A\mathbf{y}$ leads to $A\mathbf{y}=\lambda\mathbf{y}$.

It remains to verify the invariance of the generalized eigenspaces. Let E_k correspond to the eigenvalue λ_k.

$$(A-\lambda_k I)^m(A-\alpha I)^{-1}(A-\alpha I) = (A-\lambda_k I)^m = (A-\alpha I)^{-1}(A-\lambda_k I)^m(A-\alpha I)$$

and so it follows

$$(A-\lambda_k I)^m(A-\alpha I)^{-1} = (A-\alpha I)^{-1}(A-\lambda_k I)^m.$$

Now let $\mathbf{x}\in E_k$. Then this means that for some m, $(A-\lambda_k I)^m\mathbf{x}=\mathbf{0}$.

$$(A-\lambda_k I)^m(A-\alpha I)^{-1}\mathbf{x} = (A-\alpha I)^{-1}(A-\lambda_k I)^m\mathbf{x} = (A-\alpha I)^{-1}\mathbf{0}=\mathbf{0}.$$

Hence $(A-\alpha I)^{-1}\mathbf{x}\in E_k$ also. This proves the lemma.

Now assume α is closer to λ than to any other eigenvalue. Also suppose

$$\mathbf{u}_1 = \sum_{i=1}^k \mathbf{x}_i + \mathbf{y} \tag{15.33}$$

where $\mathbf{y}\neq\mathbf{0}$ and $\mathbf{y}$ is in the generalized eigenspace associated with λ and in addition is an eigenvector for A corresponding to λ. Let $\mathbf{x}_k$ be a vector in the generalized eigenspace associated with λ_k where the $\lambda_k\neq\lambda$. If $\mathbf{u}_n$ has been chosen,

$$\mathbf{u}_{n+1} \equiv \frac{(A-\alpha I)^{-1}\mathbf{u}_n}{S_{n+1}} \tag{15.34}$$

where S_{n+1} is a number with the property that $||\mathbf{u}_{n+1}||=1$. I am being vague about the particular choice of norm because it does not matter. One way to do this is to let S_{n+1} be the entry of $(A-\alpha I)^{-1}\mathbf{u}_n$ which has largest absolute value. Thus for $n>1$, $||\mathbf{u}_n||_\infty = 1$. This describes the method. Why does anything interesting happen?

$$\begin{aligned}
\mathbf{u}_2 &= \frac{\sum_{i=1}^k (A-\alpha I)^{-1}\mathbf{x}_i + (A-\alpha I)^{-1}\mathbf{y}}{S_2},\\
\mathbf{u}_3 &= \frac{\sum_{i=1}^k \left((A-\alpha I)^{-1}\right)^2\mathbf{x}_i + \left((A-\alpha I)^{-1}\right)^2\mathbf{y}}{S_2S_3}
\end{aligned}$$

and continuing this way,

$$\begin{aligned}\mathbf{u}_n &= \frac{\sum_{i=1}^k \left((A-\alpha I)^{-1}\right)^n \mathbf{x}_i + \left((A-\alpha I)^{-1}\right)^n \mathbf{y}}{S_2 S_3 \cdots S_n} \\ &= \frac{\sum_{i=1}^k \left((A-\alpha I)^{-1}\right)^n \mathbf{x}_i}{S_2 S_3 \cdots S_n} + \frac{\left((A-\alpha I)^{-1}\right)^n \mathbf{y}}{S_2 S_3 \cdots S_n} \qquad (15.35) \\ &\equiv \mathbf{r}_n + \mathbf{v}_n \qquad (15.36)\end{aligned}$$

Claim: $\mathbf{v}_n$ is an eigenvector for $(A-\alpha I)^{-1}$ and $\lim_{n\to\infty} \mathbf{r}_n = \mathbf{0}$.

Proof of claim: Consider $\mathbf{r}_n$. By invariance of $(A-\alpha I)^{-1}$ on the generalized eigenspaces, it follows from Gelfand's theorem that for n large enough

$$\begin{aligned}||\mathbf{r}_n|| &= \left|\left|\frac{\sum_{i=1}^k \left((A-\alpha I)^{-1}\right)^n \mathbf{x}_i}{S_2 S_3 \cdots S_n}\right|\right| \\ &\leq \frac{\sum_{i=1}^k \left|\left|\left((A-\alpha I)^{-1}\right)^n\right|\right| ||\mathbf{x}_i||}{S_2 S_3 \cdots S_n} \\ &\leq \frac{\sum_{i=1}^k \left|\frac{1+\eta}{\lambda_k-\alpha}\right|^n ||\mathbf{x}_i||}{S_2 S_3 \cdots S_n} = \left|\frac{1+\varepsilon}{\lambda_k-\alpha}\right|^n \frac{\sum_{i=1}^k ||\mathbf{x}_i||}{S_2 S_3 \cdots S_n}\end{aligned}$$

where η is chosen small enough that for all k,

$$\left|\frac{1+\eta}{\lambda_k-\alpha}\right| < \left|\frac{1}{\lambda-\alpha}\right| \qquad (15.37)$$

The second term on the right in 15.35 yields

$$\mathbf{v}_n = \frac{\mathbf{y}}{(\lambda-\alpha)^n S_2 S_3 \cdots S_n} = C_n \mathbf{y},$$

which is an eigenvector thanks to Lemma 15.6.2.

Now let $\varepsilon > 0$ be given, $\varepsilon < 1/2$. By 15.37, it follows that for large n,

$$\left|\frac{1}{\lambda-\alpha}\right|^n >> \left|\frac{1+\eta}{\lambda_k-\alpha}\right|^n$$

where $>>$ denotes much larger than. Therefore, for large n

$$\frac{||\mathbf{r}_n||}{||\mathbf{v}_n||} < \varepsilon. \qquad (15.38)$$

Then

$$||\mathbf{v}_n|| - ||\mathbf{u}_n|| \leq ||\mathbf{u}_n - \mathbf{v}_n|| = ||\mathbf{r}_n|| \leq \varepsilon ||\mathbf{v}_n||$$

and so $||\mathbf{v}_n|| \leq 1 + \varepsilon ||\mathbf{v}_n||$. Hence $||\mathbf{v}_n|| \leq 2$. Therefore, from 15.38,

$$||\mathbf{r}_n|| < \varepsilon ||\mathbf{v}_n|| < 2\varepsilon.$$

since $\varepsilon > 0$ is arbitrary it follows

$$\lim_{n\to\infty} \mathbf{r}_n = \mathbf{0}.$$

This verifies the claim.

Now from 15.34,

$$\begin{aligned} S_{n+1}\mathbf{u}_{n+1} &= (A-\alpha I)^{-1}\mathbf{r}_n + (A-\alpha I)^{-1}\mathbf{v}_n \\ &= (A-\alpha I)^{-1}\mathbf{r}_n + \frac{1}{\lambda-\alpha}\mathbf{v}_n \\ &= (A-\alpha I)^{-1}\mathbf{r}_n + \frac{1}{\lambda-\alpha}(\mathbf{u}_n - \mathbf{r}_n) \\ &= \left((A-\alpha I)^{-1} - \frac{I}{\lambda-\alpha}\right)\mathbf{r}_n + \frac{1}{\lambda-\alpha}\mathbf{u}_n \\ &= \mathbf{R}_n + \frac{1}{\lambda-\alpha}\mathbf{u}_n \end{aligned}$$

where $\lim_{n\to\infty}\mathbf{R}_n = \mathbf{0}$. Therefore,

$$\frac{S_{n+1}(\mathbf{u}_{n+1},\mathbf{u}_n) - (\mathbf{R}_n,\mathbf{u}_n)}{|\mathbf{u}_n|^2} = \frac{1}{\lambda-\alpha}.$$

It follows from $\lim_{n\to\infty}\mathbf{R}_n = \mathbf{0}$ that for large n,

$$\frac{S_{n+1}(\mathbf{u}_{n+1},\mathbf{u}_n)}{|\mathbf{u}_n|^2} \approx \frac{1}{\lambda-\alpha}$$

so you can solve this equation to obtain an approximate value for λ. If $\mathbf{u}_n \approx \mathbf{u}_{n+1}$, then this reduces to solving

$$S_{n+1} = \frac{1}{\lambda-\alpha}.$$

What about the case where S_{n+1} is the entry of $(A-\alpha I)^{-1}\mathbf{u}_n$ which has the largest absolute value? Can it be shown that for large $n, \mathbf{u}_n \approx \mathbf{u}_{n+1}$ and $S_n \approx S_{n+1}$? In this case, the norm is $||\cdot||_\infty$ and the construction requires that for $n > 1$,

$$\mathbf{u}_n = \begin{pmatrix} w_1 \\ \vdots \\ w_l \end{pmatrix}$$

where $|w_i| \le 1$ and for some $p, w_p = 1$. Then for large n,

$$(A-\alpha I)^{-1}\mathbf{u}_n \approx (A-\alpha I)^{-1}\mathbf{v}_n \approx \frac{1}{\lambda-\alpha}\mathbf{v}_n \approx \frac{1}{\lambda-\alpha}\mathbf{u}_n.$$

Therefore, you can take $S_{n+1} \approx \frac{1}{\lambda-\alpha}$ which shows that in this case, the scaling factors can be chosen to be a convergent sequence and that for large n they may all be considered to be approximately equal to $\frac{1}{\lambda-\alpha}$. Also,

$$\frac{1}{\lambda-\alpha}\mathbf{u}_{n+1} \approx (A-\alpha I)^{-1}\mathbf{u}_n \approx (A-\alpha I)^{-1}\mathbf{v}_n = \frac{1}{\lambda-\alpha}\mathbf{v}_n \approx \frac{1}{\lambda-\alpha}\mathbf{u}_n$$

and so for large n, $\mathbf{u}_{n+1} \approx \mathbf{u}_n \approx \mathbf{v}_n$, an eigenvector. Of course this last item could fail if $(A-\alpha I)^{-1}\mathbf{u}_n$ had more than one entry having absolute value equal to 1.

15.6.2 The Defective Case

In the case where the multiplicity of λ equals the dimension of the eigenspace for λ, the above is a good description of how to find both λ and an eigenvector which goes with λ.

This is because the whole space is the direct sum of the generalized eigenspaces and every nonzero vector in the generalized eigenspace associated with λ is an eigenvector. This is the case where λ is non defective. What of the case when the multiplicity of λ is greater than the dimension of the eigenspace associated with λ? Theoretically, the method will even work in this case although not very well. For the sake of completeness, I will give an argument which includes this case as well. As before, $||\mathbf{v}||$ will denote $||\mathbf{v}||_\infty$.

Let A be a $q \times q$ matrix and let α be closer to λ than to any other eigenvalue, the other ones being $\lambda_1, \lambda_2, \cdots, \lambda_p$. Then letting E_{λ_i} denote the generalized eigenspace associated with the eigenvector λ_i, let E denote the direct sum

$$E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_p},$$

it follows from Lemma 15.6.2 that E is invariant with respect to $(A - \alpha I)^{-1}$. Thus the eigenvalues of this linear transformation restricted to E are $\frac{1}{\lambda_i - \alpha}$ for $i = 1, \cdots, p$. Therefore, letting L_1 denote the largest of the quantities, $\frac{1}{|\lambda_i - \alpha|}$, it follows that for small enough η, $L_1 + \eta \equiv L < \left|\frac{1}{\lambda - \alpha}\right|$. Suppose then that

$$\mathbf{u}_1 \equiv \frac{(A - \alpha I)^{-1}(\mathbf{v} + \mathbf{y})}{S_1}, \ \mathbf{u}_{n+1} \equiv \frac{(A - \alpha I)^{-1} \mathbf{u}_n}{S_{n+1}}$$

where $\mathbf{v} \in E$ and $\mathbf{y} \in E_\lambda$ and it is assumed that $\mathbf{y} \neq \mathbf{0}$. As before, S_{n+1} denotes the entry of $(A - \alpha I)^{-1} \mathbf{u}_n$ which has the largest absolute value and S_1 the entry of $(A - \alpha I)^{-1}(\mathbf{v} + \mathbf{y})$ which has largest absolute value. Thus $||\mathbf{u}_n|| = 1$ for all n and $\mathbf{u}_n$ has at least one entry which equals 1. Thus

$$\mathbf{u}_{n+1} = \frac{\left((A - \alpha I)^{-1}\right)^n \mathbf{v} + \left((A - \alpha I)^{-1}\right)^n \mathbf{y}}{S_{n+1} S_n \cdots S_1}. \tag{15.39}$$

Since $\mathbf{y} \in E_\lambda$,

$$\mathbf{y} = \sum_{k=1}^{Q} \sum_{i=0}^{m_k} c_i^k \mathbf{x}_i^k$$

where $\mathbf{x}_0^k$ is an eigenvector for $1/(\lambda - \alpha)$ and $\mathbf{x}_1^k, \cdots, \mathbf{x}_{m_k}^k$ is a chain of generalized eigenvectors based on $\mathbf{x}_0^k$ satisfying

$$\left((A - \alpha I)^{-1} - \frac{1}{\lambda - \alpha} I\right) \mathbf{x}_l^k = \mathbf{x}_{l-1}^k$$

and

$$\left((A - \alpha I)^{-1} - \frac{1}{\lambda - \alpha} I\right)^l \mathbf{x}_l^k = \mathbf{x}_0^k,$$

the eigenvectors forming a linearly independent set.

Consider $\left((A - \alpha I)^{-1}\right)^n \mathbf{x}_l^k$. This equals

$$\begin{aligned}
&\left(\left((A - \alpha I)^{-1} - \frac{1}{\lambda - \alpha} I\right) + \frac{1}{\lambda - \alpha} I\right)^n \mathbf{x}_l^k \\
= &\sum_{\beta=0}^{l} \binom{n}{\beta} \left(\frac{1}{\lambda - \alpha}\right)^{n-\beta} \left((A - \alpha I)^{-1} - \frac{1}{\lambda - \alpha} I\right)^\beta \mathbf{x}_l^k \\
= &\binom{n}{l} \left(\frac{1}{\lambda - \alpha}\right)^{n-l} \mathbf{x}_0^k + \sum_{\beta=0}^{l-1} \binom{n}{\beta} \left(\frac{1}{\lambda - \alpha}\right)^{n-\beta} \mathbf{x}_{l-\beta}^k \\
\equiv &\ \mathbf{a}_n + \mathbf{b}_n
\end{aligned}$$

Of course each $l \leq q$. Also

$$\lim_{n\to\infty} \frac{\binom{n}{\beta}}{\binom{n}{l}} = 0$$

for all $\beta < l$. Therefore, $\lim_{n\to\infty} ||\mathbf{b}_n|| / ||\mathbf{a}_n|| = 0$. Separating all the terms of $\left((A-\alpha I)^{-1}\right)^n \mathbf{y}$ into the sum of eigenvectors, $\mathbf{A}_n$ and those which are not eigenvectors, $\mathbf{B}_n$, it follows that

$$\lim_{n\to\infty} ||\mathbf{B}_n|| / ||\mathbf{A}_n|| = 0.$$

Referring again to 15.39 and defining $\mathbf{C}_n \equiv \left((A-\alpha I)^{-1}\right)^n \mathbf{v}$, it follows that for n large enough,

$$||\mathbf{C}_n|| \leq \left|\left|\left((A-\alpha I)^{-1}\right)^n\right|\right| ||\mathbf{v}|| \leq L^n ||\mathbf{v}||$$

and so since $L < \left|\frac{1}{\lambda-\alpha}\right|$, it follows that

$$\lim_{n\to\infty} \frac{||\mathbf{C}_n||}{||\mathbf{A}_n||} = 0$$

also. Letting $\mathbf{D}_n = \mathbf{C}_n + \mathbf{B}_n$, it follows

$$\lim_{n\to\infty} \frac{||\mathbf{D}_n||}{||\mathbf{A}_n||} = 0$$

and

$$\mathbf{u}_{n+1} = \frac{\mathbf{A}_n + \mathbf{D}_n}{S_{n+1}S_n\cdots S_1},\ ||\mathbf{u}_{n+1}|| = 1,\ \mathbf{A}_n \text{ an eigenvector.} \tag{15.40}$$

Simplifying this further,

$$\mathbf{u}_{n+1} = \mathbf{P}_n + \mathbf{Q}_n,\ ||\mathbf{u}_{n+1}|| = 1,\ \mathbf{P}_n \text{ an eigenvector} \tag{15.41}$$

where

$$\mathbf{P}_n = \frac{\mathbf{A}_n}{S_{n+1}S_n\cdots S_1},\ \mathbf{Q}_n = \frac{\mathbf{D}_n}{S_{n+1}S_n\cdots S_1},\ \lim_{n\to\infty} \frac{||\mathbf{Q}_n||}{||\mathbf{P}_n||} = 0. \tag{15.42}$$

Claim: $\lim_{n\to\infty} ||\mathbf{Q}_n|| = 0$.

It follows from 15.42 that for all n large enough,

$$\frac{||\mathbf{Q}_n||}{||\mathbf{P}_n||} < \varepsilon < 1 \tag{15.43}$$

and so from 15.41 and large n,

$$||\mathbf{P}_n|| \leq 1 + ||\mathbf{Q}_n|| < 1 + \varepsilon ||\mathbf{P}_n||$$

so that

$$||\mathbf{P}_n|| < \frac{1}{1-\varepsilon}.$$

Therefore, from 15.43,

$$||\mathbf{Q}_n|| < \varepsilon ||\mathbf{P}_n|| < \frac{\varepsilon}{1-\varepsilon}.$$

Since ε is arbitrary, this proves the claim.

From the claim, it follows $\mathbf{u}_{n+1}$ is approximately equal to an eigenvector, $\mathbf{P}_n$ for large n. It remains to approximate the eigenvalue. This may be done by using this information.

$$(A-\alpha I)^{-1}\mathbf{u}_{n+1} \approx (A-\alpha I)^{-1}\mathbf{P}_n = \frac{1}{\lambda-\alpha}\mathbf{P}_n \approx \frac{1}{\lambda-\alpha}\mathbf{u}_{n+1} \tag{15.44}$$

and so you could estimate the eigenvalue as follows. Suppose the largest entry of $\mathbf{u}_{n+1}$ occurs in the j^{th} position. (In the above scheme, this entry will equal 1 but it really doesn't matter much how the S_m are chosen so long as they normalize the $\mathbf{u}_m$ in some norm.) Then compute λ by solving the following equation

$$\left((A-\alpha I)^{-1}\mathbf{u}_{n+1}\right)_j = \left(\frac{1}{\lambda-\alpha}\mathbf{u}_{n+1}\right)_j.$$

Another way would be to take the inner product of both ends of 15.44 with $\mathbf{u}_{n+1}$ and solve

$$\left((A-\alpha I)^{-1}\mathbf{u}_{n+1},\mathbf{u}_{n+1}\right) = \frac{1}{\lambda-\alpha}\left|\mathbf{u}_{n+1}\right|^2$$

for λ.

As before, unless something unusual happens, $S_{n+1} \approx \frac{1}{\lambda-\alpha}$ if you are using the above rule for selecting the S_m. This is because from the definition of the $\mathbf{u}_n$,

$$S_{n+1}\mathbf{u}_{n+1} \equiv (A-\alpha I)^{-1}\mathbf{u}_n \approx (A-\alpha I)^{-1}\mathbf{P}_{n-1} = \frac{1}{\lambda-\alpha}\mathbf{P}_{n-1} \approx \frac{1}{\lambda-\alpha}\mathbf{u}_n.$$

Therefore, the largest entry of $(A-\alpha I)^{-1}\mathbf{u}_n$ must be approximately equal to $\frac{1}{\lambda-\alpha}$ and so S_{n+1} is likely close to this number. Of course this could fail if $(A-\alpha I)^{-1}\mathbf{u}_n$ had many different entries all having the same absolute value.

15.6.3 The Explicit Description Of The Method

Here is how you use this 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,
$$\mathbf{u}_{k+1} = \frac{(A-\alpha I)^{-1}\mathbf{u}_k}{S_{k+1}}$$
where S_{k+1} is the entry 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
$$S_{k+1} = \frac{1}{\lambda-\alpha}$$
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 15.6.3 *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, I will start with an initial vector,

$$\mathbf{u}_1 \equiv \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Then I must solve the following equation.

$$\left(\left(\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} + 7 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right.$$

Simplifying the matrix on the left, I must solve

$$\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

and then divide by the entry which has largest absolute value to obtain

$$\mathbf{u}_2 = \begin{pmatrix} 1.0 \\ .184 \\ -.76 \end{pmatrix}$$

Now solve

$$\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0 \\ .184 \\ -.76 \end{pmatrix}$$

and divide by the largest entry, 1.0515 to get

$$\mathbf{u}_3 = \begin{pmatrix} 1.0 \\ .0266 \\ -.97061 \end{pmatrix}$$

Solve

$$\begin{pmatrix} 12 & -14 & 11 \\ -4 & 11 & -4 \\ 3 & 6 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0 \\ .0266 \\ -.97061 \end{pmatrix}$$

and divide by the largest entry, 1.01 to get

$$\mathbf{u}_4 = \begin{pmatrix} 1.0 \\ 3.8454 \times 10^{-3} \\ -.99604 \end{pmatrix}.$$

These scaling factors are pretty close after these few iterations. Therefore, the predicted eigenvalue is obtained by solving the following for λ.

$$\frac{1}{\lambda + 7} = 1.01$$

which gives $\lambda = -6.01$. You see this is pretty close. In this case the eigenvalue closest to -7 was -6.

Example 15.6.4 *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 $\alpha = -.855$. Then you solve

$$\left(\left(\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} + .855\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

or in other words,

$$\begin{pmatrix} 1.855 & 2.0 & 3.0 \\ 2.0 & 1.855 & 4.0 \\ 3.0 & 4.0 & 2.855 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

and divide by the largest entry, -67.944, to obtain

$$\mathbf{u}_2 = \begin{pmatrix} 1.0 \\ -.589\,21 \\ -.230\,44 \end{pmatrix}$$

Now solve

$$\begin{pmatrix} 1.855 & 2.0 & 3.0 \\ 2.0 & 1.855 & 4.0 \\ 3.0 & 4.0 & 2.855 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0 \\ -.589\,21 \\ -.230\,44 \end{pmatrix}$$

, Solution is : $\begin{pmatrix} -514.01 \\ 302.12 \\ 116.75 \end{pmatrix}$ and divide by the largest entry, -514.01, to obtain

$$\mathbf{u}_3 = \begin{pmatrix} 1.0 \\ -.587\,77 \\ -.227\,14 \end{pmatrix} \tag{15.45}$$

Clearly the $\mathbf{u}_k$ are not changing much. This suggests an approximate eigenvector for this eigenvalue which is close to $-.855$ is the above $\mathbf{u}_3$ and an eigenvalue is obtained by solving

$$\frac{1}{\lambda + .855} = -514.01,$$

which yields $\lambda = -.856\,9$ Lets check this.

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 2 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.587\,77 \\ -.227\,14 \end{pmatrix} = \begin{pmatrix} -.856\,96 \\ .503\,67 \\ .194\,64 \end{pmatrix}.$$

$$-.856\,9 \begin{pmatrix} 1.0 \\ -.587\,77 \\ -.227\,14 \end{pmatrix} = \begin{pmatrix} -.856\,9 \\ .503\,7 \\ .194\,6 \end{pmatrix}$$

Thus the vector of 15.45 is very close to the desired eigenvector, just as $-.856\,9$ is very close to the desired eigenvalue. For practical purposes, I have found both the eigenvector and the eigenvalue.

Example 15.6.5 *Find the eigenvalues and eigenvectors of the matrix,* $A = \begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix}$.

This is only a 3×3 matrix and so it is not hard to estimate the eigenvalues. Just get the characteristic equation, graph it using a calculator and zoom in to find the eigenvalues. If you do this, you find there is an eigenvalue near -1.2, one near $-.4$, and one near 5.5. (The characteristic equation is $2 + 8\lambda + 4\lambda^2 - \lambda^3 = 0$.) Of course I have no idea what the eigenvectors are.

Lets first try to find the eigenvector and a better approximation for the eigenvalue near -1.2. In this case, let $\alpha = -1.2$. Then

$$(A - \alpha I)^{-1} = \begin{pmatrix} -25.357\,143 & -33.928\,571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214\,286 & 30.357\,143 & -45.0 \end{pmatrix}.$$

Then for the first iteration, letting $\mathbf{u}_1 = (1,1,1)^T$,

$$\begin{pmatrix} -25.357\,143 & -33.928\,571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214\,286 & 30.357\,143 & -45.0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -9.285\,714 \\ 5.0 \\ 8.571\,429 \end{pmatrix}$$

To get $\mathbf{u}_2$, I must divide by $-9.285\,714$. Thus

$$\mathbf{u}_2 = \begin{pmatrix} 1.0 \\ -.538\,461\,56 \\ -.923\,077 \end{pmatrix}.$$

Do another iteration.

$$\begin{pmatrix} -25.357\,143 & -33.928\,571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214\,286 & 30.357\,143 & -45.0 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.538\,461\,56 \\ -.923\,077 \end{pmatrix} = \begin{pmatrix} -53.241\,762 \\ 26.153\,848 \\ 48.406\,596 \end{pmatrix}$$

Then to get $\mathbf{u}_3$ you divide by $-53.241\,762$. Thus

$$\mathbf{u}_3 = \begin{pmatrix} 1.0 \\ -.491\,228\,07 \\ -.909\,184\,71 \end{pmatrix}.$$

Now iterate again because the scaling factors are still changing quite a bit.

$$\begin{pmatrix} -25.357\,143 & -33.928\,571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214\,286 & 30.357\,143 & -45.0 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.491\,228\,07 \\ -.909\,184\,71 \end{pmatrix} = \begin{pmatrix} -54.149\,712 \\ 26.633\,127 \\ 49.215\,317 \end{pmatrix}.$$

This time the scaling factor didn't change too much. It is $-54.149\,712$. Thus

$$\mathbf{u}_4 = \begin{pmatrix} 1.0 \\ -.491\,842\,45 \\ -.908\,874\,95 \end{pmatrix}.$$

Lets do one more iteration.

$$\begin{pmatrix} -25.357\,143 & -33.928\,571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214\,286 & 30.357\,143 & -45.0 \end{pmatrix} \begin{pmatrix} 1.0 \\ -.491\,842\,45 \\ -.908\,874\,95 \end{pmatrix} = \begin{pmatrix} -54.113\,379 \\ 26.614\,631 \\ 49.182\,727 \end{pmatrix}.$$

You see at this point the scaling factors have definitely settled down and so it seems our eigenvalue would be obtained by solving

$$\frac{1}{\lambda - (-1.2)} = -54.113\,379$$

and this yields $\lambda = -1.218\,479\,7$ as an approximation to the eigenvalue and the eigenvector would be obtained by dividing by $-54.113\,379$ which gives

$$\mathbf{u}_5 = \begin{pmatrix} 1.000\,000\,2 \\ -.491\,830\,97 \\ -.908\,883\,09 \end{pmatrix}.$$

How well does it work?

$$\begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1.000\,000\,2 \\ -.491\,830\,97 \\ -.908\,883\,09 \end{pmatrix} = \begin{pmatrix} -1.218\,479\,8 \\ .599\,286\,34 \\ 1.107\,455\,6 \end{pmatrix}$$

while

$$-1.218\,479\,7 \begin{pmatrix} 1.000\,000\,2 \\ -.491\,830\,97 \\ -.908\,883\,09 \end{pmatrix} = \begin{pmatrix} -1.218\,479\,9 \\ .599\,286\,05 \\ 1.107\,455\,6 \end{pmatrix}.$$

For practical purposes, this has found the eigenvalue near -1.2 as well as an eigenvector associated with it.

Next I shall find the eigenvector and a more precise value for the eigenvalue near $-.4$. In this case,

$$(A-\alpha I)^{-1} = \begin{pmatrix} 8.064\,516\,1\times 10^{-2} & -9.274\,193\,5 & 6.451\,612\,9 \\ -.403\,225\,81 & 11.370\,968 & -7.258\,064\,5 \\ .403\,225\,81 & 3.629\,032\,3 & -2.741\,935\,5 \end{pmatrix}.$$

As before, I have no idea what the eigenvector is so I will again try $(1,1,1)^T$. Then to find $\mathbf{u}_2$,

$$\begin{pmatrix} 8.064\,516\,1\times 10^{-2} & -9.274\,193\,5 & 6.451\,612\,9 \\ -.403\,225\,81 & 11.370\,968 & -7.258\,064\,5 \\ .403\,225\,81 & 3.629\,032\,3 & -2.741\,935\,5 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -2.741\,935\,4 \\ 3.709\,677\,7 \\ 1.290\,322\,6 \end{pmatrix}$$

The scaling factor is $3.709\,677\,7$. Thus

$$\mathbf{u}_2 = \begin{pmatrix} -.739\,130\,36 \\ 1.0 \\ .347\,826\,07 \end{pmatrix}.$$

Now lets do another iteration.

$$\begin{pmatrix} 8.064\,516\,1\times 10^{-2} & -9.274\,193\,5 & 6.451\,612\,9 \\ -.403\,225\,81 & 11.370\,968 & -7.258\,064\,5 \\ .403\,225\,81 & 3.629\,032\,3 & -2.741\,935\,5 \end{pmatrix} \begin{pmatrix} -.739\,130\,36 \\ 1.0 \\ .347\,826\,07 \end{pmatrix} = \begin{pmatrix} -7.089\,761\,6 \\ 9.144\,460\,4 \\ 2.377\,279\,2 \end{pmatrix}.$$

The scaling factor is $9.144\,460\,4$. Thus

$$\mathbf{u}_3 = \begin{pmatrix} -.775\,306\,72 \\ 1.0 \\ .259\,969\,33 \end{pmatrix}.$$

Lets do another iteration. The scaling factors are still changing quite a bit.

$$\begin{pmatrix} 8.064\,516\,1\times 10^{-2} & -9.274\,193\,5 & 6.451\,612\,9 \\ -.403\,225\,81 & 11.370\,968 & -7.258\,064\,5 \\ .403\,225\,81 & 3.629\,032\,3 & -2.741\,935\,5 \end{pmatrix} \begin{pmatrix} -.775\,306\,72 \\ 1.0 \\ .259\,969\,33 \end{pmatrix} = \begin{pmatrix} -7.659\,496\,8 \\ 9.796\,717\,5 \\ 2.603\,589\,5 \end{pmatrix}.$$

The scaling factor is now $9.796\,717\,5$. Therefore,

$$\mathbf{u}_4 = \begin{pmatrix} -.781\,843\,18 \\ 1.0 \\ .265\,761\,41 \end{pmatrix}.$$

Lets do another iteration.

$$\begin{pmatrix} 8.064\,516\,1\times 10^{-2} & -9.274\,193\,5 & 6.451\,612\,9 \\ -.403\,225\,81 & 11.370\,968 & -7.258\,064\,5 \\ .403\,225\,81 & 3.629\,032\,3 & -2.741\,935\,5 \end{pmatrix} \begin{pmatrix} -.781\,843\,18 \\ 1.0 \\ .265\,761\,41 \end{pmatrix} = \begin{pmatrix} -7.622\,655\,6 \\ 9.757\,313\,9 \\ 2.585\,072\,3 \end{pmatrix}.$$

Now the scaling factor is $9.757\,313\,9$ and so

$$\mathbf{u}_5 = \begin{pmatrix} -.781\,224\,8 \\ 1.0 \\ .264\,936\,88 \end{pmatrix}.$$

I notice the scaling factors are not changing by much so the approximate eigenvalue is

$$\frac{1}{\lambda + .4} = 9.\,757\,313\,9$$

which shows $\lambda = -.\,297\,512\,78$ is an approximation to the eigenvalue near .4. How well does it work?

$$\begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} -.\,781\,224\,8 \\ 1.\,0 \\ .\,264\,936\,88 \end{pmatrix} = \begin{pmatrix} .\,232\,361\,04 \\ -.\,297\,512\,72 \\ -.0\,787\,375\,2 \end{pmatrix}.$$

$$-.\,297\,512\,78 \begin{pmatrix} -.\,781\,224\,8 \\ 1.\,0 \\ .\,264\,936\,88 \end{pmatrix} = \begin{pmatrix} .\,232\,424\,36 \\ -.\,297\,512\,78 \\ -7.\,882\,210\,8 \times 10^{-2} \end{pmatrix}.$$

It works pretty well. For practical purposes, the eigenvalue and eigenvector have now been found. If you want better accuracy, you could just continue iterating.

Next I will find the eigenvalue and eigenvector for the eigenvalue near 5.5. In this case,

$$(A - \alpha I)^{-1} = \begin{pmatrix} 29.\,2 & 16.\,8 & 23.\,2 \\ 19.\,2 & 10.\,8 & 15.\,2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix}.$$

As before, I have no idea what the eigenvector is but I am tired of always using $(1,1,1)^T$ and I don't want to give the impression that you always need to start with this vector. Therefore, I shall let $\mathbf{u}_1 = (1,2,3)^T$. What follows is the iteration without all the comments between steps.

$$\begin{pmatrix} 29.\,2 & 16.\,8 & 23.\,2 \\ 19.\,2 & 10.\,8 & 15.\,2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 1.\,324 \times 10^2 \\ 86.\,4 \\ 1.\,26 \times 10^2 \end{pmatrix}.$$

$S_2 = 86.\,4.$

$$\mathbf{u}_2 = \begin{pmatrix} 1.\,532\,407\,4 \\ 1.0 \\ 1.\,458\,333\,3 \end{pmatrix}.$$

$$\begin{pmatrix} 29.\,2 & 16.\,8 & 23.\,2 \\ 19.\,2 & 10.\,8 & 15.\,2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix} \begin{pmatrix} 1.\,532\,407\,4 \\ 1.0 \\ 1.\,458\,333\,3 \end{pmatrix} = \begin{pmatrix} 95.\,379\,629 \\ 62.\,388\,888 \\ 90.\,990\,74 \end{pmatrix}$$

$S_3 = 95.\,379\,629.$

$$\mathbf{u}_3 = \begin{pmatrix} 1.\,0 \\ .\,654\,111\,25 \\ .\,953\,985\,05 \end{pmatrix}$$

$$\begin{pmatrix} 29.\,2 & 16.\,8 & 23.\,2 \\ 19.\,2 & 10.\,8 & 15.\,2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix} \begin{pmatrix} 1.\,0 \\ .\,654\,111\,25 \\ .\,953\,985\,05 \end{pmatrix} = \begin{pmatrix} 62.\,321\,522 \\ 40.\,764\,974 \\ 59.\,453\,451 \end{pmatrix}$$

$S_4 = 62.\,321\,522.$

$$\mathbf{u}_4 = \begin{pmatrix} 1.0 \\ .\,654\,107\,48 \\ .\,953\,979\,45 \end{pmatrix}$$

$$\begin{pmatrix} 29.\,2 & 16.\,8 & 23.\,2 \\ 19.\,2 & 10.\,8 & 15.\,2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix} \begin{pmatrix} 1.0 \\ .\,654\,107\,48 \\ .\,953\,979\,45 \end{pmatrix} = \begin{pmatrix} 62.\,321\,329 \\ 40.\,764\,848 \\ 59.\,453\,268 \end{pmatrix}$$

$S_5 = 62.321\,329$. Looks like it is time to stop because this scaling factor is not changing much from S_3.

$$\mathbf{u}_5 = \begin{pmatrix} 1.0 \\ .654\,107\,49 \\ .953\,979\,46 \end{pmatrix}.$$

Then the approximation of the eigenvalue is gotten by solving

$$62.321\,329 = \frac{1}{\lambda - 5.5}$$

which gives $\lambda = 5.516\,045\,9$. Lets see how well it works.

$$\begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1.0 \\ .654\,107\,49 \\ .953\,979\,46 \end{pmatrix} = \begin{pmatrix} 5.516\,045\,9 \\ 3.608\,087 \\ 5.262\,194\,4 \end{pmatrix}$$

$$5.516\,045\,9 \begin{pmatrix} 1.0 \\ .654\,107\,49 \\ .953\,979\,46 \end{pmatrix} = \begin{pmatrix} 5.516\,045\,9 \\ 3.608\,086\,9 \\ 5.262\,194\,5 \end{pmatrix}.$$

15.6.4 Complex Eigenvalues

What about complex eigenvalues? If your matrix is real, you won't see these by graphing the characteristic equation on your calculator. Will the shifted inverse power method find these eigenvalues and their associated eigenvectors? The answer is yes. However, for a real matrix, you must pick α to be complex. This is because the eigenvalues occur in conjugate pairs so if you don't pick it complex, it will be the same distance between any conjugate pair of complex numbers and so nothing in the above argument for convergence implies you will get convergence to a complex number. Also, the process of iteration will yield only real vectors and scalars.

Example 15.6.6 *Find the complex eigenvalues and corresponding eigenvectors for the matrix,*

$$\begin{pmatrix} 5 & -8 & 6 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Here the characteristic equation is $\lambda^3 - 5\lambda^2 + 8\lambda - 6 = 0$. One solution is $\lambda = 3$. The other two are $1+i$ and $1-i$. We will apply the process to $\alpha = i$ so we will find the eigenvalue closest to i.

$$(A - \alpha I)^{-1} = \begin{pmatrix} -.02 - .14i & 1.24 + .68i & -.84 + .12i \\ -.14 + .02i & .68 - .24i & .12 + .84i \\ .02 + .14i & -.24 - .68i & .84 + .88i \end{pmatrix}$$

Then let $\mathbf{u}_1 = (1,1,1)^T$ for lack of any insight into anything better.

$$\begin{pmatrix} -.02 - .14i & 1.24 + .68i & -.84 + .12i \\ -.14 + .02i & .68 - .24i & .12 + .84i \\ .02 + .14i & -.24 - .68i & .84 + .88i \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} .38 + .66i \\ .66 + .62i \\ .62 + .34i \end{pmatrix}$$

$S_2 = .66 + .62i$.

$$\mathbf{u}_2 = \begin{pmatrix} .804\,878\,05 + .243\,902\,44i \\ 1.0 \\ .756\,097\,56 - .195\,121\,95i \end{pmatrix}$$

$$\begin{pmatrix} -.02-.14i & 1.24+.68i & -.84+.12i \\ -.14+.02i & .68-.24i & .12+.84i \\ .02+.14i & -.24-.68i & .84+.88i \end{pmatrix} \begin{pmatrix} .804\,878\,05+.243\,902\,44i \\ 1.0 \\ .756\,097\,56-.195\,121\,95i \end{pmatrix}$$
$$= \begin{pmatrix} .646\,341\,46+.817\,073\,17i \\ .817\,073\,17+.353\,658\,54i \\ .548\,780\,49-6.097\,560\,9\times 10^{-2}i \end{pmatrix}$$

$S_3 = .646\,341\,46 + .817\,073\,17i$. After more iterations, of this sort, you find $S_9 = 1.002\,748\,5 + 2.137\,621\,7\times 10^{-4}i$ and

$$\mathbf{u}_9 = \begin{pmatrix} 1.0 \\ .501\,514\,17-.499\,807\,33i \\ 1.562\,088\,1\times 10^{-3}-.499\,778\,55i \end{pmatrix}.$$

Then

$$\begin{pmatrix} -.02-.14i & 1.24+.68i & -.84+.12i \\ -.14+.02i & .68-.24i & .12+.84i \\ .02+.14i & -.24-.68i & .84+.88i \end{pmatrix} \begin{pmatrix} 1.0 \\ .501\,514\,17-.499\,807\,33i \\ 1.562\,088\,1\times 10^{-3}-.499\,778\,55i \end{pmatrix}$$
$$= \begin{pmatrix} 1.000\,407\,8+1.269\,979\times 10^{-3}i \\ .501\,077\,31-.498\,893\,66i \\ 8.848\,928\times 10^{-4}-.499\,515\,22i \end{pmatrix}$$

$S_{10} = 1.000\,407\,8 + 1.269\,979\times 10^{-3}i$.

$$\mathbf{u}_{10} = \begin{pmatrix} 1.0 \\ .500\,239\,18-.499\,325\,33i \\ 2.506\,749\,2\times 10^{-4}-.499\,311\,92i \end{pmatrix}$$

The scaling factors are not changing much at this point

$$1.000\,407\,8+1.269\,979\times 10^{-3}i = \frac{1}{\lambda - i}$$

The approximate eigenvalue is then $\lambda = .999\,590\,76 + .998\,731\,06i$. This is pretty close to $1+i$. How well does the eigenvector work?

$$\begin{pmatrix} 5 & -8 & 6 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1.0 \\ .500\,239\,18-.499\,325\,33i \\ 2.506\,749\,2\times 10^{-4}-.499\,311\,92i \end{pmatrix}$$
$$= \begin{pmatrix} .999\,590\,61+.998\,731\,12i \\ 1.0 \\ .500\,239\,18-.499\,325\,33i \end{pmatrix}$$

$$(.999\,590\,76+.998\,731\,06i) \begin{pmatrix} 1.0 \\ .500\,239\,18-.499\,325\,33i \\ 2.506\,749\,2\times 10^{-4}-.499\,311\,92i \end{pmatrix}$$
$$= \begin{pmatrix} .999\,590\,76+.998\,731\,06i \\ .998\,726\,18+4.834\,203\,9\times 10^{-4}i \\ .498\,928\,9-.498\,857\,22i \end{pmatrix}$$

It took more iterations than before because α was not very close to $1+i$.

This illustrates an interesting topic which leads to many related topics. If you have a polynomial, $x^4 + ax^3 + bx^2 + cx + d$, you can consider it as the characteristic polynomial of a certain matrix, called a **companion matrix**. In this case,

$$\begin{pmatrix} -a & -b & -c & -d \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The above example was just a companion matrix for $\lambda^3 - 5\lambda^2 + 8\lambda - 6$. You can see the pattern which will enable you to obtain a companion matrix for any polynomial of the form $\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n$. This illustrates that one way to find the complex zeros of a polynomial is to use the shifted inverse power method on a companion matrix for the polynomial. Doubtless there are better ways but this does illustrate how impressive this procedure is. Do you have a better way?

15.6.5 Rayleigh Quotients And Estimates for Eigenvalues

There are many specialized results concerning the eigenvalues and eigenvectors for Hermitian matrices. Recall a matrix, A is Hermitian if $A = A^*$ where A^* means to take the transpose of the conjugate of A. In the case of a real matrix, Hermitian reduces to symmetric. Recall also that for $\mathbf{x} \in \mathbb{F}^n$,

$$|\mathbf{x}|^2 = \mathbf{x}^*\mathbf{x} = \sum_{j=1}^{n} |x_j|^2 .$$

Recall the following corollary found on Page 170 which is stated here for convenience.

Corollary 15.6.7 *If A is Hermitian, then all the eigenvalues of A are real and there exists an orthonormal basis of eigenvectors.*

Thus for $\{\mathbf{x}_k\}_{k=1}^n$ this orthonormal basis,

$$\mathbf{x}_i^*\mathbf{x}_j = \delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

For $\mathbf{x} \in \mathbb{F}^n$, $\mathbf{x} \neq \mathbf{0}$, the Rayleigh quotient is defined by

$$\frac{\mathbf{x}^* A \mathbf{x}}{|\mathbf{x}|^2}.$$

Now let the eigenvalues of A be $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $A\mathbf{x}_k = \lambda_k \mathbf{x}_k$ where $\{\mathbf{x}_k\}_{k=1}^n$ is the above orthonormal basis of eigenvectors mentioned in the corollary. Then if $\mathbf{x}$ is an arbitrary vector, there exist constants, a_i such that

$$\mathbf{x} = \sum_{i=1}^{n} a_i \mathbf{x}_i.$$

Also,

$$\begin{aligned} |\mathbf{x}|^2 &= \sum_{i=1}^{n} \overline{a}_i \mathbf{x}_i^* \sum_{j=1}^{n} a_j \mathbf{x}_j \\ &= \sum_{ij} \overline{a}_i a_j \mathbf{x}_i^* \mathbf{x}_j = \sum_{ij} \overline{a}_i a_j \delta_{ij} = \sum_{i=1}^{n} |a_i|^2 . \end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\mathbf{x}^*A\mathbf{x}}{|\mathbf{x}|^2} &= \frac{\left(\sum_{i=1}^n \overline{a}_i\mathbf{x}_i^*\right)\left(\sum_{j=1}^n a_j\lambda_j\mathbf{x}_j\right)}{\sum_{i=1}^n |a_i|^2} \\ &= \frac{\sum_{ij}\overline{a}_i a_j\lambda_j\mathbf{x}_i^*\mathbf{x}_j}{\sum_{i=1}^n |a_i|^2} = \frac{\sum_{ij}\overline{a}_i a_j\lambda_j\delta_{ij}}{\sum_{i=1}^n |a_i|^2} \\ &= \frac{\sum_{i=1}^n |a_i|^2\lambda_i}{\sum_{i=1}^n |a_i|^2} \in [\lambda_1,\lambda_n]\,.\end{aligned}$$

In other words, the Rayleigh quotient is always between the largest and the smallest eigenvalues of A. When $\mathbf{x}=\mathbf{x}_n$, the Rayleigh quotient equals the largest eigenvalue and when $\mathbf{x}=\mathbf{x}_1$ the Rayleigh quotient equals the smallest eigenvalue. Suppose you calculate a Rayleigh quotient. How close is it to some eigenvalue?

Theorem 15.6.8 *Let $\mathbf{x}\neq\mathbf{0}$ and form the Rayleigh quotient,*

$$\frac{\mathbf{x}^*A\mathbf{x}}{|\mathbf{x}|^2}\equiv q.$$

Then there exists an eigenvalue of A, denoted here by λ_q such that

$$|\lambda_q - q| \leq \frac{|A\mathbf{x}-q\mathbf{x}|}{|\mathbf{x}|}. \tag{15.46}$$

Proof: Let $\mathbf{x}=\sum_{k=1}^n a_k\mathbf{x}_k$ where $\{\mathbf{x}_k\}_{k=1}^n$ is the orthonormal basis of eigenvectors.

$$\begin{aligned}|A\mathbf{x}-q\mathbf{x}|^2 &= (A\mathbf{x}-q\mathbf{x})^*(A\mathbf{x}-q\mathbf{x}) \\ &= \left(\sum_{k=1}^n a_k\lambda_k\mathbf{x}_k - qa_k\mathbf{x}_k\right)^*\left(\sum_{k=1}^n a_k\lambda_k\mathbf{x}_k - qa_k\mathbf{x}_k\right) \\ &= \left(\sum_{j=1}^n (\lambda_j-q)\,\overline{a}_j\mathbf{x}_j^*\right)\left(\sum_{k=1}^n (\lambda_k-q)\,a_k\mathbf{x}_k\right) \\ &= \sum_{j,k}(\lambda_j-q)\,\overline{a}_j\,(\lambda_k-q)\,a_k\mathbf{x}_j^*\mathbf{x}_k \\ &= \sum_{k=1}^n |a_k|^2(\lambda_k-q)^2\end{aligned}$$

Now pick the eigenvalue, λ_q which is closest to q. Then

$$|A\mathbf{x}-q\mathbf{x}|^2 = \sum_{k=1}^n |a_k|^2(\lambda_k-q)^2 \geq (\lambda_q-q)^2\sum_{k=1}^n |a_k|^2 = (\lambda_q-q)^2|\mathbf{x}|^2$$

which implies 15.46.

Example 15.6.9 *Consider the symmetric matrix, $A=\begin{pmatrix}1&2&3\\2&2&1\\3&1&4\end{pmatrix}$. Let $\mathbf{x}=(1,1,1)^T$. How close is the Rayleigh quotient to some eigenvalue of A? Find the eigenvector and eigenvalue to several decimal places.*

Everything is real and so there is no need to worry about taking conjugates. Therefore, the Rayleigh quotient is

$$\frac{\begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}}{3} = \frac{19}{3}$$

According to the above theorem, there is some eigenvalue of this matrix, λ_q such that

$$\begin{aligned} \left|\lambda_q - \frac{19}{3}\right| &\leq \frac{\left|\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \frac{19}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\right|}{\sqrt{3}} = \frac{1}{\sqrt{3}} \begin{pmatrix} -\frac{1}{3} \\ -\frac{4}{3} \\ \frac{5}{3} \end{pmatrix} \\ &= \frac{\sqrt{\frac{1}{9} + \left(\frac{4}{3}\right)^2 + \left(\frac{5}{3}\right)^2}}{\sqrt{3}} = 1.247\,2 \end{aligned}$$

Could you find this eigenvalue and associated eigenvector? Of course you could. This is what the shifted inverse power method is all about.

Solve

$$\left(\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} - \frac{19}{3} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

In other words solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

and divide by the entry which is largest, $3.870\,7$, to get

$$\mathbf{u}_2 = \begin{pmatrix} .699\,25 \\ .493\,89 \\ 1.0 \end{pmatrix}$$

Now solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .699\,25 \\ .493\,89 \\ 1.0 \end{pmatrix}$$

and divide by the largest entry, $2.997\,9$ to get

$$\mathbf{u}_3 = \begin{pmatrix} .714\,73 \\ .522\,63 \\ 1.0 \end{pmatrix}$$

Now solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .714\,73 \\ .522\,63 \\ 1.0 \end{pmatrix}$$

and divide by the largest entry, $3.045\,4$, to get

$$\mathbf{u}_4 = \begin{pmatrix} .713\,7 \\ .520\,56 \\ 1.0 \end{pmatrix}$$

Solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .713\,7 \\ .520\,56 \\ 1.0 \end{pmatrix}$$

and divide by the largest entry, 3.042 1 to get

$$\mathbf{u}_5 = \begin{pmatrix} .713\,78 \\ .520\,73 \\ 1.0 \end{pmatrix}$$

You can see these scaling factors are not changing much. The predicted eigenvalue is then about

$$\frac{1}{3.042\,1} + \frac{19}{3} = 6.662\,1.$$

How close is this?

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} \begin{pmatrix} .713\,78 \\ .520\,73 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 4.755\,2 \\ 3.469 \\ 6.662\,1 \end{pmatrix}$$

while

$$6.662\,1 \begin{pmatrix} .713\,78 \\ .520\,73 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 4.755\,3 \\ 3.469\,2 \\ 6.662\,1 \end{pmatrix}.$$

You see that for practical purposes, this has found the eigenvalue.

15.7 Exercises

1. In Example 15.6.9 an eigenvalue was found correct to several decimal places along with an eigenvector. Find the other eigenvalues along with their eigenvectors.

2. Find the eigenvalues and eigenvectors of the matrix, $A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 1 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.
   In this case the exact eigenvalues are $\pm\sqrt{3}, 6$. Compare with the exact answers.

3. Find the eigenvalues and eigenvectors of the matrix, $A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.
   The exact eigenvalues are $2, 4+\sqrt{15}, 4-\sqrt{15}$. Compare your numerical results with the exact values. Is it much fun to compute the exact eigenvectors?

4. Find the eigenvalues and eigenvectors of the matrix, $A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.
   I don't know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

5. Find the eigenvalues and eigenvectors of the matrix, $A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically.
   I don't know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

6. Consider the matrix, $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 0 \end{pmatrix}$ and the vector $(1,1,1)^T$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of A.

7. Consider the matrix, $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 1 & 4 \\ 1 & 4 & 5 \end{pmatrix}$ and the vector $(1,1,1)^T$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of A.

8. Consider the matrix, $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3 \end{pmatrix}$ and the vector $(1,1,1)^T$. Find the shortest distance between the Rayleigh quotient determined by this vector and some eigenvalue of A.

9. Using Gerschgorin's theorem, find upper and lower bounds for the eigenvalues of $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3 \end{pmatrix}$.

15.8 Positive Matrices

Earlier theorems about Markov matrices were presented. These were matrices in which all the entries were nonnegative and either the columns or the rows added to 1. It turns out that many of the theorems presented can be generalized to positive matrices. When this is done, the resulting theory is mainly due to Perron and Frobenius. I will give an introduction to this theory here following Karlin and Taylor [9].

Definition 15.8.1 *For A a matrix or vector, the notation, $A >> 0$ will mean every entry of A is positive. By $A > 0$ is meant that every entry is nonnegative and at least one is positive. By $A \geq 0$ is meant that every entry is nonnegative. Thus the matrix or vector consisting only of zeros is ≥ 0. An expression like $A >> B$ will mean $A - B >> 0$ with similar modifications for $>$ and $\geq$.*

For the sake of this section only, define the following for $\mathbf{x} = (x_1, \cdots, x_n)^T$, a vector.

$$|\mathbf{x}| \equiv (|x_1|, \cdots, |x_n|)^T.$$

Thus $|\mathbf{x}|$ is the vector which results by replacing each entry of $\mathbf{x}$ with its absolute value[3]*. Also define for $\mathbf{x} \in \mathbb{C}^n$,*

$$||\mathbf{x}||_1 \equiv \sum_k |x_k|.$$

Lemma 15.8.2 *Let $A >> 0$ and let $\mathbf{x} > \mathbf{0}$. Then $A\mathbf{x} >> \mathbf{0}$.*

Proof: $(A\mathbf{x})_i = \sum_j A_{ij} x_j > 0$ because all the $A_{ij} > 0$ and at least one $x_j > 0$.

[3]This notation is just about the most abominable thing imaginable. However, it saves space in the presentation of this theory of positive matrices and avoids the use of new symbols. Please forget about it when you leave this section.

Lemma 15.8.3 *Let $A >> 0$. Define*

$$S \equiv \{\lambda : A\mathbf{x} > \lambda\mathbf{x} \text{ for some } \mathbf{x} >> \mathbf{0}\},$$

and let

$$K \equiv \{\mathbf{x} \geq \mathbf{0} \text{ such that } ||\mathbf{x}||_1 = 1\}.$$

Now define

$$S_1 \equiv \{\lambda : A\mathbf{x} \geq \lambda\mathbf{x} \text{ for some } \mathbf{x} \in K\}.$$

Then

$$\sup(S) = \sup(S_1).$$

Proof: Let $\lambda \in S$. Then there exists $\mathbf{x} >> \mathbf{0}$ such that $A\mathbf{x} > \lambda\mathbf{x}$. Consider $\mathbf{y} \equiv \mathbf{x}/||\mathbf{x}||_1$. Then $||\mathbf{y}||_1 = 1$ and $A\mathbf{y} > \lambda\mathbf{y}$. Therefore, $\lambda \in S_1$ and so $S \subseteq S_1$. Therefore, $\sup(S) \leq \sup(S_1)$.

Now let $\lambda \in S_1$. Then there exists $\mathbf{x} \geq \mathbf{0}$ such that $||\mathbf{x}||_1 = 1$ so $\mathbf{x} > \mathbf{0}$ and $A\mathbf{x} > \lambda\mathbf{x}$. Letting $\mathbf{y} \equiv A\mathbf{x}$, it follows from Lemma 15.8.2 that $A\mathbf{y} >> \lambda\mathbf{y}$ and $\mathbf{y} >> \mathbf{0}$. Thus $\lambda \in S$ and so $S_1 \subseteq S$ which shows that $\sup(S_1) \leq \sup(S)$. This proves the lemma.

This lemma is significant because the set, $\{\mathbf{x} \geq \mathbf{0} \text{ such that } ||\mathbf{x}||_1 = 1\} \equiv K$ is a compact set in $\mathbb{R}^n$. Define

$$\lambda_0 \equiv \sup(S) = \sup(S_1). \tag{15.47}$$

The following theorem is due to Perron.

Theorem 15.8.4 *Let $A >> 0$ be an $n \times n$ matrix and let λ_0 be given in 15.47. Then*

1. *$\lambda_0 > 0$ and there exists $\mathbf{x}_0 >> \mathbf{0}$ such that $A\mathbf{x}_0 = \lambda_0\mathbf{x}_0$ so λ_0 is an eigenvalue for A.*
2. *If $A\mathbf{x} = \mu\mathbf{x}$ where $\mathbf{x} \neq \mathbf{0}$, and $\mu \neq \lambda_0$. Then $|\mu| < \lambda_0$.*
3. *The eigenspace for λ_0 has dimension 1.*

Proof: To see $\lambda_0 > 0$, consider the vector, $\mathbf{e} \equiv (1, \cdots, 1)^T$. Then

$$(A\mathbf{e})_i = \sum_j A_{ij} > 0$$

and so λ_0 is at least as large as

$$\min_i \sum_j A_{ij}.$$

Let $\{\lambda_k\}$ be an increasing sequence of numbers from S_1 converging to λ_0. Letting $\mathbf{x}_k$ be the vector from K which occurs in the definition of S_1, these vectors are in a compact set. Therefore, there exists a subsequence, still denoted by $\mathbf{x}_k$ such that $\mathbf{x}_k \to \mathbf{x}_0 \in K$ and $\lambda_k \to \lambda_0$. Then passing to the limit,

$$A\mathbf{x}_0 \geq \lambda_0\mathbf{x}_0, \ \mathbf{x}_0 > \mathbf{0}.$$

If $A\mathbf{x}_0 > \lambda_0\mathbf{x}_0$, then letting $\mathbf{y} \equiv A\mathbf{x}_0$, it follows from Lemma 15.8.2 that $A\mathbf{y} >> \lambda_0\mathbf{y}$ and $\mathbf{y} >> \mathbf{0}$. But this contradicts the definition of λ_0 as the supremum of the elements of S because since $A\mathbf{y} >> \lambda_0\mathbf{y}$, it follows $A\mathbf{y} >> (\lambda_0 + \varepsilon)\mathbf{y}$ for ε a small positive number. Therefore, $A\mathbf{x}_0 = \lambda_0\mathbf{x}_0$. It remains to verify that $\mathbf{x}_0 >> \mathbf{0}$. But this follows immediately from

$$0 < \sum_j A_{ij}x_{0j} = (A\mathbf{x}_0)_i = \lambda_0 x_{0i}.$$

This proves 1.

Next suppose $A\mathbf{x} = \mu\mathbf{x}$ and $\mathbf{x} \neq \mathbf{0}$ and $\mu \neq \lambda_0$. Then $|A\mathbf{x}| = |\mu|\,|\mathbf{x}|$. But this implies $A|\mathbf{x}| \geq |\mu|\,|\mathbf{x}|$. (See the above abominable definition of $|\mathbf{x}|$.)

Case 1: $|\mathbf{x}| \neq \mathbf{x}$ and $|\mathbf{x}| \neq -\mathbf{x}$.

In this case, $A|\mathbf{x}| > |A\mathbf{x}| = |\mu|\,|\mathbf{x}|$ and letting $\mathbf{y} = A|\mathbf{x}|$, it follows $\mathbf{y} >> \mathbf{0}$ and $A\mathbf{y} >> |\mu|\,\mathbf{y}$ which shows $A\mathbf{y} >> (|\mu| + \varepsilon)\,\mathbf{y}$ for sufficiently small positive ε and verifies $|\mu| < \lambda_0$.

Case 2: $|\mathbf{x}| = \mathbf{x}$ or $|\mathbf{x}| = -\mathbf{x}$

In this case, the entries of $\mathbf{x}$ are all real and have the same sign. Therefore, $A|\mathbf{x}| = |A\mathbf{x}| = |\mu|\,|\mathbf{x}|$. Now let $\mathbf{y} \equiv |\mathbf{x}| / ||\mathbf{x}||_1$. Then $A\mathbf{y} = |\mu|\,\mathbf{y}$ and so $|\mu| \in S_1$ showing that $|\mu| \leq \lambda_0$. But also, the fact the entries of $\mathbf{x}$ all have the same sign shows $\mu = |\mu|$ and so $\mu \in S_1$. Since $\mu \neq \lambda_0$, it must be that $\mu = |\mu| < \lambda_0$. This proves 2.

It remains to verify 3. Suppose then that $A\mathbf{y} = \lambda_0\mathbf{y}$ and for all scalars, $\alpha, \alpha\mathbf{x}_0 \neq \mathbf{y}$. Then

$$A\,\mathrm{Re}\,\mathbf{y} = \lambda_0\,\mathrm{Re}\,\mathbf{y},\ A\,\mathrm{Im}\,\mathbf{y} = \lambda_0\,\mathrm{Im}\,\mathbf{y}.$$

If $\mathrm{Re}\,\mathbf{y} = \alpha_1\mathbf{x}_0$ and $\mathrm{Im}\,\mathbf{y} = \alpha_2\mathbf{x}_0$ for real numbers, α_i,then $\mathbf{y} = (\alpha_1 + i\alpha_2)\,\mathbf{x}_0$ and it is assumed this does not happen. Therefore, either

$$t\,\mathrm{Re}\,\mathbf{y} \neq \mathbf{x}_0 \text{ for all } t \in \mathbb{R}$$

or

$$t\,\mathrm{Im}\,\mathbf{y} \neq \mathbf{x}_0 \text{ for all } t \in \mathbb{R}.$$

Assume the first holds. Then varying $t \in \mathbb{R}$, there exists a value of t such that $\mathbf{x}_0 + t\,\mathrm{Re}\,\mathbf{y} > \mathbf{0}$ but it is not the case that $\mathbf{x}_0 + t\,\mathrm{Re}\,\mathbf{y} >> 0$. Then $A(\mathbf{x}_0 + t\,\mathrm{Re}\,\mathbf{y}) >> 0$ by Lemma 15.8.2. But this implies $\lambda_0(\mathbf{x}_0 + t\,\mathrm{Re}\,\mathbf{y}) >> 0$ which is a contradiction. Hence there exist real numbers, α_1 and α_2 such that $\mathrm{Re}\,\mathbf{y} = \alpha_1\mathbf{x}_0$ and $\mathrm{Im}\,\mathbf{y} = \alpha_2\mathbf{x}_0$ showing that $\mathbf{y} = (\alpha_1 + i\alpha_2)\,\mathbf{x}_0$. This proves 3.

It is possible to obtain a simple corollary to the above theorem.

Corollary 15.8.5 *If $A > 0$ and $A^m >> 0$ for some $m \in \mathbb{N}$, then all the conclusions of the above theorem hold.*

Proof: There exists $\mu_0 > 0$ such that $A^m\mathbf{y}_0 = \mu_0\mathbf{y}_0$ for $\mathbf{y}_0 >> 0$ by Theorem 15.8.4 and

$$\mu_0 = \sup\{\mu : A^m\mathbf{x} \geq \mu\mathbf{x} \text{ for some } \mathbf{x} \in K\}.$$

Let $\lambda_0^m = \mu_0$. Then

$$(A - \lambda_0 I)\left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right)\mathbf{y}_0 = (A^m - \lambda_0^m I)\,\mathbf{y}_0 = \mathbf{0}$$

and so letting $\mathbf{x}_0 \equiv \left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right)\mathbf{y}_0$, it follows $\mathbf{x}_0 >> 0$ and $A\mathbf{x}_0 = \lambda_0\mathbf{x}_0$.

Suppose now that $A\mathbf{x} = \mu\mathbf{x}$ for $\mathbf{x} \neq \mathbf{0}$ and $\mu \neq \lambda_0$. Suppose $|\mu| \geq \lambda_0$. Multiplying both sides by A, it follows $A^m\mathbf{x} = \mu^m\mathbf{x}$ and $|\mu^m| = |\mu|^m \geq \lambda_0^m = \mu_0$ and so from Theorem 15.8.4, since $|\mu^m| \geq \mu_0$, and μ^m is an eigenvalue of A^m, it follows that $\mu^m = \mu_0$. But by Theorem 15.8.4 again, this implies $\mathbf{x} = c\mathbf{y}_0$ for some scalar, c and hence $A\mathbf{y}_0 = \mu\mathbf{y}_0$. Since $\mathbf{y}_0 >> \mathbf{0}$, it follows $\mu \geq 0$ and so $\mu = \lambda_0$, a contradiction. Therefore, $|\mu| < \lambda_0$.

Finally, if $A\mathbf{x} = \lambda_0\mathbf{x}$, then $A^m\mathbf{x} = \lambda_0^m\mathbf{x}$ and so $\mathbf{x} = c\mathbf{y}_0$ for some scalar, c. Consequently,

$$\begin{aligned}\left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right)\mathbf{x} &= c\left(A^{m-1} + \lambda_0 A^{m-2} + \cdots + \lambda_0^{m-1} I\right)\mathbf{y}_0 \\ &= c\mathbf{x}_0.\end{aligned}$$

Hence

$$m\lambda_0^{m-1}\mathbf{x} = c\mathbf{x}_0$$

which shows the dimension of the eigenspace for λ_0 is one. This proves the corollary.

The following corollary is an extremely interesting convergence result involving the powers of positive matrices.

Corollary 15.8.6 *Let $A > 0$ and $A^m >> 0$ for some $m \in \mathbb{N}$. Then for λ_0 given in 15.47, there exists a rank one matrix, P such that $\lim_{m\to\infty} \left|\left|\left(\frac{A}{\lambda_0}\right)^m - P\right|\right| = 0$.*

Proof: Considering A^T, and the fact that A and A^T have the same eigenvalues, Corollary 15.8.5 implies the existence of a vector, $\mathbf{v} >> \mathbf{0}$ such that

$$A^T\mathbf{v} = \lambda_0\mathbf{v}.$$

Also let $\mathbf{x}_0$ denote the vector such that $A\mathbf{x}_0 = \lambda_0\mathbf{x}_0$ with $\mathbf{x}_0 >> \mathbf{0}$. First note that $\mathbf{x}_0^T\mathbf{v} > 0$ because both these vectors have all entries positive. Therefore, $\mathbf{v}$ may be scaled such that

$$\mathbf{v}^T\mathbf{x}_0 = \mathbf{x}_0^T\mathbf{v} = 1. \tag{15.48}$$

Define

$$P \equiv \mathbf{x}_0\mathbf{v}^T.$$

Thanks to 15.48,

$$\frac{A}{\lambda_0}P = \mathbf{x}_0\mathbf{v}^T = P,\ P\left(\frac{A}{\lambda_0}\right) = \mathbf{x}_0\mathbf{v}^T\left(\frac{A}{\lambda_0}\right) = \mathbf{x}_0\mathbf{v}^T = P, \tag{15.49}$$

and

$$P^2 = \mathbf{x}_0\mathbf{v}^T\mathbf{x}_0\mathbf{v}^T = \mathbf{v}^T\mathbf{x}_0 = P. \tag{15.50}$$

Therefore,

$$\begin{aligned}\left(\frac{A}{\lambda_0} - P\right)^2 &= \left(\frac{A}{\lambda_0}\right)^2 - 2\left(\frac{A}{\lambda_0}\right)P + P^2 \\ &= \left(\frac{A}{\lambda_0}\right)^2 - P.\end{aligned}$$

Continuing this way, using 15.49 repeatedly, it follows

$$\left(\left(\frac{A}{\lambda_0}\right) - P\right)^m = \left(\frac{A}{\lambda_0}\right)^m - P. \tag{15.51}$$

The eigenvalues of $\left(\frac{A}{\lambda_0}\right) - P$ are of interest because it is powers of this matrix which determine the convergence of $\left(\frac{A}{\lambda_0}\right)^m$ to P. Therefore, let μ be a nonzero eigenvalue of this matrix. Thus

$$\left(\left(\frac{A}{\lambda_0}\right) - P\right)\mathbf{x} = \mu\mathbf{x} \tag{15.52}$$

for $\mathbf{x} \neq \mathbf{0}$, and $\mu \neq 0$. Applying P to both sides and using the second formula of 15.49 yields

$$\mathbf{0} = (P - P)\,\mathbf{x} = \left(P\left(\frac{A}{\lambda_0}\right) - P^2\right)\mathbf{x} = \mu P\mathbf{x}.$$

But since $P\mathbf{x} = \mathbf{0}$, it follows from 15.52 that

$$A\mathbf{x} = \lambda_0 \mu \mathbf{x}$$

which implies $\lambda_0\mu$ is an eigenvalue of A. Therefore, by Corollary 15.8.5 it follows that either $\lambda_0\mu = \lambda_0$ in which case $\mu = 1$, or $\lambda_0 |\mu| < \lambda_0$ which implies $|\mu| < 1$. But if $\mu = 1$, then $\mathbf{x}$ is a multiple of $\mathbf{x}_0$ and 15.52 would yield

$$\left(\left(\frac{A}{\lambda_0}\right) - P\right)\mathbf{x}_0 = \mathbf{x}_0$$

which says $\mathbf{x}_0 - \mathbf{x}_0\mathbf{v}^T\mathbf{x}_0 = \mathbf{x}_0$ and so by 15.48, $\mathbf{x}_0 = \mathbf{0}$ contrary to the property that $\mathbf{x}_0 >> \mathbf{0}$. Therefore, $|\mu| < 1$ and so this has shown that the absolute values of all eigenvalues of $\left(\frac{A}{\lambda_0}\right) - P$ are less than 1. By Gelfand's theorem, Theorem 15.2.9, it follows

$$\left|\left|\left(\left(\frac{A}{\lambda_0}\right) - P\right)^m\right|\right|^{1/m} < r < 1$$

whenever m is large enough. Now by 15.51 this yields

$$\left|\left|\left(\frac{A}{\lambda_0}\right)^m - P\right|\right| = \left|\left|\left(\left(\frac{A}{\lambda_0}\right) - P\right)^m\right|\right| \leq r^m$$

whenever m is large enough. It follows

$$\lim_{m\to\infty} \left|\left|\left(\frac{A}{\lambda_0}\right)^m - P\right|\right| = 0$$

as claimed.

What about the case when $A > 0$ but maybe it is not the case that $A >> 0$? As before,

$$K \equiv \{\mathbf{x} \geq \mathbf{0} \text{ such that } ||\mathbf{x}||_1 = 1\}.$$

Now define

$$S_1 \equiv \{\lambda : A\mathbf{x} \geq \lambda\mathbf{x} \text{ for some } \mathbf{x} \in K\}$$

and

$$\lambda_0 \equiv \sup(S_1) \tag{15.53}$$

Theorem 15.8.7 *Let $A > 0$ and let λ_0 be defined in 15.53. Then there exists $\mathbf{x}_0 > \mathbf{0}$ such that $A\mathbf{x}_0 = \lambda_0\mathbf{x}_0$.*

Proof: Let E consist of the matrix which has a one in every entry. Then from Theorem 15.8.4 it follows there exists $\mathbf{x}_\delta >> \mathbf{0}$, $||\mathbf{x}_\delta||_1 = 1$, such that $(A + \delta E)\mathbf{x}_\delta = \lambda_{0\delta}\mathbf{x}_\delta$ where

$$\lambda_{0\delta} \equiv \sup\{\lambda : (A + \delta E)\mathbf{x} \geq \lambda\mathbf{x} \text{ for some } \mathbf{x} \in K\}.$$

Now if $\alpha < \delta$

$$\{\lambda : (A + \alpha E)\mathbf{x} \geq \boldsymbol{\lambda}\mathbf{x} \text{ for some } \mathbf{x} \in K\} \subseteq$$

$$\{\lambda : (A + \delta E)\mathbf{x} \geq \boldsymbol{\lambda}\mathbf{x} \text{ for some } \mathbf{x} \in K\}$$

and so $\lambda_{0\delta} \geq \lambda_{0\alpha}$ because $\lambda_{0\delta}$ is the sup of the second set and $\lambda_{0\alpha}$ is the sup of the first. It follows the limit, $\lambda_1 \equiv \lim_{\delta\to 0+} \lambda_{0\delta}$ exists. Taking a subsequence and using the compactness

of K, there exists a subsequence, still denoted by δ such that as $\delta \to 0$, $\mathbf{x}_\delta \to \mathbf{x} \in K$. Therefore,

$$A\mathbf{x} = \lambda_1 \mathbf{x}$$

and so, in particular, $A\mathbf{x} \geq \lambda_1 \mathbf{x}$ and so $\lambda_1 \leq \lambda_0$. But also, if $\lambda \leq \lambda_0$,

$$\lambda \mathbf{x} \leq A\mathbf{x} < (A + \delta E)\,\mathbf{x}$$

showing that $\lambda_{0\delta} \geq \lambda$ for all such λ. But then $\lambda_{0\delta} \geq \lambda_0$ also. Hence $\lambda_1 \geq \lambda_0$, showing these two numbers are the same. Hence $A\mathbf{x} = \lambda_0 \mathbf{x}$ and this proves the theorem.

If $A^m >> 0$ for some m and $A > 0$, it follows that the dimension of the eigenspace for λ_0 is one and that the absolute value of every other eigenvalue of A is less than λ_0. If it is only assumed that $A > 0$, not necessarily $>> 0$, this is no longer true. However, there is something which is very interesting which can be said. First here is an interesting lemma.

Lemma 15.8.8 *Let M be a matrix of the form*

$$M = \begin{pmatrix} A & 0 \\ B & C \end{pmatrix}$$

or

$$M = \begin{pmatrix} A & B \\ 0 & C \end{pmatrix}$$

where A is an $r \times r$ matrix and C is an $(n-r) \times (n-r)$ matrix. Then $\det(M) = \det(A)\det(B)$ and $\sigma(M) = \sigma(A) \cup \sigma(C)$.

Proof: To verify the claim about the determinants, note

$$\begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ B & C \end{pmatrix}$$

Therefore,

$$\det \begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \det \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} \det \begin{pmatrix} I & 0 \\ B & C \end{pmatrix}.$$

But it is clear from the method of Laplace expansion that

$$\det \begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix} = \det A$$

and from the multilinear properties of the determinant and row operations that

$$\det \begin{pmatrix} I & 0 \\ B & C \end{pmatrix} = \det \begin{pmatrix} I & 0 \\ 0 & C \end{pmatrix} = \det C.$$

The case where M is upper block triangular is similar.

This immediately implies $\sigma(M) = \sigma(A) \cup \sigma(C)$.

Theorem 15.8.9 *Let $A > 0$ and let λ_0 be given in 15.53. If λ is an eigenvalue for A such that $|\lambda| = \lambda_0$, then λ/λ_0 is a root of unity. Thus $(\lambda/\lambda_0)^m = 1$ for some $m \in \mathbb{N}$.*

Proof: Applying Theorem 15.8.7 to A^T, there exists $\mathbf{v} > \mathbf{0}$ such that $A^T\mathbf{v} = \lambda_0 \mathbf{v}$. In the first part of the argument it is assumed $\mathbf{v} >> \mathbf{0}$. Now suppose $A\mathbf{x} = \lambda\mathbf{x}, \mathbf{x} \neq \mathbf{0}$ and that $|\lambda| = \lambda_0$. Then

$$A|\mathbf{x}| \geq |\lambda|\,|\mathbf{x}| = \lambda_0 |\mathbf{x}|$$

and it follows that if $A|\mathbf{x}| > |\lambda||\mathbf{x}|$, then since $\mathbf{v} >> \mathbf{0}$,

$$\lambda_0(\mathbf{v},|\mathbf{x}|) < (\mathbf{v},A|\mathbf{x}|) = \left(A^T\mathbf{v},|\mathbf{x}|\right) = \lambda_0(\mathbf{v},|\mathbf{x}|),$$

a contradiction. Therefore,

$$A|\mathbf{x}| = \lambda_0|\mathbf{x}|. \tag{15.54}$$

It follows that

$$\left|\sum_j A_{ij}x_j\right| = \lambda_0|\mathbf{x}_i| = \sum_j A_{ij}|x_j|$$

and so the complex numbers,

$$A_{ij}x_j,\ A_{ik}x_k$$

must have the same argument for every k, j because equality holds in the triangle inequality. Therefore, there exists a complex number, μ_i such that

$$A_{ij}x_j = \mu_i A_{ij}|x_j| \tag{15.55}$$

and so, letting $r \in \mathbb{N}$,

$$A_{ij}x_j\mu_j^r = \mu_i A_{ij}|x_j|\mu_j^r.$$

Summing on j yields

$$\sum_j A_{ij}x_j\mu_j^r = \mu_i\sum_j A_{ij}|x_j|\mu_j^r. \tag{15.56}$$

Also, summing 15.55 on j and using that λ is an eigenvalue for $\mathbf{x}$, it follows from 15.54 that

$$\lambda x_i = \sum_j A_{ij}x_j = \mu_i\sum_j A_{ij}|x_j| = \mu_i\lambda_0|x_i|. \tag{15.57}$$

From 15.56 and 15.57,

$$\begin{aligned}
\sum_j A_{ij}x_j\mu_j^r &= \mu_i\sum_j A_{ij}|x_j|\mu_j^r \\
&= \mu_i\sum_j A_{ij}\overbrace{\mu_j|x_j|}^{\text{see 15.57}}\mu_j^{r-1} \\
&= \mu_i\sum_j A_{ij}\left(\frac{\lambda}{\lambda_0}\right)x_j\mu_j^{r-1} \\
&= \mu_i\left(\frac{\lambda}{\lambda_0}\right)\sum_j A_{ij}x_j\mu_j^{r-1}
\end{aligned}$$

Now from 15.56 with r replaced by $r-1$, this equals

$$\begin{aligned}
\mu_i^2\left(\frac{\lambda}{\lambda_0}\right)\sum_j A_{ij}|x_j|\mu_j^{r-1} &= \mu_i^2\left(\frac{\lambda}{\lambda_0}\right)\sum_j A_{ij}\mu_j|x_j|\mu_j^{r-2} \\
&= \mu_i^2\left(\frac{\lambda}{\lambda_0}\right)^2\sum_j A_{ij}x_j\mu_j^{r-2}.
\end{aligned}$$

Continuing this way,

$$\sum_j A_{ij}x_j\mu_j^r = \mu_i^k\left(\frac{\lambda}{\lambda_0}\right)^k\sum_j A_{ij}x_j\mu_j^{r-k}$$

and eventually, this shows

$$\begin{aligned}\sum_j A_{ij}x_j\mu_j^r &= \mu_i^r\left(\frac{\lambda}{\lambda_0}\right)^r\sum_j A_{ij}x_j \\ &= \left(\frac{\lambda}{\lambda_0}\right)^r\lambda\left(x_i\mu_i^r\right)\end{aligned}$$

and this says $\left(\frac{\lambda}{\lambda_0}\right)^{r+1}$ is an eigenvalue for $\left(\frac{A}{\lambda_0}\right)$ with the eigenvector being $(x_1\mu_1^r,\cdots,x_n\mu_n^r)^T$. Now recall that $r\in\mathbb{N}$ was arbitrary and so this has shown that $\left(\frac{\lambda}{\lambda_0}\right)^2,\left(\frac{\lambda}{\lambda_0}\right)^3,\left(\frac{\lambda}{\lambda_0}\right)^4,\cdots$ are each eigenvalues of $\left(\frac{A}{\lambda_0}\right)$ which has only finitely many and hence this sequence must repeat. Therefore, $\left(\frac{\lambda}{\lambda_0}\right)$ is a root of unity as claimed. This proves the theorem in the case that $\mathbf{v}>>\mathbf{0}$.

Now it is necessary to consider the case where $\mathbf{v}>\mathbf{0}$ but it is not the case that $\mathbf{v}>>\mathbf{0}$. Then in this case, there exists a permutation matrix, P such that

$$P\mathbf{v}=\begin{pmatrix}v_1\\ \vdots\\ v_r\\ 0\\ \vdots\\ 0\end{pmatrix}\equiv\begin{pmatrix}\mathbf{u}\\ \mathbf{0}\end{pmatrix}\equiv\mathbf{v}_1$$

Then

$$\lambda_0\mathbf{v}=A^T\mathbf{v}=A^TP\mathbf{v}_1.$$

Therefore,

$$\lambda_0\mathbf{v}_1=PA^TP\mathbf{v}_1=G\mathbf{v}_1$$

Now $P^2=I$ because it is a permuation matrix. Therefore, the matrix, $G\equiv PA^TP$ and A are similar. Consequently, they have the same eigenvalues and it suffices from now on to consider the matrix, G rather than A. Then

$$\lambda_0\begin{pmatrix}\mathbf{u}\\ \mathbf{0}\end{pmatrix}=\begin{pmatrix}M_1 & M_2\\ M_3 & M_4\end{pmatrix}\begin{pmatrix}\mathbf{u}\\ \mathbf{0}\end{pmatrix}$$

where M_1 is $r\times r$ and M_4 is $(n-r)\times(n-r)$. It follows from block multiplication and the assumption that A and hence G are >0 that

$$G=\begin{pmatrix}A' & B\\ 0 & C\end{pmatrix}.$$

Now let λ be an eigenvalue of G such that $|\lambda|=\lambda_0$. Then from Lemma 15.8.8, either $\lambda\in\sigma(A')$ or $\lambda\in\sigma(C)$. Suppose without loss of generality that $\lambda\in\sigma(A')$. Since $A'>0$ it has a largest positive eigenvalue, λ_0' which is obtained from 15.53. Thus $\lambda_0'\leq\lambda_0$ but λ being an eigenvalue of A', has its absolute value bounded by λ_0' and so $\lambda_0=|\lambda|\leq\lambda_0'\leq\lambda_0$ showing that $\lambda_0\in\sigma(A')$. Now if there exists $\mathbf{v}>>\mathbf{0}$ such that $A'^T\mathbf{v}=\lambda_0\mathbf{v}$, then the first part of this proof applies to the matrix, A and so (λ/λ_0) is a root of unity. If such a vector, $\mathbf{v}$ does not exist, then let A' play the role of A in the above argument and reduce to the consideration of

$$G'\equiv\begin{pmatrix}A'' & B'\\ 0 & C'\end{pmatrix}$$

where G' is similar to A' and $\lambda, \lambda_0 \in \sigma(A'')$. Stop if $A''^T \mathbf{v} = \lambda_0 \mathbf{v}$ for some $\mathbf{v} >> \mathbf{0}$. Otherwise, decompose A'' similar to the above and add another prime. Continuing this way you must eventually obtain the situation where $(A'^{\cdots\prime})^T \mathbf{v} = \lambda_0 \mathbf{v}$ for some $\mathbf{v} >> \mathbf{0}$. Indeed, this happens no later than when $A'^{\cdots\prime}$ is a 1×1 matrix. This proves the theorem.

15.9 Functions Of Matrices

The existence of the Jordan form also makes it possible to define various functions of matrices. Suppose

$$f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n \tag{15.58}$$

for all $|\lambda| < R$. There is a formula for $f(A) \equiv \sum_{n=0}^{\infty} a_n A^n$ which makes sense whenever $\rho(A) < R$. Thus you can speak of $\sin(A)$ or e^A for A an $n \times n$ matrix. To begin with, define

$$f_P(\lambda) \equiv \sum_{n=0}^{P} a_n \lambda^n$$

so for $k < P$

$$\begin{aligned} f_P^{(k)}(\lambda) &= \sum_{n=k}^{P} a_n n \cdots (n-k+1) \lambda^{n-k} \\ &= \sum_{n=k}^{P} a_n \binom{n}{k} k! \lambda^{n-k}. \end{aligned} \tag{15.59}$$

To begin with consider $f(J_m(\lambda))$ where $J_m(\lambda)$ is an $m \times m$ Jordan block. Thus $J_m(\lambda) = D + N$ where $N^m = 0$ and N commutes with D. Therefore, letting $P > m$

$$\begin{aligned} \sum_{n=0}^{P} a_n J_m(\lambda)^n &= \sum_{n=0}^{P} a_n \sum_{k=0}^{n} \binom{n}{k} D^{n-k} N^k \\ &= \sum_{k=0}^{P} \sum_{n=k}^{P} a_n \binom{n}{k} D^{n-k} N^k \\ &= \sum_{k=0}^{m-1} N^k \sum_{n=k}^{P} \binom{n}{k} D^{n-k}. \end{aligned} \tag{15.60}$$

Now for $k = 0, \cdots, m-1$, define $\mathrm{diag}_k(a_1, \cdots, a_{m-k})$ the $m \times m$ matrix which equals zero everywhere except on the k^{th} super diagonal where this diagonal is filled with the numbers, $\{a_1, \cdots, a_{m-k}\}$ from the upper left to the lower right. Thus in 4×4 matrices, $\mathrm{diag}_2(1,2)$ would be the matrix,

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

Then from 15.60 and 15.59,

$$\sum_{n=0}^{P} a_n J_m(\lambda)^n = \sum_{k=0}^{m-1} \mathrm{diag}_k \left(\frac{f_P^{(k)}(\lambda)}{k!}, \cdots, \frac{f_P^{(k)}(\lambda)}{k!} \right).$$

Therefore, $\sum_{n=0}^{P} a_n J_m(\lambda)^n =$

$$\begin{pmatrix} f_P(\lambda) & \frac{f_P'(\lambda)}{1!} & \frac{f_P^{(2)}(\lambda)}{2!} & \cdots & \frac{f_P^{(m-1)}(\lambda)}{(m-1)!} \\ & f_P(\lambda) & \frac{f_P'(\lambda)}{1!} & \ddots & \vdots \\ & & f_P(\lambda) & \ddots & \frac{f_P^{(2)}(\lambda)}{2!} \\ & & & \ddots & \frac{f_P'(\lambda)}{1!} \\ 0 & & & & f_P(\lambda) \end{pmatrix} \tag{15.61}$$

Now let A be an $n \times n$ matrix with $\rho(A) < R$ where R is given above. Then the Jordan form of A is of the form

$$J = \begin{pmatrix} J_1 & & 0 \\ & J_2 & \\ & & \ddots \\ 0 & & J_r \end{pmatrix} \tag{15.62}$$

where $J_k = J_{m_k}(\lambda_k)$ is an $m_k \times m_k$ Jordan block and $A = S^{-1}JS$. Then, letting $P > m_k$ for all k,

$$\sum_{n=0}^{P} a_n A^n = S^{-1} \sum_{n=0}^{P} a_n J^n S,$$

and because of block multiplication of matrices,

$$\sum_{n=0}^{P} a_n J^n = \begin{pmatrix} \sum_{n=0}^{P} a_n J_1^n & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & \sum_{n=0}^{P} a_n J_r^n \end{pmatrix}$$

and from 15.61 $\sum_{n=0}^{P} a_n J_k^n$ converges as $P \to \infty$ to the $m_k \times m_k$ matrix,

$$\begin{pmatrix} f(\lambda_k) & \frac{f'(\lambda_k)}{1!} & \frac{f^{(2)}(\lambda_k)}{2!} & \cdots & \frac{f^{(m-1)}(\lambda_k)}{(m_k-1)!} \\ 0 & f(\lambda_k) & \frac{f'(\lambda_k)}{1!} & \ddots & \vdots \\ 0 & 0 & f(\lambda_k) & \ddots & \frac{f^{(2)}(\lambda_k)}{2!} \\ \vdots & & \ddots & \ddots & \frac{f'(\lambda_k)}{1!} \\ 0 & 0 & \cdots & 0 & f(\lambda_k) \end{pmatrix} \tag{15.63}$$

There is no convergence problem because $|\lambda| < R$ for all $\lambda \in \sigma(A)$. This has proved the following theorem.

Theorem 15.9.1 *Let f be given by 15.58 and suppose $\rho(A) < R$ where R is the radius of convergence of the power series in 15.58. Then the series,*

$$\sum_{k=0}^{\infty} a_n A^n \tag{15.64}$$

converges in the space $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ with respect to any of the norms on this space and furthermore,

$$\sum_{k=0}^{\infty} a_n A^n = S^{-1} \begin{pmatrix} \sum_{n=0}^{\infty} a_n J_1^n & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & \sum_{n=0}^{\infty} a_n J_r^n \end{pmatrix} S$$

where $\sum_{n=0}^{\infty} a_n J_k^n$ is an $m_k \times m_k$ matrix of the form given in 15.63 where $A = S^{-1}JS$ and the Jordan form of A, J is given by 15.62. Therefore, you can define $f(A)$ by the series in 15.64.

Here is a simple example.

Example 15.9.2 *Find* $\sin(A)$ *where* $A = \begin{pmatrix} 4 & 1 & -1 & 1 \\ 1 & 1 & 0 & -1 \\ 0 & -1 & 1 & -1 \\ -1 & 2 & 1 & 4 \end{pmatrix}$.

In this case, the Jordan canonical form of the matrix is not too hard to find.

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

$$\begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{8} & -\frac{3}{8} & 0 & -\frac{1}{8} \\ 0 & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Then from the above theorem $\sin(J)$ is given by

$$\sin\begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} \sin 4 & 0 & 0 & 0 \\ 0 & \sin 2 & \cos 2 & \frac{-\sin 2}{2} \\ 0 & 0 & \sin 2 & \cos 2 \\ 0 & 0 & 0 & \sin 2 \end{pmatrix}.$$

Therefore, $\sin(A) =$

$$\begin{pmatrix} 2 & 0 & -2 & -1 \\ 1 & -4 & -2 & -1 \\ 0 & 0 & -2 & 1 \\ -1 & 4 & 4 & 2 \end{pmatrix} \begin{pmatrix} \sin 4 & 0 & 0 & 0 \\ 0 & \sin 2 & \cos 2 & \frac{-\sin 2}{2} \\ 0 & 0 & \sin 2 & \cos 2 \\ 0 & 0 & 0 & \sin 2 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{8} & -\frac{3}{8} & 0 & -\frac{1}{8} \\ 0 & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} =$$

$$\begin{pmatrix} \sin 4 & \sin 4 - \sin 2 - \cos 2 & -\cos 2 & \sin 4 - \sin 2 - \cos 2 \\ \frac{1}{2}\sin 4 - \frac{1}{2}\sin 2 & \frac{1}{2}\sin 4 + \frac{3}{2}\sin 2 - 2\cos 2 & \sin 2 & \frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 - 2\cos 2 \\ 0 & -\cos 2 & \sin 2 - \cos 2 & -\cos 2 \\ -\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 & -\frac{1}{2}\sin 4 - \frac{1}{2}\sin 2 + 3\cos 2 & \cos 2 - \sin 2 & -\frac{1}{2}\sin 4 + \frac{1}{2}\sin 2 + 3\cos 2 \end{pmatrix}.$$

Perhaps this isn't the first thing you would think of. Of course the ability to get this nice closed form description of $\sin(A)$ was dependent on being able to find the Jordan form along with a similarity transformation which will yield the Jordan form.

The following corollary is known as the spectral mapping theorem.

Corollary 15.9.3 *Let A be an $n \times n$ matrix and let $\rho(A) < R$ where for $|\lambda| < R$,*

$$f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n.$$

Then $f(A)$ is also an $n \times n$ matrix and furthermore, $\sigma(f(A)) = f(\sigma(A))$. Thus the eigenvalues of $f(A)$ are exactly the numbers $f(\lambda)$ where λ is an eigenvalue of A. Furthermore, the algebraic multiplicity of $f(\lambda)$ coincides with the algebraic multiplicity of λ.

All of these things can be generalized to linear transformations defined on infinite dimensional spaces and when this is done the main tool is the Dunford integral along with the methods of complex analysis. It is good to see it done for finite dimensional situations first because it gives an idea of what is possible. Actually, some of the most interesting functions in applications do not come in the above form as a power series expanded about 0. One example of this situation has already been encountered in the proof of the right polar decomposition with the square root of an Hermitian transformation which had all nonnegative eigenvalues. Another example is that of taking the positive part of an Hermitian matrix. This is important in some physical models where something may depend on the positive part of the strain which is a symmetric real matrix. Obviously there is no way to consider this as a power series expanded about 0 because the function $f(r) = r^+$ is not even differentiable at 0. Therefore, a totally different approach must be considered. First the notion of a positive part is defined.

Definition 15.9.4 *Let A be an Hermitian matrix. Thus it suffices to consider A as an element of $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ according to the usual notion of matrix multiplication. Then there exists an orthonormal basis of eigenvectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ such that*

$$A = \sum_{j=1}^{n} \lambda_j \mathbf{u}_j \otimes \mathbf{u}_j,$$

for λ_j the eigenvalues of A, all real. Define

$$A^+ \equiv \sum_{j=1}^{n} \lambda_j^+ \mathbf{u}_j \otimes \mathbf{u}_j$$

where $\lambda^+ \equiv \frac{|\lambda| + \lambda}{2}$.

This gives us a nice definition of what is meant but it turns out to be very important in the applications to determine how this function depends on the choice of symmetric matrix, A. The following addresses this question.

Theorem 15.9.5 *If A, B be Hermitian matrices, then for $|\cdot|$ the Frobenius norm,*

$$\left|A^+ - B^+\right| \leq |A - B|.$$

Proof: Let $A = \sum_i \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i$ and let $B = \sum_j \mu_j \mathbf{w}_j \otimes \mathbf{w}_j$ where $\{\mathbf{v}_i\}$ and $\{\mathbf{w}_j\}$ are orthonormal bases of eigenvectors.

$$\left|A^+ - B^+\right|^2 = trace\left(\sum_i \lambda_i^+ \mathbf{v}_i \otimes \mathbf{v}_i - \sum_j \mu_j^+ \mathbf{w}_j \otimes \mathbf{w}_j\right)^2 =$$

$$trace\left[\sum_i \left(\lambda_i^+\right)^2 \mathbf{v}_i \otimes \mathbf{v}_i + \sum_j \left(\mu_j^+\right)^2 \mathbf{w}_j \otimes \mathbf{w}_j\right.$$

$$\left. - \sum_{i,j} \lambda_i^+ \mu_j^+ \left(\mathbf{w}_j, \mathbf{v}_i\right) \mathbf{v}_i \otimes \mathbf{w}_j - \sum_{i,j} \lambda_i^+ \mu_j^+ \left(\mathbf{v}_i, \mathbf{w}_j\right) \mathbf{w}_j \otimes \mathbf{v}_i\right]$$

Since the trace of $\mathbf{v}_i \otimes \mathbf{w}_j$ is $\left(\mathbf{v}_i, \mathbf{w}_j\right)$, a fact which follows from $\left(\mathbf{v}_i, \mathbf{w}_j\right)$ being the only possibly nonzero eigenvalue,

$$= \sum_i \left(\lambda_i^+\right)^2 + \sum_j \left(\mu_j^+\right)^2 - 2\sum_{i,j} \lambda_i^+ \mu_j^+ \left|\left(\mathbf{v}_i, \mathbf{w}_j\right)\right|^2. \tag{15.65}$$

Since these are orthonormal bases,

$$\sum_i |(\mathbf{v}_i, \mathbf{w}_j)|^2 = 1 = \sum_j |(\mathbf{v}_i, \mathbf{w}_j)|^2$$

and so 15.65 equals

$$= \sum_i \sum_j \left(\left(\lambda_i^+\right)^2 + \left(\mu_j^+\right)^2 - 2\lambda_i^+ \mu_j^+ \right) |(\mathbf{v}_i, \mathbf{w}_j)|^2 .$$

Similarly,

$$|A - B|^2 = \sum_i \sum_j \left((\lambda_i)^2 + \left(\mu_j\right)^2 - 2\lambda_i \mu_j \right) |(\mathbf{v}_i, \mathbf{w}_j)|^2 .$$

Now it is easy to check that $(\lambda_i)^2 + \left(\mu_j\right)^2 - 2\lambda_i\mu_j \geq \left(\lambda_i^+\right)^2 + \left(\mu_j^+\right)^2 - 2\lambda_i^+\mu_j^+$ and so this proves the theorem.

Applications To Differential Equations

16.1 Theory Of Ordinary Differntial Equations

Here I will present fundamental existence and uniqueness theorems for initial value problems for the differential equation,

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x}).$$

Suppose that $\mathbf{f} : [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$ satisfies the following two conditions.

$$|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{x}_1)| \leq K |\mathbf{x} - \mathbf{x}_1|, \tag{16.1}$$

$$\mathbf{f} \text{ is continuous.} \tag{16.2}$$

The first of these conditions is known as a Lipschitz condition.

Lemma 16.1.1 *Suppose $\mathbf{x} : [a,b] \to \mathbb{R}^n$ is a continuous function and $c \in [a,b]$. Then $\mathbf{x}$ is a solution to the initial value problem,*

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x}), \ \mathbf{x}(c) = \mathbf{x}_0 \tag{16.3}$$

if and only if $\mathbf{x}$ is a solution to the integral equation,

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_c^t \mathbf{f}(s, \mathbf{x}(s))\, ds. \tag{16.4}$$

Proof: If $\mathbf{x}$ solves 16.4, then since $\mathbf{f}$ is continuous, we may apply the fundamental theorem of calculus to differentiate both sides and obtain $\mathbf{x}'(t) = \mathbf{f}(t, \mathbf{x}(t))$. Also, letting $t = c$ on both sides, gives $\mathbf{x}(c) = \mathbf{x}_0$. Conversely, if $\mathbf{x}$ is a solution of the initial value problem, we may integrate both sides from c to t to see that $\mathbf{x}$ solves 16.4. This proves the lemma.

Theorem 16.1.2 *Let $\mathbf{f}$ satisfy 16.1 and 16.2. Then there exists a unique solution to the initial value problem, 16.3 on the interval $[a,b]$.*

Proof: Let $||\mathbf{x}||_\lambda \equiv \sup\left\{e^{\lambda t} |\mathbf{x}(t)| : t \in [a,b]\right\}$. Then this norm is equivalent to the usual norm on $BC([a,b], \mathbb{F}^n)$ described in Example 15.4.2. This means that for $||\cdot||$ the norm given there, there exists constants δ and Δ such that

$$||\mathbf{x}||_\lambda \delta \leq ||\mathbf{x}|| \leq \Delta ||\mathbf{x}||$$

for all $\mathbf{x} \in BC([a,b],\mathbb{F}^n)$. In fact, you can take $\delta \equiv e^{\lambda a}$ and $\Delta \equiv e^{\lambda b}$ in case $\lambda > 0$ with the two reversed in case $\lambda < 0$. Thus $BC([a,b],\mathbb{F}^n)$ is a Banach space with this norm, $||\cdot||_\lambda$. Then let $F : BC([a,b],\mathbb{F}^n) \to BC([a,b],\mathbb{F}^n)$ be defined by

$$F\mathbf{x}(t) \equiv \mathbf{x}_0 + \int_c^t \mathbf{f}(s,\mathbf{x}(s))\,ds.$$

Let $\lambda < 0$. It follows

$$\begin{aligned} e^{\lambda t}|F\mathbf{x}(t) - F\mathbf{y}(t)| &\leq \left| e^{\lambda t}\int_c^t |\mathbf{f}(s,\mathbf{x}(s)) - \mathbf{f}(s,\mathbf{y}(s))|\,ds \right| \\ &\leq \left| \int_c^t Ke^{\lambda(t-s)}|\mathbf{x}(s) - \mathbf{y}(s)|\,e^{\lambda s}ds \right| \\ &\leq ||\mathbf{x}-\mathbf{y}||_\lambda \int_a^t Ke^{\lambda(t-s)}ds \\ &\leq ||\mathbf{x}-\mathbf{y}||_\lambda \frac{K}{|\lambda|} \end{aligned}$$

and therefore,

$$||F\mathbf{x} - F\mathbf{y}||_\lambda \leq ||\mathbf{x}-\mathbf{y}||\frac{K}{|\lambda|}.$$

If $|\lambda|$ is chosen larger than K, this implies F is a contraction mapping on $BC([a,b],\mathbb{F}^n)$. Therefore, there exists a unique fixed point. With Lemma 16.1.1 this proves the theorem.

16.2 Linear Systems

As an example of the above theorem, consider for $t \in [a,b]$ the system

$$\mathbf{x}' = A(t)\mathbf{x}(t) + \mathbf{g}(t),\ \mathbf{x}(c) = \mathbf{x}_0 \tag{16.5}$$

where $A(t)$ is an $n \times n$ matrix whose entries are continuous functions of $t, (a_{ij}(t))$ and $\mathbf{g}(t)$ is a vector whose components are continuous functions of t satisfies the conditions of Theorem 16.1.2 with $\mathbf{f}(t,\mathbf{x}) = A(t)\mathbf{x} + \mathbf{g}(t)$. To see this, let $\mathbf{x} = (x_1,\cdots,x_n)^T$ and $\mathbf{x}_1 = (x_{11},\cdots x_{1n})^T$. Then letting $M = \max\{|a_{ij}(t)| : t \in [a,b], i,j \leq n\}$,

$$\begin{aligned} |\mathbf{f}(t,\mathbf{x}) - \mathbf{f}(t,\mathbf{x}_1)| &= |A(t)(\mathbf{x}-\mathbf{x}_1)| \\ &= \left| \left(\sum_{i=1}^n \left| \sum_{j=1}^n a_{ij}(t)(x_j - x_{1j}) \right|^2 \right)^{1/2} \right| \\ &\leq M \left| \left(\sum_{i=1}^n \left(\sum_{j=1}^n |x_j - x_{1j}| \right)^2 \right)^{1/2} \right| \\ &\leq M \left| \left(\sum_{i=1}^n n \sum_{j=1}^n |x_j - x_{1j}|^2 \right)^{1/2} \right| \\ &= Mn \left(\sum_{j=1}^n |x_j - x_{1j}|^2 \right)^{1/2} = Mn|\mathbf{x}-\mathbf{x}_1|. \end{aligned}$$

Therefore, let $K = Mn$. This proves

Theorem 16.2.1 *Let $A(t)$ be a continuous $n \times n$ matrix and let $\mathbf{g}(t)$ be a continuous vector for $t \in [a,b]$ and let $c \in [a,b]$ and $\mathbf{x}_0 \in \mathbb{F}^n$. Then there exists a unique solution to 16.5 valid for $t \in [a,b]$.*

This includes more examples of linear equations than are typically encountered in an entire differential equations course.

16.3 Local Solutions

Lemma 16.3.1 *Let $D(\mathbf{x}_0, r) \equiv \{\mathbf{x} \in \mathbb{F}^n : |\mathbf{x} - \mathbf{x}_0| \leq r\}$ and suppose U is an open set containing $D(\mathbf{x}_0, r)$ such that $\mathbf{f} : U \to \mathbb{F}^n$ is $C^1(U)$. (Recall this means all partial derivatives of $\mathbf{f}$ exist and are continuous.) Then for $K = Mn$, where M denotes the maximum of $\left|\frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{z})\right|$ for $\mathbf{z} \in D(\mathbf{x}_0, r)$, it follows that for all $\mathbf{x}, \mathbf{y} \in D(\mathbf{x}_0, r)$,*

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq K |\mathbf{x} - \mathbf{y}|.$$

Proof: Let $\mathbf{x}, \mathbf{y} \in D(\mathbf{x}_0, r)$ and consider the line segment joining these two points, $\mathbf{x} + t(\mathbf{y} - \mathbf{x})$ for $t \in [0,1]$. Letting $\mathbf{h}(t) = \mathbf{f}(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$ for $t \in [0,1]$, then

$$\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x}) = \mathbf{h}(1) - \mathbf{h}(0) = \int_0^1 \mathbf{h}'(t)\, dt.$$

Also, by the chain rule,

$$\mathbf{h}'(t) = \sum_{i=1}^{n} \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(y_i - x_i).$$

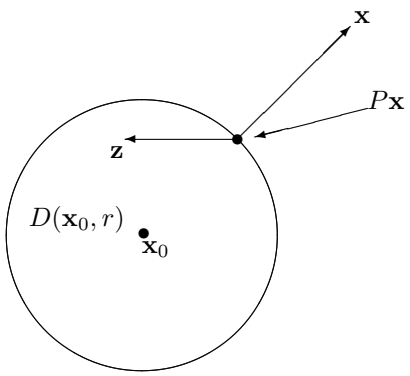
Therefore,

$$|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})| =$$

$$\begin{aligned} & \left| \int_0^1 \sum_{i=1}^{n} \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(y_i - x_i)\, dt \right| \\ \leq & \int_0^1 \sum_{i=1}^{n} \left| \frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) \right| |y_i - x_i|\, dt \\ \leq & M \sum_{i=1}^{n} |y_i - x_i| \leq Mn |\mathbf{x} - \mathbf{y}|. \end{aligned}$$

This proves the lemma.

Now consider the map, P which maps all of $\mathbb{R}^n$ to $D(\mathbf{x}_0, r)$ given as follows. For $\mathbf{x} \in D(\mathbf{x}_0, r)$, $P\mathbf{x} = \mathbf{x}$. For $\mathbf{x} \notin D(\mathbf{x}_0, r)$, $P\mathbf{x}$ will be the closest point in $D(\mathbf{x}_0, r)$ to $\mathbf{x}$. Such a closest point exists because $D(\mathbf{x}_0, r)$ is a closed and bounded set. Taking $f(\mathbf{y}) \equiv |\mathbf{y} - \mathbf{x}|$, it follows f is a continuous function defined on $D(\mathbf{x}_0, r)$ which must achieve its minimum value by the extreme value theorem from calculus.



Lemma 16.3.2 *For any pair of points,* $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$, $|P\mathbf{x} - P\mathbf{y}| \leq |\mathbf{x} - \mathbf{y}|$.

Proof: The above picture suggests the geometry of what is going on. Letting $\mathbf{z} \in D(\mathbf{x}_0, r)$, it follows that for all $t \in [0,1]$,

$$|\mathbf{x} - P\mathbf{x}|^2 \leq |\mathbf{x} - (P\mathbf{x} + t(\mathbf{z} - P\mathbf{x}))|^2$$

$$= |\mathbf{x} - P\mathbf{x}|^2 + 2t \operatorname{Re}((\mathbf{x} - P\mathbf{x}) \cdot (P\mathbf{x} - \mathbf{z})) + t^2 |\mathbf{z} - P\mathbf{x}|^2$$

Hence

$$2t \operatorname{Re}((\mathbf{x} - P\mathbf{x}) \cdot (P\mathbf{x} - \mathbf{z})) + t^2 |\mathbf{z} - P\mathbf{x}|^2 \geq 0$$

and this can only happen if

$$\operatorname{Re}((\mathbf{x} - P\mathbf{x}) \cdot (P\mathbf{x} - \mathbf{z})) \geq 0.$$

Therefore,

$$\begin{aligned} \operatorname{Re}((\mathbf{x} - P\mathbf{x}) \cdot (P\mathbf{x} - P\mathbf{y})) &\geq 0 \\ \operatorname{Re}((\mathbf{y} - P\mathbf{y}) \cdot (P\mathbf{y} - P\mathbf{x})) &\geq 0 \end{aligned}$$

and so

$$\operatorname{Re}(\mathbf{x} - P\mathbf{x} - (\mathbf{y} - P\mathbf{y})) \cdot (P\mathbf{x} - P\mathbf{y}) \geq 0$$

which implies

$$\operatorname{Re}(\mathbf{x} - \mathbf{y}) \cdot (P\mathbf{x} - P\mathbf{y}) \geq |P\mathbf{x} - P\mathbf{y}|^2$$

Then using the Cauchy Schwarz inequality it follows

$$|\mathbf{x} - \mathbf{y}| \geq |P\mathbf{x} - P\mathbf{y}|.$$

This proves the lemma.

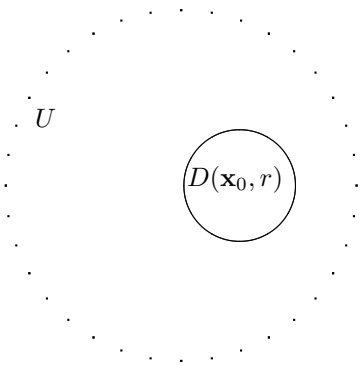
With this here is the local existence and uniqueness theorem.

Theorem 16.3.3 *Let* $[a,b]$ *be a closed interval and let* U *be an open subset of* $\mathbb{F}^n$. *Let* $\mathbf{f} : [a,b] \times U \to \mathbb{F}^n$ *be continuous and suppose that for each* $t \in [a,b]$, *the map* $\mathbf{x} \to \frac{\partial \mathbf{f}}{\partial x_i}(t, \mathbf{x})$ *is continuous. Also let* $\mathbf{x}_0 \in U$ *and* $c \in [a,b]$. *Then there exists an interval,* $I \subseteq [a,b]$ *such that* $c \in I$ *and there exists a unique solution to the initial value problem,*

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x}),\ \mathbf{x}(c) = \mathbf{x}_0 \tag{16.6}$$

valid for $t \in I$.

Proof: Consider the following picture.



The large dotted circle represents U and the little solid circle represents $D(\mathbf{x}_0, r)$ as indicated. Here r is so small that $D(\mathbf{x}_0, r)$ is contained in U as shown. Now let P denote the projection map defined above. Consider the initial value problem

$$\mathbf{x}' = \mathbf{f}(t, P\mathbf{x}),\ \mathbf{x}(c) = \mathbf{x}_0. \tag{16.7}$$

From Lemma 16.3.1 and the continuity of $\mathbf{x} \to \frac{\partial \mathbf{f}}{\partial x_i}(t, \mathbf{x})$, there exists a constant, K such that if $\mathbf{x}, \mathbf{y} \in D(\mathbf{x}_0, r)$, then $|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})| \leq K |\mathbf{x} - \mathbf{y}|$ for all $t \in [a, b]$. Therefore, by Lemma 16.3.2

$$|\mathbf{f}(t, P\mathbf{x}) - \mathbf{f}(t, P\mathbf{y})| \leq K |P\mathbf{x} - P\mathbf{y}| \leq K |\mathbf{x} - \mathbf{y}|.$$

It follows from Theorem 16.1.2 that 16.7 has a unique solution valid for $t \in [a, b]$. Since $\mathbf{x}$ is continuous, it follows that there exists an interval, I containing c such that for $t \in I$, $\mathbf{x}(t) \in D(\mathbf{x}_0, r)$. Therefore, for these values of t, $\mathbf{f}(t, P\mathbf{x}) = \mathbf{f}(t, \mathbf{x})$ and so there is a unique solution to 16.6 on I. This proves the theorem.

Now suppose $\mathbf{f}$ has the property that for every $R > 0$ there exists a constant, K_R such that for all $\mathbf{x}, \mathbf{x}_1 \in \overline{B(0, R)}$,

$$|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{x}_1)| \leq K_R |\mathbf{x} - \mathbf{x}_1|. \tag{16.8}$$

Corollary 16.3.4 *Let $\mathbf{f}$ satisfy 16.8 and suppose also that $(t, \mathbf{x}) \to \mathbf{f}(t, \mathbf{x})$ is continuous. Suppose now that $\mathbf{x}_0$ is given and there exists an estimate of the form $|\mathbf{x}(t)| < R$ for all $t \in [0, T)$ where $T \leq \infty$ on the local solution to*

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x}),\ \mathbf{x}(0) = \mathbf{x}_0. \tag{16.9}$$

Then there exists a unique solution to the initial value problem, 16.9 valid on $[0, T)$.

Proof: Replace $\mathbf{f}(t, \mathbf{x})$ with $\mathbf{f}(t, P\mathbf{x})$ where P is the projection onto $\overline{B(0, R)}$. Then by Theorem 16.1.2 there exists a unique solution to the system

$$\mathbf{x}' = \mathbf{f}(t, P\mathbf{x}),\ \mathbf{x}(0) = \mathbf{x}_0$$

valid on $[0, T_1]$ for every $T_1 < T$. Therefore, the above system has a unique solution on $[0, T)$ and from the estimate, $P\mathbf{x} = \mathbf{x}$. This proves the corollary.

16.4 First Order Linear Systems

Here is a discussion of linear systems of the form

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}(t)$$

where A is a constant $n \times n$ matrix and $\mathbf{f}$ is a vector valued function having all entries continuous. Of course the existence theory is a very special case of the general considerations above but I will give a self contained presentation based on elementary first order scalar differential equations and linear algebra.

Definition 16.4.1 *Suppose $t \to M(t)$ is a matrix valued function of t. Thus $M(t) = (m_{ij}(t))$. Then define*

$$M'(t) \equiv \left(m'_{ij}(t)\right).$$

In words, the derivative of $M(t)$ is the matrix whose entries consist of the derivatives of the entries of $M(t)$. Integrals of matrices are defined the same way. Thus

$$\int_a^b M(t)\, di \equiv \left(\int_a^b m_{ij}(t)\, dt\right).$$

In words, the integral of $M(t)$ is the matrix obtained by replacing each entry of $M(t)$ by the integral of that entry.

With this definition, it is easy to prove the following theorem.

Theorem 16.4.2 *Suppose $M(t)$ and $N(t)$ are matrices for which $M(t)N(t)$ makes sense. Then if $M'(t)$ and $N'(t)$ both exist, it follows that*

$$(M(t)N(t))' = M'(t)N(t) + M(t)N'(t).$$

Proof:

$$\begin{aligned}
\left((M(t)N(t))'\right)_{ij} &\equiv \left((M(t)N(t))_{ij}\right)' \\
&= \left(\sum_k M(t)_{ik} N(t)_{kj}\right)' \\
&= \sum_k (M(t)_{ik})' N(t)_{kj} + M(t)_{ik}\left(N(t)_{kj}\right)' \\
&\equiv \sum_k \left(M(t)'\right)_{ik} N(t)_{kj} + M(t)_{ik}\left(N(t)'\right)_{kj} \\
&\equiv \left(M'(t)N(t) + M(t)N'(t)\right)_{ij}
\end{aligned}$$

and this proves the theorem.

In the study of differential equations, one of the most important theorems is Gronwall's inequality which is next.

Theorem 16.4.3 *Suppose $u(t) \geq 0$ and for all $t \in [0,T]$,*

$$u(t) \leq u_0 + \int_0^t Ku(s)\, ds. \tag{16.10}$$

where K is some constant. Then

$$u(t) \leq u_0 e^{Kt}. \tag{16.11}$$

Proof: Let $w(t) = \int_0^t u(s)\,ds$. Then using the fundamental theorem of calculus, 16.10 $w(t)$ satisfies the following.

$$u(t) - Kw(t) = w'(t) - Kw(t) \leq u_0,\ w(0) = 0. \tag{16.12}$$

Multiply both sides of this inequality by e^{-Kt} and using the product rule and the chain rule,

$$e^{-Kt}(w'(t) - Kw(t)) = \frac{d}{dt}\left(e^{-Kt}w(t)\right) \leq u_0 e^{-Kt}.$$

Integrating this from 0 to t,

$$e^{-Kt}w(t) \leq u_0 \int_0^t e^{-Ks}ds = u_0\left(-\frac{e^{-tK}-1}{K}\right).$$

Now multiply through by e^{Kt} to obtain

$$w(t) \leq u_0\left(-\frac{e^{-tK}-1}{K}\right)e^{Kt} = -\frac{u_0}{K} + \frac{u_0}{K}e^{tK}.$$

Therefore, 16.12 implies

$$u(t) \leq u_0 + K\left(-\frac{u_0}{K} + \frac{u_0}{K}e^{tK}\right) = u_0 e^{Kt}.$$

This proves the theorem.

With Gronwall's inequality, here is a theorem on uniqueness of solutions to the initial value problem,

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}(t),\ \mathbf{x}(a) = \mathbf{x}_a, \tag{16.13}$$

in which A is an $n \times n$ matrix and $\mathbf{f}$ is a continuous function having values in $\mathbb{C}^n$.

Theorem 16.4.4 *Suppose* $\mathbf{x}$ *and* $\mathbf{y}$ *satisfy 16.13. Then* $\mathbf{x}(t) = \mathbf{y}(t)$ *for all* t.

Proof: Let $\mathbf{z}(t) = \mathbf{x}(t+a) - \mathbf{y}(t+a)$. Then for $t \geq 0$,

$$\mathbf{z}' = A\mathbf{z},\ \mathbf{z}(0) = \mathbf{0}. \tag{16.14}$$

Note that for $K = \max\{|a_{ij}|\}$, where $A = (a_{ij})$,

$$|(A\mathbf{z},\mathbf{z})| = \left|\sum_{ij} a_{ij}z_j\overline{z_i}\right| \leq K\sum_{ij}|z_i|\,|z_j|$$

$$\leq K\sum_{ij}\left(\frac{|z_i|^2}{2} + \frac{|z_j|^2}{2}\right) = nK|\mathbf{z}|^2.$$

(For x and y real numbers, $xy \leq \frac{x^2}{2} + \frac{y^2}{2}$ because this is equivalent to saying $(x-y)^2 \geq 0$.) Similarly,

$$|(\mathbf{z},A\mathbf{z})| \leq nK|\mathbf{z}|^2$$

Thus,

$$|(\mathbf{z},A\mathbf{z})|, |(A\mathbf{z},\mathbf{z})| \leq nK|\mathbf{z}|^2. \tag{16.15}$$

Now multiplying 16.14 by $\mathbf{z}$ and observing that

$$\frac{d}{dt}\left(|\mathbf{z}|^2\right) = (\mathbf{z}',\mathbf{z}) + (\mathbf{z},\mathbf{z}') = (A\mathbf{z},\mathbf{z}) + (\mathbf{z},A\mathbf{z}),$$

it follows from 16.15 and the observation that $\mathbf{z}(0) = 0$,

$$|\mathbf{z}(t)|^2 \leq \int_0^t 2nK |\mathbf{z}(s)|^2 ds$$

and so by Gronwall's inequality, $|\mathbf{z}(t)|^2 = 0$ for all $t \geq 0$. Thus,

$$\mathbf{x}(t) = \mathbf{y}(t)$$

for all $t \geq a$.

Now let $\mathbf{w}(t) = \mathbf{x}(a-t) - \mathbf{y}(a-t)$ for $t \geq 0$. Then $\mathbf{w}'(t) = (-A)\mathbf{w}(t)$ and you can repeat the argument which was just given to conclude that $\mathbf{x}(t) = \mathbf{y}(t)$ for all $t \leq a$. This proves the theorem.

Definition 16.4.5 *Let A be an $n \times n$ matrix. We say $\Phi(t)$ is a fundamental matrix for A if*

$$\Phi'(t) = A\Phi(t), \ \Phi(0) = I, \tag{16.16}$$

and $\Phi(t)^{-1}$ exists for all $t \in \mathbb{R}$.

Why should anyone care about a fundamental matrix? The reason is that such a matrix valued function makes possible a convenient description of the solution of the initial value problem,

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f}(t), \ \mathbf{x}(0) = \mathbf{x}_0, \tag{16.17}$$

on the interval, $[0, T]$. First consider the special case where $n = 1$. This is the first order linear differential equation,

$$r' = \lambda r + g, \ r(0) = r_0, \tag{16.18}$$

where g is a continuous scalar valued function. First consider the case where $g = 0$.

Lemma 16.4.6 *There exists a unique solution to the initial value problem,*

$$r' = \lambda r, \ r(0) = 1, \tag{16.19}$$

and the solution for $\lambda = a + ib$ is given by

$$r(t) = e^{at}(\cos bt + i \sin bt). \tag{16.20}$$

This solution to the initial value problem is denoted as $e^{\lambda t}$. (If λ is real, $e^{\lambda t}$ as defined here reduces to the usual exponential function so there is no contradiction between this and earlier notation seen in Calculus.)

Proof: From the uniqueness theorem presented above, Theorem 16.4.4, applied to the case where $n = 1$, there can be no more than one solution to the initial value problem, 16.19. Therefore, it only remains to verify 16.20 is a solution to 16.19. However, this is an easy calculus exercise. This proves the Lemma.

Note the differential equation in 16.19 says

$$\frac{d}{dt}\left(e^{\lambda t}\right) = \lambda e^{\lambda t}. \tag{16.21}$$

With this lemma, it becomes possible to easily solve the case in which $g \neq 0$.

Theorem 16.4.7 *There exists a unique solution to 16.18 and this solution is given by the formula,*

$$r(t) = e^{\lambda t} r_0 + e^{\lambda t} \int_0^t e^{-\lambda s} g(s) ds. \tag{16.22}$$

Proof: By the uniqueness theorem, Theorem 16.4.4, there is no more than one solution. It only remains to verify that 16.22 is a solution. But $r(0) = e^{\lambda 0} r_0 + \int_0^0 e^{-\lambda s} g(s)\, ds = r_0$ and so the initial condition is satisfied. Next differentiate this expression to verify the differential equation is also satisfied. Using 16.21, the product rule and the fundamental theorem of calculus,

$$\begin{aligned} r'(t) &= \lambda e^{\lambda t} r_0 + \lambda e^{\lambda t} \int_0^t e^{-\lambda s} g(s)\, ds + e^{\lambda t} e^{-\lambda t} g(t) \\ &= \lambda r(t) + g(t). \end{aligned}$$

This proves the Theorem.

Now consider the question of finding a fundamental matrix for A. When this is done, it will be easy to give a formula for the general solution to 16.17 known as the variation of constants formula, arguably the most important result in differential equations.

The next theorem gives a formula for the fundamental matrix16.16. It is known as Putzer's method [1].

Theorem 16.4.8 *Let A be an $n \times n$ matrix whose eigenvalues are $\{\lambda_1, \cdots, \lambda_n\}$. Define*

$$P_k(A) \equiv \prod_{m=1}^{k} (A - \lambda_m I),\ P_0(A) \equiv I,$$

and let the scalar valued functions, $r_k(t)$ be defined as the solutions to the following initial value problem

$$\begin{pmatrix} r_0'(t) \\ r_1'(t) \\ r_2'(t) \\ \vdots \\ r_n'(t) \end{pmatrix} = \begin{pmatrix} 0 \\ \lambda_1 r_1(t) + r_0(t) \\ \lambda_2 r_2(t) + r_1(t) \\ \vdots \\ \lambda_n r_n(t) + r_{n-1}(t) \end{pmatrix}, \quad \begin{pmatrix} r_0(0) \\ r_1(0) \\ r_2(0) \\ \vdots \\ r_n(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Note the system amounts to a list of single first order linear differential equations. Now define

$$\Phi(t) \equiv \sum_{k=0}^{n-1} r_{k+1}(t) P_k(A).$$

Then

$$\Phi'(t) = A\Phi(t),\ \Phi(0) = I. \tag{16.23}$$

Furthermore, if $\Phi(t)$ is a solution to 16.23 for all t, then it follows $\Phi(t)^{-1}$ exists for all t and $\Phi(t)$ is the unique fundamental matrix for A.

Proof: The first part of this follows from a computation. First note that by the Cayley Hamilton theorem, $P_n(A) = 0$. Now for the computation:

$$\Phi'(t) = \sum_{k=0}^{n-1} r_{k+1}'(t) P_k(A) = \sum_{k=0}^{n-1} (\lambda_{k+1} r_{k+1}(t) + r_k(t)) P_k(A) =$$

$$\sum_{k=0}^{n-1} \lambda_{k+1} r_{k+1}(t) P_k(A) + \sum_{k=0}^{n-1} r_k(t) P_k(A) = \sum_{k=0}^{n-1} (\lambda_{k+1} I - A) r_{k+1}(t) P_k(A) +$$

$$\sum_{k=0}^{n-1} r_k'(t) P_k(A) + \sum_{k=0}^{n-1} A r_{k+1}(t) P_k(A)$$

$$= -\sum_{k=0}^{n-1} r_{k+1}(t) P_{k+1}(A) + \sum_{k=0}^{n-1} r_k(t) P_k(A) + A\sum_{k=0}^{n-1} r_{k+1}(t) P_k(A). \tag{16.24}$$

Now using $r_0(t) = 0$, the first term equals

$$\begin{aligned} -\sum_{k=1}^{n} r_k(t) P_k(A) &= -\sum_{k=1}^{n-1} r_k(t) P_k(A) \\ &= -\sum_{k=0}^{n-1} r_k(t) P_k(A) \end{aligned}$$

and so 16.24 reduces to

$$A\sum_{k=0}^{n-1} r_{k+1}(t) P_k(A) = A\Phi(t).$$

This shows $\Phi'(t) = A\Phi(t)$. That $\Phi(0) = 0$ follows from

$$\Phi(0) = \sum_{k=0}^{n-1} r_{k+1}(0) P_k(A) = r_1(0) P_0 = I.$$

It remains to verify that if 16.23 holds, then $\Phi(t)^{-1}$ exists for all t. To do so, consider $\mathbf{v} \neq \mathbf{0}$ and suppose for some t_0, $\Phi(t_0)\mathbf{v} = \mathbf{0}$. Let $\mathbf{x}(t) \equiv \Phi(t_0 + t)\mathbf{v}$. Then

$$\mathbf{x}'(t) = A\Phi(t_0 + t)\mathbf{v} = A\mathbf{x}(t),\ \mathbf{x}(0) = \Phi(t_0)\mathbf{v} = \mathbf{0}.$$

But also $\mathbf{z}(t) \equiv \mathbf{0}$ also satisfies

$$\mathbf{z}'(t) = A\mathbf{z}(t),\ \mathbf{z}(0) = \mathbf{0},$$

and so by the theorem on uniqueness, it must be the case that $\mathbf{z}(t) = \mathbf{x}(t)$ for all t, showing that $\Phi(t + t_0)\mathbf{v} = \mathbf{0}$ for all t, and in particular for $t = -t_0$. Therefore,

$$\Phi(-t_0 + t_0)\mathbf{v} = I\mathbf{v} = \mathbf{0}$$

and so $\mathbf{v} = \mathbf{0}$, a contradiction. It follows that $\Phi(t)$ must be one to one for all t and so, $\Phi(t)^{-1}$ exists for all t.

It only remains to verify the solution to 16.23 is unique. Suppose Ψ is another fundamental matrix solving 16.23. Then letting $\mathbf{v}$ be an arbitrary vector,

$$\mathbf{z}(t) \equiv \Phi(t)\mathbf{v},\ \mathbf{y}(t) \equiv \Psi(t)\mathbf{v}$$

both solve the initial value problem,

$$\mathbf{x}' = A\mathbf{x},\ \mathbf{x}(0) = \mathbf{v},$$

and so by the uniqueness theorem, $\mathbf{z}(t) = \mathbf{y}(t)$ for all t showing that $\Phi(t)\mathbf{v} = \Psi(t)\mathbf{v}$ for all t. Since $\mathbf{v}$ is arbitrary, this shows that $\Phi(t) = \Psi(t)$ for every t. This proves the theorem.

It is useful to consider the differential equations for the r_k for $k \geq 1$. As noted above, $r_0(t) = 0$ and $r_1(t) = e^{\lambda_1 t}$.

$$r_{k+1}' = \lambda_{k+1} r_{k+1} + r_k,\ r_{k+1}(0) = 0.$$

Thus

$$r_{k+1}(t) = \int_0^t e^{\lambda_{k+1}(t-s)} r_k(s)\, ds.$$

Therefore,

$$r_2(t) = \int_0^t e^{\lambda_2(t-s)} e^{\lambda_1 s} ds = \frac{e^{\lambda_1 t} - e^{\lambda_2 t}}{-\lambda_2 + \lambda_1}$$

assuming $\lambda_1 \neq \lambda_2$.

Sometimes people define a fundamental matrix to be a matrix, $\Phi(t)$ such that $\Phi'(t) = A\Phi(t)$ and $\det(\Phi(t)) \neq 0$ for all t. Thus this avoids the initial condition, $\Phi(0) = I$. The next proposition has to do with this situation.

Proposition 16.4.9 *Suppose A is an $n \times n$ matrix and suppose $\Phi(t)$ is an $n \times n$ matrix for each $t \in \mathbb{R}$ with the property that*

$$\Phi'(t) = A\Phi(t). \tag{16.25}$$

Then either $\Phi(t)^{-1}$ exists for all $t \in \mathbb{R}$ or $\Phi(t)^{-1}$ fails to exist for all $t \in \mathbb{R}$.

Proof: Suppose $\Phi(0)^{-1}$ exists and 16.25 holds. Let $\Psi(t) \equiv \Phi(t)\Phi(0)^{-1}$. Then $\Psi(0) = I$ and

$$\Psi'(t) = \Phi'(t)\Phi(0)^{-1} = A\Phi(t)\Phi(0)^{-1} = A\Psi(t)$$

so by Theorem 16.4.8, $\Psi(t)^{-1}$ exists for all t. Therefore, $\Phi(t)^{-1}$ also exists for all t.

Next suppose $\Phi(0)^{-1}$ does not exist. I need to show $\Phi(t)^{-1}$ does not exist for any t. Suppose then that $\Phi(t_0)^{-1}$ does exist. Then let$\Psi(t) \equiv \Phi(t_0 + t)\Phi(t_0)^{-1}$. Then $\Psi(0) = I$ and $\Psi' = A\Psi$ so by Theorem 16.4.8 it follows $\Psi(t)^{-1}$ exists for all t and so for all $t, \Phi(t + t_0)^{-1}$ must also exist, even for $t = -t_0$ which implies $\Phi(0)^{-1}$ exists after all. This proves the proposition.

The conclusion of this proposition is usually referred to as the Wronskian alternative and another way to say it is that if 16.25 holds, then either $\det(\Phi(t)) = 0$ for all t or $\det(\Phi(t))$ is never equal to 0. The Wronskian is the usual name of the function, $t \to \det(\Phi(t))$.

The following theorem gives the variation of constants formula,.

Theorem 16.4.10 *Let $\mathbf{f}$ be continuous on $[0, T]$ and let A be an $n \times n$ matrix and $\mathbf{x}_0$ a vector in $\mathbb{C}^n$. Then there exists a unique solution to 16.17, $\mathbf{x}$, given by the variation of constants formula,*

$$\mathbf{x}(t) = \Phi(t)\mathbf{x}_0 + \Phi(t)\int_0^t \Phi(s)^{-1}\mathbf{f}(s)\, ds \tag{16.26}$$

for $\Phi(t)$ the fundamental matrix for A. Also, $\Phi(t)^{-1} = \Phi(-t)$ and $\Phi(t+s) = \Phi(t)\Phi(s)$ for all t, s and the above variation of constants formula can also be written as

$$\begin{aligned} \mathbf{x}(t) &= \Phi(t)\mathbf{x}_0 + \int_0^t \Phi(t-s)\mathbf{f}(s)\, ds && (16.27) \\ &= \Phi(t)\mathbf{x}_0 + \int_0^t \Phi(s)\mathbf{f}(t-s)\, ds && (16.28) \end{aligned}$$

Proof: From the uniqueness theorem there is at most one solution to 16.17. Therefore, if 16.26 solves 16.17, the theorem is proved. The verification that the given formula works is identical with the verification that the scalar formula given in Theorem 16.4.7 solves the initial value problem given there. $\Phi(s)^{-1}$ is continuous because of the formula for the inverse of a matrix in terms of the transpose of the cofactor matrix. Therefore, the integrand in

16.26 is continuous and the fundamental theorem of calculus applies. To verify the formula for the inverse, fix s and consider $\mathbf{x}(t) = \Phi(s+t)\mathbf{v}$, and $\mathbf{y}(t) = \Phi(t)\Phi(s)\mathbf{v}$. Then

$$\mathbf{x}'(t) = A\Phi(t+s)\mathbf{v} = A\mathbf{x}(t),\ \mathbf{x}(0) = \Phi(s)\mathbf{v}$$

$$\mathbf{y}'(t) = A\Phi(t)\Phi(s)\mathbf{v} = A\mathbf{y}(t),\ \mathbf{y}(0) = \Phi(s)\mathbf{v}.$$

By the uniqueness theorem, $\mathbf{x}(t) = \mathbf{y}(t)$ for all t. Since s and $\mathbf{v}$ are arbitrary, this shows $\Phi(t+s) = \Phi(t)\Phi(s)$ for all t, s. Letting $s = -t$ and using $\Phi(0) = I$ verifies $\Phi(t)^{-1} = \Phi(-t)$.

Next, note that this also implies $\Phi(t-s)\Phi(s) = \Phi(t)$ and so $\Phi(t-s) = \Phi(t)\Phi(s)^{-1}$. Therefore, this yields 16.27 and then 16.28follows from changing the variable. This proves the theorem.

If $\Phi' = A\Phi$ and $\Phi(t)^{-1}$ exists for all t, you should verify that the solution to the initial value problem

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f},\ \mathbf{x}(t_0) = \mathbf{x}_0$$

is given by

$$\mathbf{x}(t) = \Phi(t-t_0)\mathbf{x}_0 + \int_{t_0}^{t} \Phi(t-s)\mathbf{f}(s)\,ds.$$

Theorem 16.4.10 is general enough to include all constant coefficient linear differential equations or any order. Thus it includes as a special case the main topics of an entire elementary differential equations class. This is illustrated in the following example. One can reduce an arbitrary linear differential equation to a first order system and then apply the above theory to solve the problem. The next example is a differential equation of damped vibration.

Example 16.4.11 *The differential equation is* $y'' + 2y' + 2y = \cos t$ *and initial conditions,* $y(0) = 1$ *and* $y'(0) = 0$.

To solve this equation, let $x_1 = y$ and $x_2 = x_1' = y'$. Then, writing this in terms of these new variables, yields the following system.

$$\begin{aligned} x_2' + 2x_2 + 2x_1 &= \cos t \\ x_1' &= x_2 \end{aligned}$$

This system can be written in the above form as

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}' &= \begin{pmatrix} x_2 \\ -2x_2 - 2x_1 \end{pmatrix} + \begin{pmatrix} 0 \\ \cos t \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ -2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ \cos t \end{pmatrix}. \end{aligned}$$

and the initial condition is of the form

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Now $P_0(A) \equiv I$. The eigenvalues are $-1+i, -1-i$ and so

$$\begin{aligned} P_1(A) &= \left(\begin{pmatrix} 0 & 1 \\ -2 & -2 \end{pmatrix} - (-1+i) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \\ &= \begin{pmatrix} 1-i & 1 \\ -2 & -1-i \end{pmatrix}. \end{aligned}$$

Recall $r_0(t) \equiv 0$ and $r_1(t) = e^{(-1+i)t}$. Then

$$r_2' = (-1-i)\, r_2 + e^{(-1+i)t},\ r_2(0) = 0$$

and so

$$r_2(t) = \frac{e^{(-1+i)t} - e^{(-1-i)t}}{2i} = e^{-t}\sin(t)$$

Putzer's method yields the fundamental matrix as

$$\begin{aligned}\Phi(t) &= e^{(-1+i)t}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + e^{-t}\sin(t)\begin{pmatrix} 1-i & 1 \\ -2 & -1-i \end{pmatrix} \\ &= \begin{pmatrix} e^{-t}(\cos(t)+\sin(t)) & e^{-t}\sin t \\ -2e^{-t}\sin t & e^{-t}(\cos(t)-\sin(t)) \end{pmatrix}\end{aligned}$$

From variation of constants formula the desired solution is

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}(t) = \begin{pmatrix} e^{-t}(\cos(t)+\sin(t)) & e^{-t}\sin t \\ -2e^{-t}\sin t & e^{-t}(\cos(t)-\sin(t)) \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$+\int_0^t \begin{pmatrix} e^{-s}(\cos(s)+\sin(s)) & e^{-s}\sin s \\ -2e^{-s}\sin s & e^{-s}(\cos(s)-\sin(s)) \end{pmatrix}\begin{pmatrix} 0 \\ \cos(t-s) \end{pmatrix}$$

$$= \begin{pmatrix} e^{-t}(\cos(t)+\sin(t)) \\ -2e^{-t}\sin t \end{pmatrix} + \int_0^t \begin{pmatrix} e^{-s}\sin(s)\cos(t-s) \\ e^{-s}(\cos s - \sin s)\cos(t-s) \end{pmatrix} ds$$

$$= \begin{pmatrix} e^{-t}(\cos(t)+\sin(t)) \\ -2e^{-t}\sin t \end{pmatrix} + \begin{pmatrix} -\frac{1}{5}(\cos t)\, e^{-t} - \frac{3}{5}e^{-t}\sin t + \frac{1}{5}\cos t + \frac{2}{5}\sin t \\ -\frac{2}{5}(\cos t)\, e^{-t} + \frac{4}{5}e^{-t}\sin t + \frac{2}{5}\cos t - \frac{1}{5}\sin t \end{pmatrix}$$

$$= \begin{pmatrix} \frac{4}{5}(\cos t)\, e^{-t} + \frac{2}{5}e^{-t}\sin t + \frac{1}{5}\cos t + \frac{2}{5}\sin t \\ -\frac{6}{5}e^{-t}\sin t - \frac{2}{5}(\cos t)\, e^{-t} + \frac{2}{5}\cos t - \frac{1}{5}\sin t \end{pmatrix}$$

Thus $y(t) = x_1(t) = \frac{4}{5}(\cos t)\, e^{-t} + \frac{2}{5}e^{-t}\sin t + \frac{1}{5}\cos t + \frac{2}{5}\sin t$.

16.5 Geometric Theory Of Autonomous Systems

Here a sufficient condition is given for stability of a first order system. First of all, here is a fundamental estimate for the entries of a fundamental matrix.

Lemma 16.5.1 *Let the functions, r_k be given in the statement of Theorem 16.4.8 and suppose that A is an $n \times n$ matrix whose eigenvalues are $\{\lambda_1, \cdots, \lambda_n\}$. Suppose that these eigenvalues are ordered such that*

$$\mathrm{Re}(\lambda_1) \le \mathrm{Re}(\lambda_2) \le \cdots \le \mathrm{Re}(\lambda_n) < 0.$$

Then if $0 > -\delta > \mathrm{Re}(\lambda_n)$ is given, there exists a constant, C such that for each $k = 0, 1, \cdots, n$,

$$|r_k(t)| \le Ce^{-\delta t} \tag{16.29}$$

for all $t > 0$.

Proof: This is obvious for $r_0(t)$ because it is identically equal to 0. From the definition of the r_k,

$$r_1' = \lambda_1 r_1, r_1(0) = 1$$

and so

$$r_1(t) = e^{\lambda_1 t}$$

which implies

$$|r_1(t)| \leq e^{\mathrm{Re}(\lambda_1)t}.$$

Suppose for some $m \geq 1$ there exists a constant, C_m such that

$$|r_k(t)| \leq C_m t^m e^{\mathrm{Re}(\lambda_m)t}$$

for all $k \leq m$ for all $t > 0$. Then

$$r'_{m+1}(t) = \lambda_{m+1} r_{m+1}(t) + r_m(t),\ r_{m+1}(0) = 0$$

and so

$$r_{m+1}(t) = e^{\lambda_{m+1}t} \int_0^t e^{-\lambda_{m+1}s} r_m(s)\, ds.$$

Then by the induction hypothesis,

$$\begin{aligned} |r_{m+1}(t)| &\leq e^{\mathrm{Re}(\lambda_{m+1})t} \int_0^t \left|e^{-\lambda_{m+1}s}\right| C_m s^m e^{\mathrm{Re}(\lambda_m)s} ds \\ &\leq e^{\mathrm{Re}(\lambda_{m+1})t} \int_0^t s^m C_m e^{-\mathrm{Re}(\lambda_{m+1})s} e^{\mathrm{Re}(\lambda_m)s} ds \\ &\leq e^{\mathrm{Re}(\lambda_{m+1})t} \int_0^t s^m C_m ds = \frac{C_m}{m+1} t^{m+1} e^{\mathrm{Re}(\lambda_{m+1})t} \end{aligned}$$

It follows by induction there exists a constant, C such that for all $k \leq n$,

$$|r_k(t)| \leq C t^n e^{\mathrm{Re}(\lambda_n)t}$$

and this obviously implies the conclusion of the lemma.

The proof of the above lemma yields the following corollary.

Corollary 16.5.2 *Let the functions, r_k be given in the statement of Theorem 16.4.8 and suppose that A is an $n \times n$ matrix whose eigenvalues are $\{\lambda_1, \cdots, \lambda_n\}$. Suppose that these eigenvalues are ordered such that*

$$\mathrm{Re}(\lambda_1) \leq \mathrm{Re}(\lambda_2) \leq \cdots \leq \mathrm{Re}(\lambda_n).$$

Then there exists a constant C such that for all $k \leq m$

$$|r_k(t)| \leq C t^m e^{\mathrm{Re}(\lambda_m)t}.$$

With the lemma, the following sloppy estimate is available for a fundamental matrix.

Theorem 16.5.3 *Let A be an $n \times n$ matrix and let $\Phi(t)$ be the fundamental matrix for A. That is,*

$$\Phi'(t) = A\Phi(t),\ \Phi(0) = I.$$

Suppose also the eigenvalues of A are $\{\lambda_1, \cdots, \lambda_n\}$ where these eigenvalues are ordered such that

$$\mathrm{Re}(\lambda_1) \leq \mathrm{Re}(\lambda_2) \leq \cdots \leq \mathrm{Re}(\lambda_n) < 0.$$

Then if $0 > -\delta > \mathrm{Re}(\lambda_n)$, is given, there exists a constant, C such that

$$\left|\Phi(t)_{ij}\right| \leq C e^{-\delta t}$$

for all $t > 0$. Also

$$|\Phi(t)\mathbf{x}| \leq C n^{3/2} e^{-\delta t} |\mathbf{x}|. \tag{16.30}$$

Proof: Let

$$M \equiv \max\left\{P_k\left(A\right)_{ij} \text{ for all } i,j,k\right\}.$$

Then from Putzer's formula for $\Phi\left(t\right)$ and Lemma 16.5.1, there exists a constant, C such that

$$\left|\Phi\left(t\right)_{ij}\right| \leq \sum_{k=0}^{n-1} Ce^{-\delta t}M.$$

Let the new C be given by nCM. This proves the theorem.

Next,

$$\begin{aligned}
\left|\Phi\left(t\right)\mathbf{x}\right|^2 &\equiv \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\Phi_{ij}\left(t\right)x_j\right)^2 \\
&\leq \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\left|\Phi_{ij}\left(t\right)\right|\left|x_j\right|\right)^2 \\
&\leq \sum_{i=1}^{n}\left(\sum_{j=1}^{n}Ce^{-\delta t}\left|\mathbf{x}\right|\right)^2 \\
&= C^2e^{-2\delta t}\sum_{i=1}^{n}\left(n\left|\mathbf{x}\right|\right)^2 = C^2e^{-2\delta t}n^3\left|\mathbf{x}\right|^2
\end{aligned}$$

This proves 16.30 and completes the proof.

Definition 16.5.4 *Let $\mathbf{f}: U \to \mathbb{R}^n$ where U is an open subset of $\mathbb{R}^n$ such that $\mathbf{a} \in U$ and $\mathbf{f}\left(\mathbf{a}\right) = \mathbf{0}$. A point, $\mathbf{a}$ where $\mathbf{f}\left(\mathbf{a}\right) = \mathbf{0}$ is called an equilibrium point. Then $\mathbf{a}$ is assymptotically stable if for any $\varepsilon > 0$ there exists $r > 0$ such that whenever $\left|\mathbf{x}_0 - \mathbf{a}\right| < r$ and $\mathbf{x}\left(t\right)$ the solution to the initial value problem,*

$$\mathbf{x}' = \mathbf{f}\left(\mathbf{x}\right),\ \mathbf{x}\left(0\right) = \mathbf{x}_0,$$

it follows

$$\lim_{t\to\infty}\mathbf{x}\left(t\right) = \mathbf{a},\ \left|\mathbf{x}\left(t\right) - \mathbf{a}\right| < \varepsilon$$

A differential equation of the form $\mathbf{x}' = \mathbf{f}\left(\mathbf{x}\right)$ is called autonomous as opposed to a nonautonomous equation of the form $\mathbf{x}' = \mathbf{f}\left(t,\mathbf{x}\right)$. The equilibrium point $\mathbf{a}$ is stable if for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\left|\mathbf{x}_0 - \mathbf{a}\right| < \delta$, then if $\mathbf{x}$ is the solution of

$$\mathbf{x}' = \mathbf{f}\left(\mathbf{x}\right),\ \mathbf{x}\left(0\right) = \mathbf{x}_0, \tag{16.31}$$

then $\left|\mathbf{x}\left(t\right) - \mathbf{a}\right| < \varepsilon$ for all $t > 0$.

Obviously assymptotic stability implies stability.

An ordinary differential equation is called almost linear if it is of the form

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}\left(\mathbf{x}\right)$$

where A is an $n \times n$ matrix and

$$\lim_{\mathbf{x}\to\mathbf{0}}\frac{\mathbf{g}\left(\mathbf{x}\right)}{\left|\mathbf{x}\right|} = \mathbf{0}.$$

Now the stability of an equilibrium point of an autonomous system,

$$\mathbf{x}' = \mathbf{f}(\mathbf{x})$$

can always be reduced to the consideration of the stability of $\mathbf{0}$ for an almost linear system. Here is why. If you are considering the equilibrium point, $\mathbf{a}$ for $\mathbf{x}' = \mathbf{f}(\mathbf{x})$, you could define a new variable, $\mathbf{y}$ by

$$\mathbf{a} + \mathbf{y} = \mathbf{x}.$$

Then assymptotic stability would involve $|\mathbf{y}(t)| < \varepsilon$ and $\lim_{t\to\infty} \mathbf{y}(t) = \mathbf{0}$ while stability would only require $|\mathbf{y}(t)| < \varepsilon$. Then since $\mathbf{a}$ is an equilibrium point, $\mathbf{y}$ solves the following initial value problem.

$$\mathbf{y}' = \mathbf{f}(\mathbf{a} + \mathbf{y}) - \mathbf{f}(\mathbf{a}),\ \mathbf{y}(0) = \mathbf{y}_0,$$

where $\mathbf{y}_0 = \mathbf{x}_0 - \mathbf{a}$.

Let $A = D\mathbf{f}(\mathbf{a})$. Then from the definition of the derivative of a function,

$$\mathbf{y}' = A\mathbf{y} + \mathbf{g}(\mathbf{y}),\ \mathbf{y}(0) = \mathbf{y}_0 \tag{16.32}$$

where

$$\lim_{\mathbf{y}\to\mathbf{0}} \frac{\mathbf{g}(\mathbf{y})}{|\mathbf{y}|} = \mathbf{0}.$$

Thus there is never any loss of generality in considering only the equilibrium point $\mathbf{0}$ for an almost linear system.[1] Therefore, from now on I will only consider the case of almost linear systems and the equilibrium point $\mathbf{0}$.

Theorem 16.5.5 *Consider the almost linear system of equations,*

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x})$$

where

$$\lim_{\mathbf{x}\to\mathbf{0}} \frac{\mathbf{g}(\mathbf{x})}{|\mathbf{x}|} = \mathbf{0}$$

and $\mathbf{g}$ *is a* C^1 *function. Suppose that for all* λ *an eigenvalue of* A, $\operatorname{Re}\lambda < 0$. *Then* $\mathbf{0}$ *is assymptotically stable.*

Proof: By Theorem 16.5.3 there exist constants $\delta > 0$ and K such that for $\Phi(t)$ the fundamental matrix for A,

$$|\Phi(t)\mathbf{x}| \le Ke^{-\delta t}|\mathbf{x}|.$$

Let $\varepsilon > 0$ be given and let r be small enough that $Kr < \varepsilon$ and for $|\mathbf{x}| < (K+1)r, |\mathbf{g}(\mathbf{x})| < \eta|\mathbf{x}|$ where η is so small that $K\eta < \delta$, and let $|\mathbf{y}_0| < r$. Then by the variation of constants formula, the solution to **??**, at least for small t satisfies

$$\mathbf{y}(t) = \Phi(t)\mathbf{y}_0 + \int_0^t \Phi(t-s)\mathbf{g}(\mathbf{y}(s))\,ds.$$

The following estimate holds.

$$\begin{aligned} |\mathbf{y}(t)| &\le Ke^{-\delta t}|\mathbf{y}_0| + \int_0^t Ke^{-\delta(t-s)}\eta|\mathbf{y}(s)|\,ds \\ &< Ke^{-\delta t}r + \int_0^t Ke^{-\delta(t-s)}\eta|\mathbf{y}(s)|\,ds. \end{aligned}$$

[1] This is no longer true when you study partial differential equatioins as ordinary differential equationis in infinite dimensional spaces.

Therefore,

$$e^{\delta t}\left|\mathbf{y}\left(t\right)\right| < Kr + \int_0^t K\eta e^{\delta s}\left|\mathbf{y}\left(s\right)\right| ds.$$

By Gronwall's inequality,

$$e^{\delta t}\left|\mathbf{y}\left(t\right)\right| < Kre^{K\eta t}$$

and so

$$\left|\mathbf{y}\left(t\right)\right| < Kre^{(K\eta-\delta)t} < \varepsilon e^{(K\eta-\delta)t}$$

Therefore, $\left|\mathbf{y}\left(t\right)\right| < Kr < \varepsilon$ for all t and so from Corollary 16.3.4, the solution to **??** exists for all $t \geq 0$ and since $K\eta - \delta < 0$,

$$\lim_{t\to\infty}\left|\mathbf{y}\left(t\right)\right| = 0.$$

This proves the theorem.

16.6 General Geometric Theory

Here I will consider the case where the matrix, A has both postive and negative eigenvalues. First here is a useful lemma.

Lemma 16.6.1 *Suppose A is an $n \times n$ matrix and there exists $\delta > 0$ such that*

$$0 < \delta < \mathrm{Re}\left(\lambda_1\right) \leq \cdots \leq \mathrm{Re}\left(\lambda_n\right)$$

where $\{\lambda_1, \cdots, \lambda_n\}$ are the eigenvalues of A, with possibly some repeated. Then there exists a constant, C such that for all $t < 0$,

$$\left|\Phi\left(t\right)\mathbf{x}\right| \leq Ce^{\delta t}\left|\mathbf{x}\right|$$

Proof: I want an estimate on the solutions to the system

$$\Phi'\left(t\right) = A\Phi\left(t\right),\ \Phi\left(0\right) = I.$$

for $t < 0$. Let $s = -t$ and let $\Psi\left(s\right) = \Phi\left(t\right)$. Then writing this in terms of Ψ,

$$\Psi'\left(s\right) = -A\Psi\left(s\right),\ \Psi\left(0\right) = I.$$

Now the eigenvalues of $-A$ have real parts less than $-\delta$ because these eigenvalues are obtained from the eigenvalues of A by multiplying by -1. Then by Theorem 16.5.3 there exists a constant, C such that for any $\mathbf{x}$,

$$\left|\Psi\left(s\right)\mathbf{x}\right| \leq Ce^{-\delta s}\left|\mathbf{x}\right|.$$

Therefore, from the definition of Ψ,

$$\left|\Phi\left(t\right)\mathbf{x}\right| \leq Ce^{\delta t}\left|\mathbf{x}\right|.$$

This proves the lemma.

Here is another essential lemma which is found in Coddington and Levinson [3]

Lemma 16.6.2 *Let $p_j(t)$ be polynomials with complex coefficients and let*

$$f(t) = \sum_{j=1}^{m} p_j(t) e^{\lambda_j t}$$

where $m \geq 1, \lambda_j \neq \lambda_k$ for $j \neq k$, and none of the $p_j(t)$ vanish identically. Let

$$\sigma = \max(\mathrm{Re}(\lambda_1), \cdots, \mathrm{Re}(\lambda_m)).$$

Then there exists a positive number, r and arbitrarily large positive values of t such that

$$e^{-\sigma t} |f(t)| > r.$$

In particular, $|f(t)|$ is unbounded.

Proof: Suppose the largest exponent of any of the p_j is M and let $\lambda_j = a_j + ib_j$. First assume each $a_j = 0$. This is convenient because $\sigma = 0$ in this case and the largest of the $\mathrm{Re}(\lambda_j)$ occurs in every λ_j.

Then arranging the above sum as a sum of decreasing powers of t,

$$f(t) = t^M f_M(t) + \cdots + t f_1(t) + f_0(t).$$

Then

$$t^{-M} f(t) = f_M(t) + O\left(\frac{1}{t}\right)$$

where the last term means that $tO\left(\frac{1}{t}\right)$ is bounded. Then

$$f_M(t) = \sum_{j=1}^{m} c_j e^{ib_j t}$$

It can't be the case that all the c_j are equal to 0 because then M would not be the highest power exponent. Suppose $c_k \neq 0$. Then

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T t^{-M} f(t) e^{-ib_k t} dt = \sum_{j=1}^{m} c_j \frac{1}{T} \int_0^T e^{i(b_j - b_k)t} dt = c_k \neq 0.$$

Letting $r = |c_k/2|$, it follows $\left|t^{-M} f(t) e^{-ib_k t}\right| > r$ for arbitrarily large values of t. Thus it is also true that $|f(t)| > r$ for arbitrarily large values of t.

Next consider the general case in which σ is given above. Thus

$$e^{-\sigma t} f(t) = \sum_{j:a_j=\sigma} p_j(t) e^{b_j t} + g(t)$$

where $\lim_{t\to\infty} g(t) = 0$, $g(t)$ being of the form $\sum_s p_s(t) e^{(a_s - \sigma + ib_s)t}$ where $a_s - \sigma < 0$. Then this reduces to the case above in which $\sigma = 0$. Therefore, there exists $r > 0$ such that

$$\left|e^{-\sigma t} f(t)\right| > r$$

for arbitrarily large values of t. This proves the lemma.

Next here is a Banach space which will be useful.

Lemma 16.6.3 *For $\gamma > 0$, let*

$$E_\gamma = \left\{ \mathbf{x} \in BC\left([0,\infty), \mathbb{F}^n\right) : t \to e^{\gamma t}\mathbf{x}\left(t\right) \text{ is also in } BC\left([0,\infty), \mathbb{F}^n\right) \right\}$$

and let the norm be given by

$$||\mathbf{x}||_\gamma \equiv \sup\left\{ \left| e^{\gamma t}\mathbf{x}\left(t\right) \right| : t \in [0,\infty) \right\}$$

Then E_γ is a Banach space.

Proof: Let $\{\mathbf{x}_k\}$ be a Cauchy sequence in E_γ. Then since $BC\left([0,\infty), \mathbb{F}^n\right)$ is a Banach space, there exists $\mathbf{y} \in BC\left([0,\infty), \mathbb{F}^n\right)$ such that $e^{\gamma t}\mathbf{x}_k\left(t\right)$ converges uniformly on $[0,\infty)$ to $\mathbf{y}\left(t\right)$. Therefore $e^{-\gamma t}e^{\gamma t}\mathbf{x}_k\left(t\right) = \mathbf{x}_k\left(t\right)$ converges uniformly to $e^{-\gamma t}\mathbf{y}\left(t\right)$ on $[0,\infty)$. Define $\mathbf{x}\left(t\right) \equiv e^{-\gamma t}\mathbf{y}\left(t\right)$. Then $\mathbf{y}\left(t\right) = e^{\gamma t}\mathbf{x}\left(t\right)$ and by definition,

$$||\mathbf{x}_k - \mathbf{x}||_\gamma \to 0.$$

This proves the lemma.

16.7 The Stable Manifold

Here assume

$$A = \begin{pmatrix} A_- & 0 \\ 0 & A_+ \end{pmatrix} \tag{16.33}$$

where A_- and A_+ are square matrices of size $k \times k$ and $(n-k) \times (n-k)$ respectively. Also assume A_- has eigenvalues whose real parts are all less than $-\alpha$ while A_+ has eigenvalues whose real parts are all larger than α. Assume also that each of A_- and A_+ is upper triangular.

Also, I will use the following convention. For $\mathbf{v} \in \mathbb{F}^n$,

$$\mathbf{v} = \begin{pmatrix} \mathbf{v}_- \\ \mathbf{v}_+ \end{pmatrix}$$

where $\mathbf{v}_-$ consists of the first k entries of $\mathbf{v}$.

Then from Theorem 16.5.3 and Lemma 16.6.1 the following lemma is obtained.

Lemma 16.7.1 *Let A be of the form given in 16.33 as explained above and let $\Phi_+\left(t\right)$ and $\Phi_-\left(t\right)$ be the fundamental matrices corresponding to A_+ and A_- respectively. Then there exist positive constants, α and γ such that*

$$\left|\Phi_+\left(t\right)\mathbf{y}\right| \leq Ce^{\alpha t} \text{ for all } t < 0 \tag{16.34}$$

$$\left|\Phi_-\left(t\right)\mathbf{y}\right| \leq Ce^{-(\alpha+\gamma)t} \text{ for all } t > 0. \tag{16.35}$$

Also for any nonzero $\mathbf{x} \in \mathbb{C}^{n-k}$,

$$\left|\Phi_+\left(t\right)\mathbf{x}\right| \text{ is unbounded.} \tag{16.36}$$

Proof: The first two claims have been established already. It suffices to pick α and γ such that $-\left(\alpha+\gamma\right)$ is larger than all eigenvalues of A_- and α is smaller than all eigenvalues of A_+. It remains to verify 16.36. From the Putzer formula for $\Phi_+\left(t\right)$,

$$\Phi_+\left(t\right)\mathbf{x} = \sum_{k=0}^{n-1} r_{k+1}\left(t\right) P_k\left(A\right)\mathbf{x}$$

where $P_0(A) \equiv I$. Now each r_k is a polynomial (possibly a constant) times an exponential. This follows easily from the definition of the r_k as solutions of the differential equations

$$r'_{k+1} = \lambda_{k+1} r_{k+1} + r_k.$$

Now by assumption the eigenvalues have positive real parts so

$$\sigma \equiv \max(\mathrm{Re}(\lambda_1), \cdots, \mathrm{Re}(\lambda_{n-k})) > 0.$$

It can also be assumed

$$\mathrm{Re}(\lambda_1) \geq \cdots \geq \mathrm{Re}(\lambda_{n-k})$$

By Lemma 16.6.2 it follows $|\Phi_+(t)\mathbf{x}|$ is unbounded. This follows because

$$\Phi_+(t)\mathbf{x} = r_1(t)\mathbf{x} + \sum_{k=1}^{n-1} r_{k+1}(t)\mathbf{y}_k,\ r_1(t) = e^{\lambda_1 t}.$$

Since $\mathbf{x} \neq \mathbf{0}$, it has a nonzero entry, say $x_m \neq 0$. Consider the m^{th} entry of the vector $\Phi_+(t)\mathbf{x}$. By this Lemma the m^{th} entry is unbounded and this is all it takes for $\mathbf{x}(t)$ to be unbounded. This proves the lemma.

Lemma 16.7.2 *Consider the initial value problem for the almost linear system*

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}),\ \mathbf{x}(0) = \mathbf{x}_0,$$

where $\mathbf{g}$ *is* C^1 *and* A *is of the special form*

$$A = \begin{pmatrix} A_- & 0 \\ 0 & A_+ \end{pmatrix}$$

in which A_- *is a* $k \times k$ *matrix which has eigenvalues for which the real parts are all negative and* A_+ *is a* $(n-k) \times (n-k)$ *matrix for which the real parts of all the eigenvalues are positive. Then* $\mathbf{0}$ *is not stable. More precisely, there exists a set of points* $(\mathbf{a}_-, \boldsymbol{\psi}(\mathbf{a}_-))$ *for* $\mathbf{a}_-$ *small such that for* $\mathbf{x}_0$ *on this set,*

$$\lim_{t\to\infty} \mathbf{x}(t, \mathbf{x}_0) = \mathbf{0}$$

and for $\mathbf{x}_0$ *not on this set, there exists a* $\delta > 0$ *such that* $|\mathbf{x}(t, \mathbf{x}_0)|$ *cannot remain less than* δ *for all positive* t.

Proof: Consider the initial value problem for the almost linear equation,

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}),\ \mathbf{x}(0) = \mathbf{a} = \begin{pmatrix} \mathbf{a}_- \\ \mathbf{a}_+ \end{pmatrix}.$$

Then by the variation of constants formula, a local solution has the form

$$\begin{aligned} \mathbf{x}(t, \mathbf{a}) &= \begin{pmatrix} \Phi_-(t) & 0 \\ 0 & \Phi_+(t) \end{pmatrix} \begin{pmatrix} \mathbf{a}_- \\ \mathbf{a}_+ \end{pmatrix} \\ &\quad + \int_0^t \begin{pmatrix} \Phi_-(t-s) & 0 \\ 0 & \Phi_+(t-s) \end{pmatrix} \mathbf{g}(\mathbf{x}(s, \mathbf{a}))\, ds \end{aligned} \tag{16.37}$$

Write $\mathbf{x}(t)$ for $\mathbf{x}(t,\mathbf{a})$ for short. Let $\varepsilon>0$ be given and suppose δ is such that if $|\mathbf{x}|<\delta$, then $|\mathbf{g}_{\pm}(\mathbf{x})|<\varepsilon|\mathbf{x}|$. Assume from now on that $|\mathbf{a}|<\delta$. Then suppose $|\mathbf{x}(t)|<\delta$ for all $t>0$. Writing 16.37 differently yields

$$\begin{aligned}\mathbf{x}(t,\mathbf{a}) &= \begin{pmatrix}\Phi_-(t) & 0\\ 0 & \Phi_+(t)\end{pmatrix}\begin{pmatrix}\mathbf{a}_-\\ \mathbf{a}_+\end{pmatrix}+\begin{pmatrix}\int_0^t \Phi_-(t-s)\,\mathbf{g}_-(\mathbf{x}(s,\mathbf{a}))\,ds\\ 0\end{pmatrix}\\ &\quad+\begin{pmatrix}0\\ \int_0^t \Phi_+(t-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds\end{pmatrix}\\ &=\begin{pmatrix}\Phi_-(t) & 0\\ 0 & \Phi_+(t)\end{pmatrix}\begin{pmatrix}\mathbf{a}_-\\ \mathbf{a}_+\end{pmatrix}+\begin{pmatrix}\int_0^t \Phi_-(t-s)\,\mathbf{g}_-(\mathbf{x}(s,\mathbf{a}))\,ds\\ 0\end{pmatrix}\\ &\quad+\begin{pmatrix}0\\ \int_0^\infty \Phi_+(t-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds-\int_t^\infty \Phi_+(t-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds\end{pmatrix}.\end{aligned}$$

These improper integrals converge thanks to the assumption that $\mathbf{x}$ is bounded and the estimates 16.34 and 16.35. Continuing the rewriting,

$$\begin{aligned}\begin{pmatrix}\mathbf{x}_-(t)\\ \mathbf{x}_+(t)\end{pmatrix} &= \begin{pmatrix}\left(\Phi_-(t)\,\mathbf{a}_-+\int_0^t \Phi_-(t-s)\,\mathbf{g}_-(\mathbf{x}(s,\mathbf{a}))\,ds\right)\\ \Phi_+(t)\left(\mathbf{a}_++\int_0^\infty \Phi_+(-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds\right)\end{pmatrix}\\ &\quad+\begin{pmatrix}0\\ -\int_t^\infty \Phi_+(t-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds\end{pmatrix}.\end{aligned}$$

It follows from Lemma 16.7.1 that if $|\mathbf{x}(t,\mathbf{a})|$ is bounded by δ as asserted, then it must be the case that $\mathbf{a}_++\int_0^\infty \Phi_+(-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds=\mathbf{0}$. Consequently, it must be the case that

$$\mathbf{x}(t)=\Phi(t)\begin{pmatrix}\mathbf{a}_-\\ \mathbf{0}\end{pmatrix}+\begin{pmatrix}\int_0^t \Phi_-(t-s)\,\mathbf{g}_-(\mathbf{x}(s,\mathbf{a}))\,ds\\ -\int_t^\infty \Phi_+(t-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds\end{pmatrix} \tag{16.38}$$

Letting $t\to 0$, this requires that for a solution to the initial value problem to exist and also satisfy $|\mathbf{x}(t)|<\delta$ for all $t>0$ it must be the case that

$$\mathbf{x}(0)=\begin{pmatrix}\mathbf{a}_-\\ -\int_0^\infty \Phi_+(-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds\end{pmatrix}$$

where $\mathbf{x}(t,\mathbf{a})$ is the solution of

$$\mathbf{x}'=A\mathbf{x}+\mathbf{g}(\mathbf{x}),\ \mathbf{x}(0)=\begin{pmatrix}\mathbf{a}_-\\ -\int_0^\infty \Phi_+(-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds\end{pmatrix}$$

This is because in 16.38, if $\mathbf{x}$ is bounded by δ then the reverse steps show $\mathbf{x}$ is a solution of the above differential equation and initial condition.

It follows if I can show that for all $\mathbf{a}_-$ sufficiently small and $\mathbf{a}=(\mathbf{a}_-,\mathbf{0})^T$, there exists a solution to 16.38 $\mathbf{x}(s,\mathbf{a})$ on $(0,\infty)$ for which $|\mathbf{x}(s,\mathbf{a})|<\delta$, then I can define

$$\boldsymbol{\psi}(\mathbf{a})\equiv-\int_0^\infty \Phi_+(-s)\,\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds$$

and conclude that $|\mathbf{x}(t,\mathbf{x}_0)|<\delta$ for all $t>0$ if and only if $\mathbf{x}_0=(\mathbf{a}_-,\boldsymbol{\psi}(\mathbf{a}_-))^T$ for some sufficiently small $\mathbf{a}_-$.

Let C,α,γ be the constants of Lemma 16.7.1. Let η be a small positive number such that

$$\frac{C\eta}{\alpha}<\frac{1}{6}$$

Note that $\frac{\partial \mathbf{g}}{\partial x_i}(0) = \mathbf{0}$. Therefore, by Lemma 16.3.1, there exists $\delta > 0$ such that if $|\mathbf{x}|, |\mathbf{y}| \leq \delta$, then

$$|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})| < \eta |\mathbf{x} - \mathbf{y}|$$

and in particular,

$$|\mathbf{g}_{\pm}(\mathbf{x}) - \mathbf{g}_{\pm}(\mathbf{y})| < \eta |\mathbf{x} - \mathbf{y}| \tag{16.39}$$

because each $\frac{\partial \mathbf{g}}{\partial x_i}(\mathbf{x})$ is very small. In particular, this implies

$$|\mathbf{g}_{-}(\mathbf{x})| < \eta |\mathbf{x}|, |\mathbf{g}_{+}(\mathbf{x})| < \eta |\mathbf{x}|.$$

For $\mathbf{x} \in E_\gamma$ defined in Lemma 16.6.3 and $|\mathbf{a}_-| < \frac{\delta}{2C}$,

$$F\mathbf{x}(t) \equiv \begin{pmatrix} \Phi_-(t)\mathbf{a}_- + \int_0^t \Phi_-(t-s)\mathbf{g}_-(\mathbf{x}(s))\,ds \\ -\int_t^\infty \Phi_+(t-s)\mathbf{g}_+(\mathbf{x}(s))\,ds \end{pmatrix}.$$

I need to find a fixed point of F. Letting $||\mathbf{x}||_\gamma < \delta$, and using the estimates of Lemma 16.7.1,

$$\begin{aligned} e^{\gamma t}|F\mathbf{x}(t)| &\leq e^{\gamma t}|\Phi_-(t)\mathbf{a}_-| + e^{\gamma t}\int_0^t Ce^{-(\alpha+\gamma)(t-s)}\eta|\mathbf{x}(s)|\,ds \\ &\quad + e^{\gamma t}\int_t^\infty Ce^{\alpha(t-s)}\eta|\mathbf{x}(s)|\,ds \\ &\leq e^{\gamma t}C\frac{\delta}{2C}e^{-(\alpha+\gamma)t} + e^{\gamma t}||\mathbf{x}||_\gamma C\eta\int_0^t e^{-(\alpha+\gamma)(t-s)}e^{-\gamma s}ds \\ &\quad + e^{\gamma t}C\eta\int_t^\infty e^{\alpha(t-s)}e^{-\gamma s}ds\,||\mathbf{x}||_\gamma \\ &< \frac{\delta}{2} + \delta C\eta\int_0^t e^{-\alpha(t-s)}ds + C\eta\delta\int_t^\infty e^{(\alpha+\gamma)(t-s)}ds \\ &< \frac{\delta}{2} + \delta C\eta\frac{1}{\alpha} + \frac{\delta C\eta}{\alpha+\gamma} \leq \delta\left(\frac{1}{2} + \frac{C\eta}{\alpha}\right) < \frac{2\delta}{3}. \end{aligned}$$

Thus F maps every $\mathbf{x} \in E_\gamma$ having $||\mathbf{x}||_\gamma < \delta$ to $F\mathbf{x}$ where $||F\mathbf{x}||_\gamma \leq \frac{2\delta}{3}$.

Now let $\mathbf{x}, \mathbf{y} \in E_\gamma$ where $||\mathbf{x}||_\gamma, ||\mathbf{y}||_\gamma < \delta$. Then

$$\begin{aligned} e^{\gamma t}|F\mathbf{x}(t) - F\mathbf{y}(t)| &\leq e^{\gamma t}\int_0^t |\Phi_-(t-s)|\,\eta e^{-\gamma s}e^{\gamma s}|\mathbf{x}(s) - \mathbf{y}(s)|\,ds \\ &\quad + e^{\gamma t}\int_t^\infty |\Phi_+(t-s)|\,e^{-\gamma s}e^{\gamma s}\eta|\mathbf{x}(s) - \mathbf{y}(s)|\,ds \\ &\leq C\eta||\mathbf{x} - \mathbf{y}||_\gamma\left(\int_0^t e^{-\alpha(t-s)}ds\right) + \int_t^\infty e^{(\alpha+\gamma)(t-s)}ds \\ &\leq C\eta\left(\frac{1}{\alpha} + \frac{1}{\alpha+\gamma}\right)||\mathbf{x} - \mathbf{y}||_\gamma < \frac{2C\eta}{\alpha}||\mathbf{x} - \mathbf{y}||_\gamma < \frac{1}{3}||\mathbf{x} - \mathbf{y}||_\gamma. \end{aligned}$$

It follows from Lemma 15.4.4, for each $\mathbf{a}_-$ such that $|\mathbf{a}_-| < \frac{\delta}{2C}$, there exists a unique solution to 16.38 in E_γ.

As pointed out earlier, if

$$\boldsymbol{\psi}(\mathbf{a}) \equiv -\int_0^\infty \Phi_+(-s)\mathbf{g}_+(\mathbf{x}(s,\mathbf{a}))\,ds$$

then for $\mathbf{x}(t,\mathbf{x}_0)$ the solution to the initial value problem

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}),\ \mathbf{x}(0) = \mathbf{x}_0$$

has the property that if $\mathbf{x}_0$ is not of the form $\begin{pmatrix} \mathbf{a}_- \\ \boldsymbol{\psi}(\mathbf{a}_-) \end{pmatrix}$, then $|\mathbf{x}(t,\mathbf{x}_0)|$ cannot be less than δ for all $t > 0$.

On the other hand, if $\mathbf{x}_0 = \begin{pmatrix} \mathbf{a}_- \\ \boldsymbol{\psi}(\mathbf{a}_-) \end{pmatrix}$ for $|\mathbf{a}_-| < \frac{\delta}{2C}$, then $\mathbf{x}(t,\mathbf{x}_0)$,the solution to 16.38 is the unique solution to the initial value problem

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}),\ \mathbf{x}(0) = \mathbf{x}_0.$$

and it was shown that $||\mathbf{x}(\cdot,\mathbf{x}_0)||_\gamma < \delta$ and so in fact,

$$|\mathbf{x}(t,\mathbf{x}_0)| \leq \delta e^{-\gamma t}$$

showing that

$$\lim_{t\to\infty} \mathbf{x}(t,\mathbf{x}_0) = \mathbf{0}.$$

This proves the Lemma.

The following theorem is the main result. It involves a use of linear algebra and the above lemma.

Theorem 16.7.3 *Consider the initial value problem for the almost linear system*

$$\mathbf{x}' = A\mathbf{x} + \mathbf{g}(\mathbf{x}),\ \mathbf{x}(0) = \mathbf{x}_0$$

in which $\mathbf{g}$ *is* C^1 *and where at there are* $k < n$ *eigenvalues of* A *which have negative real parts and* $n-k$ *eigenvalues of* A *which have positive real parts. Then* $\mathbf{0}$ *is not stable. More precisely, there exists a set of points* $(\mathbf{a}, \boldsymbol{\psi}(\mathbf{a}))$ *for* $\mathbf{a}$ *small and in a* k *dimensional subspace such that for* $\mathbf{x}_0$ *on this set,*

$$\lim_{t\to\infty} \mathbf{x}(t,\mathbf{x}_0) = \mathbf{0}$$

and for $\mathbf{x}_0$ *not on this set, there exists a* $\delta > 0$ *such that* $|\mathbf{x}(t,\mathbf{x}_0)|$ *cannot remain less than* δ *for all positive* t.

Proof: This involves nothing more than a reduction to the situation of Lemma 16.7.2. From Corollary 11.4.4 on Page 11.4.4 A is similar to a matrix of the form described in Lemma 16.7.2. Thus $A = S^{-1}\begin{pmatrix} A_- & 0 \\ 0 & A_+ \end{pmatrix} S$. Letting $\mathbf{y} = S\mathbf{x}$, it follows

$$\mathbf{y}' = \begin{pmatrix} A_- & 0 \\ 0 & A_+ \end{pmatrix}\mathbf{y} + \mathbf{g}\left(S^{-1}\mathbf{y}\right)$$

Now $|\mathbf{x}| = \left|S^{-1}S\mathbf{x}\right| \leq \left|\left|S^{-1}\right|\right| |\mathbf{y}|$ and $|\mathbf{y}| = \left|SS^{-1}\mathbf{y}\right| \leq ||S||\,|\mathbf{x}|$. Therefore,

$$\frac{1}{||S||}|\mathbf{y}| \leq |\mathbf{x}| \leq \left|\left|S^{-1}\right|\right| |\mathbf{y}|.$$

It follows all conclusions of Lemma 16.7.2 are valid for this theorem. This proves the theorem.

The set of points $(\mathbf{a}, \boldsymbol{\psi}(\mathbf{a}))$ for $\mathbf{a}$ small is called the stable manifold. Much more can be said about the stable manifold and you should look at a good differential equations book for this.

The Fundamental Theorem Of Algebra

The fundamental theorem of algebra states that every non constant polynomial having coefficients in $\mathbb{C}$ has a zero in $\mathbb{C}$. If $\mathbb{C}$ is replaced by $\mathbb{R}$, this is not true because of the example, $x^2+1=0$. This theorem is a very remarkable result and notwithstanding its title, all the best proofs of it depend on either analysis or topology. It was first proved by Gauss in 1797. The proof given here follows Rudin [11]. See also Hardy [7] for another proof, more discussion and references. Recall De Moivre's theorem on Page 10 which is listed below for convenience.

Theorem A.0.4 *Let $r>0$ be given. Then if n is a positive integer,*

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

Now from this theorem, the following corollary on Page 1.2.5 is obtained.

Corollary A.0.5 *Let z be a non zero complex number and let k be a positive integer. Then there are always exactly k k^{th} roots of z in $\mathbb{C}$.*

Lemma A.0.6 *Let $a_k\in\mathbb{C}$ for $k=1,\cdots,n$ and let $p(z)\equiv\sum_{k=1}^n a_k z^k$. Then p is continuous.*

Proof:

$$|az^n-aw^n|\le|a||z-w|\left|z^{n-1}+z^{n-2}w+\cdots+w^{n-1}\right|.$$

Then for $|z-w|<1$, the triangle inequality implies $|w|<1+|z|$ and so if $|z-w|<1$,

$$|az^n-aw^n|\le|a||z-w|\,n(1+|z|)^n.$$

If $\varepsilon>0$ is given, let

$$\delta<\min\left(1,\frac{\varepsilon}{|a|\,n(1+|z|)^n}\right).$$

It follows from the above inequality that for $|z-w|<\delta$, $|az^n-aw^n|<\varepsilon$. The function of the lemma is just the sum of functions of this sort and so it follows that it is also continuous.

Theorem A.0.7 *(Fundamental theorem of Algebra) Let $p(z)$ be a nonconstant polynomial. Then there exists $z\in\mathbb{C}$ such that $p(z)=0$.*

Proof: Suppose not. Then

$$p(z)=\sum_{k=0}^n a_k z^k$$

where $a_n \neq 0,\ n > 0$. Then

$$|p(z)| \geq |a_n| |z|^n - \sum_{k=0}^{n-1} |a_k| |z|^k$$

and so

$$\lim_{|z|\to\infty} |p(z)| = \infty. \tag{1.1}$$

Now let

$$\lambda \equiv \inf \{|p(z)| : z \in \mathbb{C}\}.$$

By 1.1, there exists an $R > 0$ such that if $|z| > R$, it follows that $|p(z)| > \lambda + 1$. Therefore,

$$\lambda \equiv \inf \{|p(z)| : z \in \mathbb{C}\} = \inf \{|p(z)| : |z| \leq R\}.$$

The set $\{z : |z| \leq R\}$ is a closed and bounded set and so this infimum is achieved at some point w with $|w| \leq R$. A contradiction is obtained if $|p(w)| = 0$ so assume $|p(w)| > 0$. Then consider

$$q(z) \equiv \frac{p(z+w)}{p(w)}.$$

It follows $q(z)$ is of the form

$$q(z) = 1 + c_k z^k + \cdots + c_n z^n$$

where $c_k \neq 0$, because $q(0) = 1$. It is also true that $|q(z)| \geq 1$ by the assumption that $|p(w)|$ is the smallest value of $|p(z)|$. Now let $\theta \in \mathbb{C}$ be a complex number with $|\theta| = 1$ and

$$\theta c_k w^k = -|w|^k |c_k|.$$

If

$$w \neq 0, \theta = \frac{-|w^k| |c_k|}{w^k c_k}$$

and if $w = 0$, $\theta = 1$ will work. Now let $\eta^k = \theta$ and let t be a small positive number.

$$q(t\eta w) \equiv 1 - t^k |w|^k |c_k| + \cdots + c_n t^n (\eta w)^n$$

which is of the form

$$1 - t^k |w|^k |c_k| + t^k (g(t,w))$$

where $\lim_{t\to 0} g(t,w) = 0$. Letting t be small enough,

$$|g(t,w)| < |w|^k |c_k| /2$$

and so for such t,

$$|q(t\eta w)| < 1 - t^k |w|^k |c_k| + t^k |w|^k |c_k| /2 < 1,$$

a contradiction to $|q(z)| \geq 1$. This proves the theorem.

Bibliography

[1] **Apostol T.,** *Calculus Volume II Second edition,* Wiley *1969.*

[2] **Baker, Roger**, *Linear Algebra*, Rinton Press 2001.

[3] **Coddington and Levinson**, *Theory of Ordinary Differential Equations* McGraw Hill 1955.

[4] **Davis H. and Snider A.,** *Vector Analysis* Wm. C. Brown 1995.

[5] **Edwards C.H.,** *Advanced Calculus of several Variables,* Dover 1994.

[6] **Gurtin M.,** *An introduction to continuum mechanics,* Academic press 1981.

[7] **Hardy G.,** *A Course Of Pure Mathematics, Tenth edition,* Cambridge University Press 1992.

[8] **Horn R. and Johnson C.,** *matrix Analysis,* Cambridge University Press, 1985.

[9] **Karlin S. and Taylor H.,** *A First Course in Stochastic Processes,* Academic Press, 1975.

[10] **Nobel B. and Daniel J.,** *Applied Linear Algebra, Prentice Hall, 1977.*

[11] **Rudin W.,** *Principles of Mathematical Analysis*, McGraw Hill, 1976.

[12] **Salas S. and Hille E.,** *Calculus One and Several Variables,* Wiley 1990.

[13] **Strang Gilbert**, *Linear Algebra and its Applications,* Harcourt Brace Jovanovich 1980.

Index