Elementary Linear Algebra

Kuttler

June 29, 2023

CONTENTS

Preface

This is an introduction to linear algebra. The main part of the book features row operations and everything is done in terms of the row reduced echelon form and specific algorithms. At the end, the more abstract notions of vector spaces and linear transformations on vector spaces are presented. However, this is intended to be a first course in linear algebra for students who are sophomores or juniors who have had a course in one variable calculus and a reasonable background in college algebra. I have given complete proofs of all the fundamental ideas, but some topics such as Markov matrices are not complete in this book but receive a plausible introduction. The book contains a complete treatment of determinants and a simple proof of the Cayley Hamilton theorem although these are optional topics. The Jordan form is presented as an appendix. I see this theorem as the beginning of more advanced topics in linear algebra and not really part of a beginning linear algebra course. There are extensions of many of the topics of this book in my on line book [13]. I have also not emphasized that linear algebra can be carried out with any field although there is an optional section on this topic, most of the book being devoted to either the real numbers or the complex numbers. It seems to me this is a reasonable specialization for a first course in linear algebra.

Should Linear algebra be encountered for the first time before Calculus or after Calculus? I think good arguments can be given either way. This book will work for either order. If the book is used for a course on linear algebra which comes after multi-variable calculus, one can skip the two chapters preceding Chapter 4 and if the book is used for a course on linear algebra which comes before multi-variable calculus, it would be a good idea to include these two chapters and end no later than Chapter 13.

Linear algebra is a wonderful interesting subject. It is a shame when it degenerates into nothing more than a challenge to do the arithmetic correctly. It seems to me that the use of a computer algebra system can be a great help in avoiding this sort of tedium. I don't want to over emphasize the use of technology, which is easy to do if you are not careful, but there are certain standard things which are best done by the computer. Some of these include the row reduced echelon form, *PLU* factorization, and *QR* factorization. It is much more fun to let the machine do the tedious calculations than to suffer with them yourself. However, it is not good when the use of the computer algebra system degenerates into simply asking it for the answer without understanding what the oracular software is doing. With this in mind, there are a few interactive links which explain how to use a computer algebra system to accomplish some of these more tedious standard tasks. These are obtained by clicking on the symbol ►. I have included how to do it using maple and scientific notebook because these are the two systems I am familiar with and have on my computer. Also, I have included the very easy to use matrix calculator which is available on the web and have given directions for MATLAB at the end of relevant chapters. Other systems could be featured as well. It is expected that people will use such computer algebra systems to do the exercises in this book whenever it would be helpful to do so, rather than wasting huge amounts of time doing computations by hand. However, this is not a book on numerical analysis so no effort is made to consider many important numerical analysis issues.

I appreciate those who have found errors and needed corrections over the years that this has been available.

There is a pdf file of this book on my web page and a more advanced linear algebra book. This book, as well as the more advanced text, is also available as an electronic version at
http://www.saylor.org/archivedcourses/ma211/ where it is used as an open access textbook. In addition, it is available for free at BookBoon under their linear algebra offerings.

Chapter 1

Some Prerequisite Topics

The reader should be familiar with most of the topics in this chapter. However, it is often the case that set notation is not familiar and so a short discussion of this is included first. Many of the applications of linear algebra require the use of complex numbers, so this is the reason for the discussion of complex numbers.

1.1 Sets And Set Notation

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

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

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

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

Be sure you understand that something which is in both A and B is in the union. It is not an exclusive or.

The intersection of two sets, A and B consists of everything which is in both of the sets. Thus $\{1,2,3,8\} \cap \{3,4,7,8\} = \{3,8\}$ because 3 and 8 are those elements the two sets have in common. In general,

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

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

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

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

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

Set notation is used whenever convenient.

To illustrate the use of this notation relative to intervals consider three examples of inequalities. Their solutions will be written in the notation just described.

Example 1.1.1 *Solve the inequality* $2x+4 \leq x-8$

$x \leq -12$ is the answer. This is written in terms of an interval as $(-\infty,-12]$.

Example 1.1.2 *Solve the inequality* $(x+1)(2x-3) \geq 0$.

The solution is $x \leq -1$ or $x \geq \frac{3}{2}$. In terms of set notation this is denoted by $(-\infty,-1] \cup [\frac{3}{2},\infty)$.

Example 1.1.3 *Solve the inequality* $x(x+2) \geq -4$.

This is true for any value of x. It is written as $\mathbb{R}$ or $(-\infty,\infty)$.

1.2 Well Ordering And Induction

Mathematical induction and well ordering are two extremely important principles in math. They are often used to prove significant things which would be hard to prove otherwise.Typically these are theorems about integers.

Definition 1.2.1 *A set is well ordered if every nonempty subset S, contains a smallest element z having the property that $z \leq x$ for all $x \in S$.*

Axiom 1.2.2 *Any set of integers larger than a given number is well ordered.*

In particular, the natural numbers defined as

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

is well ordered.

The above axiom implies the principle of mathematical induction. The symbol $\mathbb{Z}$ denotes the set of all integers. Note that if a is an integer, then there are no integers between a and $a+1$.

Theorem 1.2.3 *(Mathematical induction) A set $S \subseteq \mathbb{Z}$, having the property that $a \in S$ and $n+1 \in S$ whenever $n \in S$ contains all integers $x \in \mathbb{Z}$ such that $x \geq a$.*

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

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

Example 1.2.4 *Prove by induction that* $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$.

►

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

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

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

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

This equals

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

and

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

Therefore,

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

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

Example 1.2.5 *Show that for all $n \in \mathbb{N}$, $\frac{1}{2} \cdot \frac{3}{4} \cdots \frac{2n-1}{2n} < \frac{1}{\sqrt{2n+1}}$.*

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

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

The theorem will be proved if this last expression is less than $\frac{1}{\sqrt{2n+3}}$. This happens if and only if

$$\left(\frac{1}{\sqrt{2n+3}} \right)^2 = \frac{1}{2n+3} > \frac{2n+1}{(2n+2)^2}$$

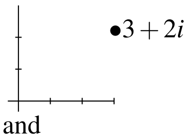
which occurs if and only if $(2n+2)^2 > (2n+3)(2n+1)$ and this is clearly true which may be seen from expanding both sides. This proves the inequality.

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

1.3 The Complex Numbers

Recall that a real number is a point on the real number line. Just as a real number should be considered as a point on the line, a complex number is considered a point in the plane which can be identified in the usual way using the Cartesian coordinates of the point. Thus (a, b) identifies a point whose x coordinate is a and whose y coordinate is b. In dealing with complex numbers, such a point is written as $a+ib$. For example, in the following picture, I have graphed the point $3+2i$. You see it corresponds to the point in the plane whose coordinates are

$$(3,2).$$



Multiplication and addition are defined in the most obvious way subject to the convention that $i^2 = -1$. Thus,

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

and

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

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

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

You should prove the following theorem.

Theorem 1.3.1 *The complex numbers with multiplication and addition defined as above form a field satisfying all the field axioms. These are the following list of properties.*

1. $x+y=y+x$, *(commutative law for addition)*
2. $x+0=x$, *(additive identity).*
3. *For each* $x\in\mathbb{R}$, *there exists* $-x\in\mathbb{R}$ *such that* $x+(-x)=0$, *(existence of additive inverse).*
4. $(x+y)+z=x+(y+z)$, *(associative law for addition).*
5. $xy=yx$, *(commutative law for multiplication). You could write this as* $x\times y=y\times x$.
6. $(xy)z=x(yz)$, *(associative law for multiplication).*
7. $1x=x$, *(multiplicative identity).*
8. *For each* $x\neq 0$, *there exists* x^{-1} *such that* $xx^{-1}=1$.*(existence of multiplicative inverse).*
9. $x(y+z)=xy+xz$.*(distributive law).*

Something which satisfies these axioms is called a field. Linear algebra is all about fields, although in this book, the field of most interest will be the field of complex numbers or the field of real numbers. You have seen in earlier courses that the real numbers also satisfies the above axioms. The field of complex numbers is denoted as $\mathbb{C}$ and the field of real numbers is denoted as $\mathbb{R}$. In general, we usually assume $1\neq 0$ since otherwise, the "field" is of little interest.

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.

$$\begin{aligned}\left(\overline{a+ib}\right)(a+ib) &= (a-ib)(a+ib)\\ &= a^2+b^2-i(ab-ab)=a^2+b^2.\end{aligned}$$

Definition 1.3.2 *Define the absolute value of a complex number as follows.*

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

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

$$|z|=(z\bar{z})^{1/2}.$$

Also from the definition, if $z=x+iy$ and $w=u+iv$ are two complex numbers, then $|zw|=|z|\,|w|$. You should verify this. ▶

Notation 1.3.3 *Recall the following notation.*

$$\sum_{j=1}^{n} a_j \equiv a_1 + \cdots + a_n$$

There is also a notation which is used to denote a product.

$$\prod_{j=1}^{n} a_j \equiv a_1 a_2 \cdots a_n$$

The triangle inequality holds for the absolute value for complex numbers just as it does for the ordinary absolute value.

Proposition 1.3.4 *Let z,w be complex numbers. Then the triangle inequality holds.*

$$|z+w| \leq |z| + |w|, \ \ ||z| - |w|| \leq |z-w|.$$

Proof: Let $z = x + iy$ and $w = u + iv$. First note that

$$z\overline{w} = (x+iy)(u-iv) = xu + yv + i(yu - xv)$$

and so $|xu + yv| \leq |z\overline{w}| = |z||w|$.

$$|z+w|^2 = (x+u+i(y+v))(x+u-i(y+v))$$
$$= (x+u)^2 + (y+v)^2 = x^2 + u^2 + 2xu + 2yv + y^2 + v^2$$
$$\leq |z|^2 + |w|^2 + 2|z||w| = (|z| + |w|)^2,$$

so this shows the first version of the triangle inequality. To get the second,

$$z = z - w + w, \ w = w - z + z$$

and so by the first form of the inequality

$$|z| \leq |z-w| + |w|, \ |w| \leq |z-w| + |z|$$

and so both $|z| - |w|$ and $|w| - |z|$ are no larger than $|z-w|$ and this proves the second version because $||z| - |w||$ is one of $|z| - |w|$ or $|w| - |z|$. ■

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

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

1.4 Polar Form Of Complex Numbers

Complex numbers, are often written in the so called polar form which is described next. Suppose $z = x + iy$ 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

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

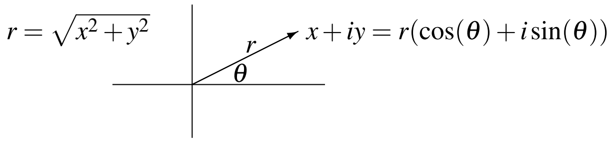
and so

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

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 $r(\cos\theta + i\sin\theta)$ where θ is this angle just described and $r = \sqrt{x^2 + y^2} \equiv |z|$.



1.5 Roots Of Complex Numbers

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

Theorem 1.5.1 *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 1.5.2 *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 1.5.3 *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

$$\cos\left(\frac{\pi}{6}\right) + i\sin\left(\frac{\pi}{6}\right), \cos\left(\frac{5}{6}\pi\right) + i\sin\left(\frac{5}{6}\pi\right), \cos\left(\frac{3}{2}\pi\right) + i\sin\left(\frac{3}{2}\pi\right).$$

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 1.5.4 *Factor the polynomial $x^3 - 27$.*

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

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

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

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

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

Note that even though the polynomial x^3-27 has all real coefficients, it has some complex zeros, $\frac{-1}{2}+i\frac{\sqrt{3}}{2}$ and $\frac{-1}{2}-i\frac{\sqrt{3}}{2}$. These zeros are complex conjugates of each other. It is **always** this way. You should show this is the case. To see how to do this, see Problems 17 and 18 below.

Another fact for your information is the fundamental theorem of algebra. This theorem says that any polynomial of degree at least 1 having any complex coefficients always has a root in $\mathbb{C}$. This is sometimes referred to by saying $\mathbb{C}$ is algebraically complete. Gauss is usually credited with giving a proof of this theorem in 1797 but many others worked on it and the first completely correct proof was due to Argand in 1806. For more on this theorem, you can google fundamental theorem of algebra and look at the interesting Wikipedia article on it. Proofs of this theorem usually involve the use of techniques from calculus even though it is really a result in algebra. A proof and plausibility explanation is given later.

1.6 The Quadratic Formula

The quadratic formula

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

gives the solutions x to

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

where a,b,c are real numbers. It holds even if $b^2-4ac<0$. This is easy to show from the above. There are exactly two square roots to this number b^2-4ac from the above methods using De Moivre's theorem. These roots are of the form

$$\sqrt{4ac-b^2}\left(\cos\left(\frac{\pi}{2}\right)+i\sin\left(\frac{\pi}{2}\right)\right)=i\sqrt{4ac-b^2}$$

and

$$\sqrt{4ac-b^2}\left(\cos\left(\frac{3\pi}{2}\right)+i\sin\left(\frac{3\pi}{2}\right)\right)=-i\sqrt{4ac-b^2}$$

Thus the solutions, according to the quadratic formula are still given correctly by the above formula.

Do these solutions predicted by the quadratic formula continue to solve the quadratic equation? Yes, they do. You only need to observe that when you square a square root of a complex number z, you recover z. Thus

$$a\left(\frac{-b+\sqrt{b^2-4ac}}{2a}\right)^2+b\left(\frac{-b+\sqrt{b^2-4ac}}{2a}\right)+c$$

$$= a\left(\frac{1}{2a^2}b^2 - \frac{1}{a}c - \frac{1}{2a^2}b\sqrt{b^2-4ac}\right) + b\left(\frac{-b+\sqrt{b^2-4ac}}{2a}\right) + c$$

$$= \left(-\frac{1}{2a}\left(b\sqrt{b^2-4ac} + 2ac - b^2\right)\right) + \frac{1}{2a}\left(b\sqrt{b^2-4ac} - b^2\right) + c = 0$$

Similar reasoning shows directly that $\frac{-b-\sqrt{b^2-4ac}}{2a}$ also solves the quadratic equation.

What if the coefficients of the quadratic equation are actually complex numbers? Does the formula hold even in this case? The answer is yes. This is a hint on how to do Problem 27 below, a special case of the fundamental theorem of algebra, and an ingredient in the proof of some versions of this theorem.

Example 1.6.1 *Find the solutions to $x^2 - 2ix - 5 = 0$.*

Formally, from the quadratic formula, these solutions are

$$x = \frac{2i \pm \sqrt{-4+20}}{2} = \frac{2i \pm 4}{2} = i \pm 2.$$

Now you can check that these really do solve the equation. In general, this will be the case. See Problem 27 below.

1.7 The Complex Exponential

It was shown above that every complex number can be written in the form

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

where $r \geq 0$. Laying aside the zero complex number, this shows that every non zero complex number is of the form $e^{\alpha}(\cos\beta + i\sin\beta)$. We write this in the form $e^{\alpha+i\beta}$. Having done so, does it follow that the expression preserves the most important property of the function $t \to e^{(\alpha+i\beta)t}$ for t real, that

$$\left(e^{(\alpha+i\beta)t}\right)' = (\alpha+i\beta)e^{(\alpha+i\beta)t}?$$

By the definition just given which does not contradict the usual definition in case $\beta = 0$ and the usual rules of differentiation in calculus,

$$\begin{aligned}\left(e^{(\alpha+i\beta)t}\right)' &= \left(e^{\alpha t}(\cos(\beta t) + i\sin(\beta t))\right)' \\ &= e^{\alpha t}\left[\alpha(\cos(\beta t) + i\sin(\beta t)) + (-\beta\sin(\beta t) + i\beta\cos(\beta t))\right]\end{aligned}$$

Now consider the other side. From the definition it equals

$$(\alpha+i\beta)\left(e^{\alpha t}(\cos(\beta t) + i\sin(\beta t))\right) = e^{\alpha t}\left[(\alpha+i\beta)(\cos(\beta t) + i\sin(\beta t))\right]$$
$$= e^{\alpha t}\left[\alpha(\cos(\beta t) + i\sin(\beta t)) + (-\beta\sin(\beta t) + i\beta\cos(\beta t))\right]$$

which is the same thing. This is of fundamental importance in differential equations. It shows that there is no change in going from real to complex numbers for ω in the consideration of the problem $y' = \omega y,\ y(0) = 1$. The solution is always $e^{\omega t}$. The formula just discussed, that

$$e^{\alpha}(\cos\beta + i\sin\beta) = e^{\alpha+i\beta}$$

is Euler's formula.

1.8 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 most straightforward proofs depend on either analysis or topology. It was first mostly proved by Gauss in 1797. The first complete proof was given by Argand in 1806. The proof given here follows Rudin [15]. See also Hardy [9] for a similar proof, more discussion and references. The shortest proof is found in the theory of complex analysis. First I will give an informal explanation of this theorem which shows why it is is reasonable to believe in the fundamental theorem of algebra.

Theorem 1.8.1 *Let $p(z)=a_nz^n+a_{n-1}z^{n-1}+\cdots+a_1z+a_0$ where each a_k is a complex number and $a_n\neq 0, n\geq 1$. Then there exists $w\in\mathbb{C}$ such that $p(w)=0$.*

To begin with, here is the informal explanation. Dividing by the leading coefficient a_n, there is no loss of generality in assuming that the polynomial is of the form

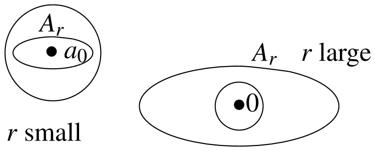
$$p(z)=z^n+a_{n-1}z^{n-1}+\cdots+a_1z+a_0$$

If $a_0=0$, there is nothing to prove because $p(0)=0$. Therefore, assume $a_0\neq 0$. From the polar form of a complex number z, it can be written as $|z|(\cos\theta+i\sin\theta)$. Thus, by DeMoivre's theorem,

$$z^n=|z|^n(\cos(n\theta)+i\sin(n\theta))$$

It follows that z^n is some point on the circle of radius $|z|^n$

Denote by C_r the circle of radius r in the complex plane which is centered at 0. Then if r is sufficiently large and $|z|=r$, the term z^n is far larger than the rest of the polynomial. It is on the circle of radius $|z|^n$ while the other terms are on circles of fixed multiples of $|z|^k$ for $k\leq n-1$. Thus, for r large enough, $A_r=\{p(z):z\in C_r\}$ describes a closed curve which misses the inside of some circle having 0 as its center. It won't be as simple as suggested in the following picture, but it will be a closed curve thanks to De Moivre's theorem and the observation that the cosine and sine are periodic. Now shrink r. Eventually, for r small enough, the non constant terms are negligible and so A_r is a curve which is contained in some circle centered at a_0 which has 0 on the outside. Thus it is reasonable to believe that for some r during this shrinking process, the set A_r must hit 0. It follows that $p(z)=0$ for some z.



For example, consider the polynomial $x^3+x+1+i$. It has no real zeros. However, you could let $z=r(\cos t+i\sin t)$ and insert this into the polynomial. Thus you would want to find a point where

$$(r(\cos t+i\sin t))^3+r(\cos t+i\sin t)+1+i=0+0i$$

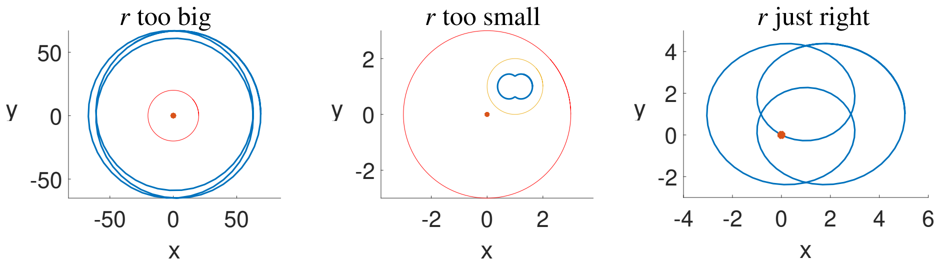
Expanding this expression on the left to write it in terms of real and imaginary parts, you get on the left

$$r^3\cos^3 t-3r^3\cos t\sin^2 t+r\cos t+1+i\left(3r^3\cos^2 t\sin t-r^3\sin^3 t+r\sin t+1\right)$$

Thus you need to have both the real and imaginary parts equal to 0. In other words, you need to have

$$\left(r^3\cos^3 t-3r^3\cos t\sin^2 t+r\cos t+1, 3r^3\cos^2 t\sin t-r^3\sin^3 t+r\sin t+1\right)=(0,0)$$

for some value of r and t. First here is a graph of this parametric function of t for $t \in [0,2\pi]$ on the left, when $r = 4$. Note how the graph misses the origin $0+i0$. In fact, the closed curve surrounds a small circle which has the point $0+i0$ on its inside.



Next is the graph when $r = .5$. Note how the closed curve is included in a circle which has $0+i0$ on its outside. As you shrink r you get closed curves. At first, these closed curves enclose $0+i0$ and later, they exclude $0+i0$. Thus one of them should pass through this point. In fact, consider the curve which results when $r = 1.386$ which is the graph on the right. Note how for this value of r the curve passes through the point $0+i0$. Thus for some t, $1.3862(\cos t + i \sin t)$ is a solution of the equation $p(z) = 0$.

Now here is a rigorous proof for those who have studied analysis. It depends on the extreme value theorem from calculus applied to the continuous function $f(x,y) \equiv |p(x+iy)|$.

Proof: Suppose the nonconstant polynomial $p(z) = a_0 + a_1 z + \cdots + a_n z^n, a_n \neq 0$, has no zero in $\mathbb{C}$. Since $\lim_{|z|\to\infty} |p(z)| = \infty$, there is a z_0 with

$$|p(z_0)| = \min_{z\in\mathbb{C}} |p(z)| > 0$$

Then let $q(z) = \frac{p(z+z_0)}{p(z_0)}$. This is also a polynomial which has no zeros and the minimum of $|q(z)|$ is 1 and occurs at $z = 0$. Since $q(0) = 1$, it follows $q(z) = 1 + a_k z^k + r(z)$ where $r(z)$ is of the form

$$r(z) = a_m z^m + a_{m+1} z^{m+1} + \ldots + a_n z^n \text{ for } m > k.$$

Choose a sequence, $z_n \to 0$, such that $a_k z_n^k < 0$. For example, let $-a_k z_n^k = (1/n)$ so $z_n = (-a_k)^{1/k} \left(\frac{1}{n}\right)^{1/k}$ and Then

$$\begin{aligned} |q(z_n)| &= \left|1 + a_k z^k + r(z)\right| \leq 1 - 1/n + |r(z_n)| \\ &\leq 1 - \frac{1}{n} + \frac{1}{n} \sum_{j=m}^{n} |a_j| |a_k|^{1/k} \left(\frac{1}{n}\right)^{(j-k)/k} < 1 \end{aligned}$$

for all n large enough because the sum is smaller than 1 whenever n is large enough, showing $|q(z_n)| < 1$ whenever n is large enough. This is a contradiction to $|q(z)| \geq 1$. ■

At this point, you could skip Chapters 2 and 3 and go directly to Chapter 4 if desired.

1.9 Exercises

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

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

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

4. The binomial theorem states $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$ where
$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \text{ if } k \in [1,n]\,,\ \binom{n}{0} \equiv 1 \equiv \binom{n}{n}$$
Prove the binomial theorem by induction. Next show that
$$\binom{n}{k} = \frac{n!}{(n-k)!k!},\ 0! \equiv 1$$
▶

5. Let $z = 5+i9$. Find z^{-1}.

6. Let $z = 2+i7$ and let $w = 3-i8$. Find $zw, z+w, z^2$, and w/z.

7. Give the complete solution to $x^4+16 = 0$.

8. Graph the complex cube roots of 8 in the complex plane. Do the same for the four fourth roots of 16. ▶

9. If z is a complex number, show there exists ω a complex number with $|\omega| = 1$ and $\omega z = |z|$.

10. 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. ▶

11. 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 $\cos(5x)$. ▶

12. 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, $r = |z|$.

13. Factor x^3+8 as a product of linear factors.

14. 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.

15. Completely factor x^4+16 as a product of linear factors.

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

17. If z, w are complex numbers prove $\overline{zw} = \overline{z}\overline{w}$ and then show by induction that $\overline{\prod_{j=1}^{n} z_j} = \prod_{j=1}^{n} \overline{z_j}$. 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.

18. 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 $p(\bar{z}) = 0$ also.

19. Show that $1+i, 2+i$ are the only two zeros to
$$p(x) = x^2 - (3+2i)x + (1+3i)$$
so the zeros do not necessarily come in conjugate pairs if the coefficients are not real.

20. 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?

21. 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?

22. Review Problem 10 at this point. Now here is another question: If n is an integer, is it always true that $(\cos\theta - i\sin\theta)^n = \cos(n\theta) - i\sin(n\theta)$? Explain.

23. Suppose you have any polynomial in $\cos\theta$ and $\sin\theta$. By this I mean an expression of the form $\sum_{\alpha=0}^{m}\sum_{\beta=0}^{n} a_{\alpha\beta}\cos^{\alpha}\theta\sin^{\beta}\theta$ where $a_{\alpha\beta} \in \mathbb{C}$. Can this always be written in the form $\sum_{\gamma=-(n+m)}^{m+n} b_\gamma \cos\gamma\theta + \sum_{\tau=-(n+m)}^{n+m} c_\tau \sin\tau\theta$? Explain.

24. Suppose $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ is a polynomial and it has n zeros,
$$z_1, z_2, \cdots, z_n$$
listed according to multiplicity. (z is a root of multiplicity m if the polynomial $f(x) = (x-z)^m$ divides $p(x)$ but $(x-z)f(x)$ does not.) Show that
$$p(x) = a_n(x-z_1)(x-z_2)\cdots(x-z_n).$$

25. Give the solutions to the following quadratic equations having real coefficients.

(a) $x^2 - 2x + 2 = 0$

(b) $3x^2 + x + 3 = 0$

(c) $x^2 - 6x + 13 = 0$

(d) $x^2 + 4x + 9 = 0$

(e) $4x^2 + 4x + 5 = 0$

26. Give the solutions to the following quadratic equations having complex coefficients. Note how the solutions do not come in conjugate pairs as they do when the equation has real coefficients.

    (a) $x^2+2x+1+i=0$

    (b) $4x^2+4ix-5=0$

    (c) $4x^2+(4+4i)x+1+2i=0$

    (d) $x^2-4ix-5=0$

    (e) $3x^2+(1-i)x+3i=0$

27. Prove the fundamental theorem of algebra for quadratic polynomials having coefficients in $\mathbb{C}$. That is, show that an equation of the form $ax^2+bx+c=0$ where a,b,c are complex numbers, $a\neq 0$ has a complex solution. **Hint:** Consider the fact, noted earlier that the expressions given from the quadratic formula do in fact serve as solutions.

Chapter 2

$\mathbb{F}^n$

One can skip this chapter and the next and go directly to Chapter 4 if desired. This material is here because in this book the main emphasis in the first part of the book is on vectors in $\mathbb{R}^n$ and so this particular vector space is discussed. However, if the reader has already seen these things, for example in a calculus course, there is absolutely no harm in skipping ahead.

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 2.0.1 *Define* $\mathbb{F}^n \equiv \left\{(x_1,\cdots,x_n) : x_j \in \mathbb{F} \text{ for } j = 1,\cdots,n\right\}$.

$$(x_1,\cdots,x_n) = (y_1,\cdots,y_n)$$

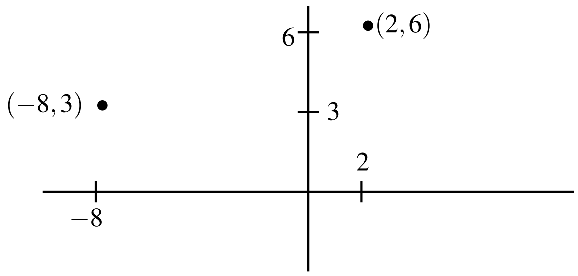
if and only if for all $j = 1,\cdots,n$, $x_j = y_j$. *When* $(x_1,\cdots,x_n) \in \mathbb{F}^n$, *it is conventional to denote* $(x_1,\cdots,x_n)$ *by the single bold face letter,* $\mathbf{x}$. *The numbers,* x_j *are called the coordinates. Elements in* $\mathbb{F}^n$ *are called* ***vectors.*** *The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

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

for t *in the* i^{th} *slot is called the* i^{th} *coordinate axis in the case of* $\mathbb{R}^n$. *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 pre-calculus. Here is a short review. First consider the case when $n = 1$. Then from the definition, $\mathbb{R}^1 = \mathbb{R}$. Recall that $\mathbb{R}$ is identified with the points of a line. Look at the number line again. Observe that this amounts to identifying a point on this line with a real number. In other words a real number determines where you are on this line. Now suppose $n = 2$ and consider two lines which intersect each other at right angles as shown in the following picture.



Notice how you can identify a point shown in the plane with the ordered pair, $(2,6)$. You go to the right a distance of 2 and then up a distance of 6. Similarly, you can identify another point in the plane with the ordered pair $(-8,3)$. Starting at 0, go to the left a distance of 8 on the horizontal line and then up a distance of 3. The reason you go to the left is that there is a $-$ sign on the eight. From this reasoning, every ordered pair determines a unique point in the plane. Conversely, taking a point in the plane, you could draw two lines through the point, one vertical and the other horizontal and determine unique points, x_1 on the horizontal line in the above picture and x_2 on the vertical line in the above picture, such that the point of interest is identified with the ordered pair, (x_1,x_2). In short, points in the plane can be identified with ordered pairs similar to the way that points on the real line are identified with real numbers. Now suppose $n = 3$. As just explained, the first two coordinates determine a point in a plane. Letting the third component determine how far up or down you go, depending on whether this number is positive or negative, this determines a point in space. Thus, $(1,4,-5)$ would mean to determine the point in the plane that goes with $(1,4)$ and then to go below this plane a distance of 5 to obtain a unique point in space. You see that the ordered triples correspond to points in space just as the ordered pairs correspond to points in a plane and single real numbers correspond to points on a line.

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

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

[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.

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

2.1 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 2.1.1 *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*

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

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

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

With this definition, vector addition and scalar multiplication satisfy the conclusions of the following theorem. More generally, these properties are called the **vector space axioms.**

Theorem 2.1.2 *For $\mathbf{v}, \mathbf{w} \in \mathbb{F}^n$ and α, β scalars, (real numbers), the following hold.*

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

the commutative law of addition,

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

the associative law for addition,

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

the existence of an additive identity,

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

the existence of an additive inverse, Also

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

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

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

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

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

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

$$\alpha(\mathbf{v}+\mathbf{w}) = \alpha(v_1 + w_1, \cdots, v_n + w_n)$$

$$= (\alpha(v_1 + w_1), \cdots, \alpha(v_n + w_n)) = (\alpha v_1 + \alpha w_1, \cdots, \alpha v_n + \alpha w_n)$$

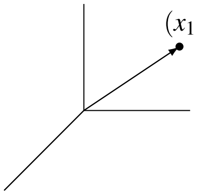
$$= (\alpha v_1, \cdots, \alpha v_n) + (\alpha w_1, \cdots, \alpha w_n) = \alpha\mathbf{v} + \alpha\mathbf{w}.$$

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

2.2 Geometric Meaning Of Vectors

The geometric meaning is especially significant in the case of $\mathbb{R}^n$ for $n = 2,3$. Here is a short discussion of this topic.

Definition 2.2.1 *Let* $\mathbf{x} = (x_1, \cdots, x_n)$ *be the coordinates of a point in* $\mathbb{R}^n$. *Imagine an arrow (line segment with a point) with its tail at* $\mathbf{0} = (0, \cdots, 0)$ *and its point at* $\mathbf{x}$ *as shown in the following picture in the case of* $\mathbb{R}^3$.

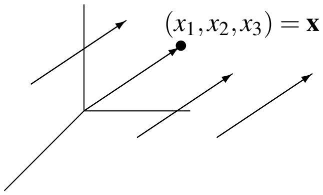


Then this arrow is called the ***position vector*** *of the point* $\mathbf{x}$. *Given two points,* P,Q *whose coordinates are* $(p_1, \cdots, p_n)$ *and* $(q_1, \cdots, q_n)$ *respectively, one can also determine the position vector from P to Q defined as follows.*

$$\overrightarrow{PQ} \equiv (q_1 - p_1, \cdots, q_n - p_n)$$

Thus every point in $\mathbb{R}^n$ determines a vector and conversely, every such position vector (arrow) which has its tail at $\mathbf{0}$ determines a point of $\mathbb{R}^n$, namely the point of $\mathbb{R}^n$ which coincides with the point of the position vector. Also two different points determine a position vector going from one to the other as just explained.

Imagine taking the above position vector and moving it around, always keeping it pointing in the same direction as shown in the following picture. After moving it around, it is regarded as the same vector because it points in the same direction and has the same length.[2]Thus each of the arrows in the above picture is regarded as the same vector. The **components** of this vector are the numbers, $x_1, \cdots, x_n$ obtained by placing the initial point of an arrow representing the vector at the origin. You should think of these numbers as directions for obtaining such a vector illustrated above. Starting at some point $(a_1, a_2, \cdots, a_n)$ in $\mathbb{R}^n$, you move to the point $(a_1 + x_1, \cdots, a_n)$ and from there to the point $(a_1 + x_1, a_2 + x_2, a_3 \cdots, a_n)$ and then to $(a_1 + x_1, a_2 + x_2, a_3 + x_3, \cdots, a_n)$ and continue this way until you obtain the point



$$(a_1 + x_1, a_2 + x_2, \cdots, a_n + x_n).$$

The arrow having its tail at $(a_1, a_2, \cdots, a_n)$ and its point at

$$(a_1 + x_1, a_2 + x_2, \cdots, a_n + x_n)$$

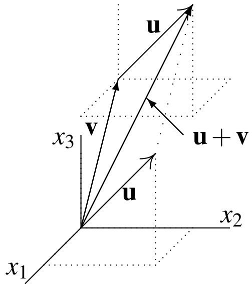
looks just like (same length and direction) the arrow which has its tail at $\mathbf{0}$ and its point at $(x_1, \cdots, x_n)$ so it is regarded as representing the same vector.

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

2.3 Geometric Meaning Of Vector Addition

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

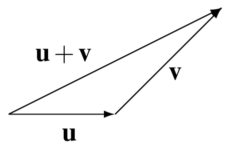
Consider the case where $n = 3$ for now. If you draw an arrow from the point in three dimensional space determined by $(0,0,0)$ to the point (a,b,c) with its tail sitting at the point $(0,0,0)$ and its point at the point (a,b,c), it is obtained by starting at $(0,0,0)$, moving parallel to the x_1 axis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



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

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

Procedure 2.3.1 *Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be two vectors. Slide* $\mathbf{v}$ *so that the tail of* $\mathbf{v}$ *is on the point of* $\mathbf{u}$. *Then draw the arrow which goes from the tail of* $\mathbf{u}$ *to the point of the slid vector* $\mathbf{v}$. *This arrow represents the vector* $\mathbf{u}+\mathbf{v}$.



Note that $P+\overrightarrow{PQ}=Q$.

2.4 Distance Between Points In $\mathbb{R}^n$ Length Of A Vector

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

Definition 2.4.1 *Let $\mathbf{x}=(x_1,\cdots,x_n)$ and $\mathbf{y}=(y_1,\cdots,y_n)$ be two points in $\mathbb{R}^n$. Then $|\mathbf{x}-\mathbf{y}|$ to indicates the distance between these points and is defined as*

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

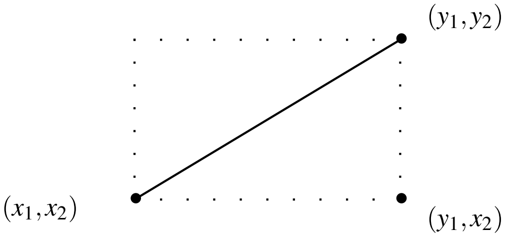
*This is called the **distance formula**. Thus $|\mathbf{x}|\equiv|\mathbf{x}-\mathbf{0}|$. The symbol, $B(\mathbf{a},r)$ is defined by*

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

*This is called an **open ball** of radius r centered at $\mathbf{a}$. It means all points in $\mathbb{R}^n$ which are closer to $\mathbf{a}$ than r. The length of a vector $\mathbf{x}$ is the distance between $\mathbf{x}$ and $\mathbf{0}$.*

First of all, note this is a generalization of the notion of distance in $\mathbb{R}$. There the distance between two points, x and y was given by the absolute value of their difference. Thus $|x-y|$ is equal to the distance between these two points on $\mathbb{R}$. Now $|x-y|=\left((x-y)^2\right)^{1/2}$ where the square root is always the positive square root. Thus it is the same formula as the above definition except there is only one term in the sum. Geometrically, this is the right way to define distance which is seen from the Pythagorean theorem. This is known as the Euclidean norm. Often people use two lines to denote this distance $||\mathbf{x}-\mathbf{y}||$. However, I want to emphasize that this is really just like the absolute value, so when the norm is defined in this way, I will usually write $|\cdot|$.

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

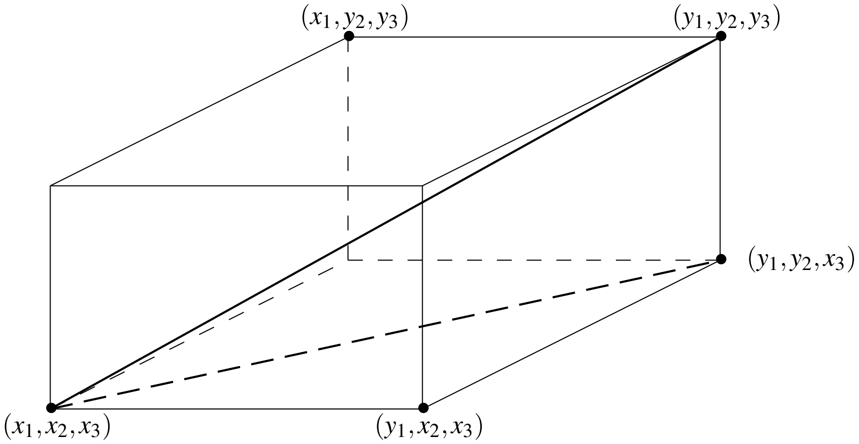


There are two points in the plane whose Cartesian coordinates are (x_1,x_2) and (y_1,y_2) respectively. Then the solid line joining these two points is the hypotenuse of a right triangle which is half of the rectangle shown in dotted lines. What is its length? Note the lengths of the sides of this triangle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

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

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

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

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

which is again just the distance formula above.

This completes the argument that the above definition is reasonable. Of course you cannot continue drawing pictures in ever higher dimensions but there is no problem with the formula for distance in any number of dimensions. Here is an example.

Example 2.4.2 *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 very much in this book but sometimes this sort of thing is done.

Another convention which is usually followed, especially in $\mathbb{R}^2$ and $\mathbb{R}^3$ is to denote the first component of a point in $\mathbb{R}^2$ by x and the second component by y. In $\mathbb{R}^3$ it is customary to denote the first and second components as just described while the third component is called z.

Example 2.4.3 *Describe the points which are at the same distance between* $(1,2,3)$ *and* $(0,1,2)$.

Let (x,y,z) be such a point. Then

$$\sqrt{(x-1)^2+(y-2)^2+(z-3)^2} = \sqrt{x^2+(y-1)^2+(z-2)^2}.$$

Squaring both sides

$$(x-1)^2+(y-2)^2+(z-3)^2 = x^2+(y-1)^2+(z-2)^2$$

and so

$$x^2-2x+14+y^2-4y+z^2-6z = x^2+y^2-2y+5+z^2-4z$$

which implies

$$-2x+14-4y-6z = -2y+5-4z$$

and so

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

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

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

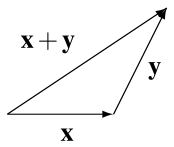
$$|\mathbf{x}-\mathbf{y}| = |\mathbf{y}-\mathbf{x}|,$$

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

The third fundamental property of distance is known as the triangle inequality. Recall that in any triangle the sum of the lengths of two sides is always at least as large as the third side. I will show you a proof of this later. This is usually stated as

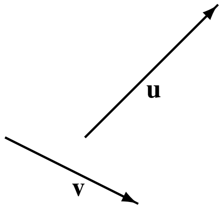
$$|\mathbf{x}+\mathbf{y}| \leq |\mathbf{x}|+|\mathbf{y}|.$$

Here is a picture which illustrates the statement of this inequality in terms of geometry. Later, this is proved, but for now, the geometric motivation will suffice. When you have a vector $\mathbf{u}$,



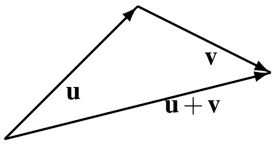
its additive inverse $-\mathbf{u}$ will be the vector which has the same magnitude as $\mathbf{u}$ but the opposite direction. When one writes $\mathbf{u}-\mathbf{v}$, the meaning is $\mathbf{u}+(-\mathbf{v})$ as with real numbers. The following example is art which illustrates these definitions and conventions.

Example 2.4.4 *Here is a picture of two vectors,* $\mathbf{u}$ *and* $\mathbf{v}$.

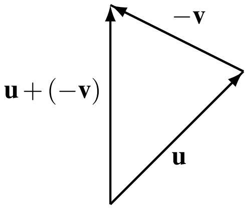


Sketch a picture of $\mathbf{u}+\mathbf{v},\mathbf{u}-\mathbf{v}$.

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



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



2.5 Geometric Meaning Of Scalar Multiplication

As discussed earlier, $\mathbf{x}=(x_1,x_2,x_3)$ determines a vector. You draw the line from $\mathbf{0}$ to $\mathbf{x}$ placing the point of the vector on $\mathbf{x}$. What is the length of this vector? The length of this vector is defined to equal $|\mathbf{x}|$ as in Definition 2.4.1. Thus the length of $\mathbf{x}$ equals $\sqrt{x_1^2+x_2^2+x_3^2}$. When you multiply $\mathbf{x}$ by a scalar α, you get $(\alpha x_1, \alpha x_2, \alpha x_3)$ and the length of this vector is defined as

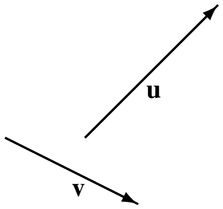
$$\sqrt{\left((\alpha x_1)^2+(\alpha x_2)^2+(\alpha x_3)^2\right)} = |\alpha| \sqrt{x_1^2+x_2^2+x_3^2}.$$

Thus the following holds.

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

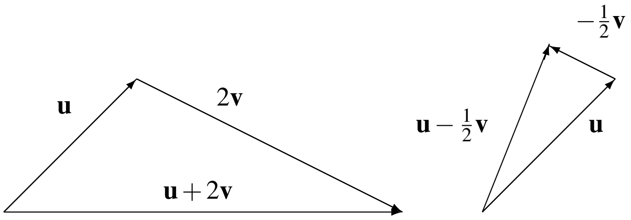
In other words, multiplication by a scalar magnifies or shrinks the length of the vector. What about the direction? You should convince yourself by drawing a picture that if α is negative, it causes the resulting vector to point in the opposite direction while if $\alpha > 0$ it preserves the direction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

Exercise 2.5.1 *Here is a picture of two vectors,* $\mathbf{u}$ *and* $\mathbf{v}$.



Sketch a picture of $\mathbf{u}+2\mathbf{v}, \mathbf{u}-\frac{1}{2}\mathbf{v}$.

The two vectors are shown below.



2.6 Parametric Lines

To begin with, suppose you have a typical equation for a line in the plane. For example,

$$y = 2x+1$$

A typical point on this line is of the form $(x, 2x+1)$ where $x \in \mathbb{R}$. You could just as well write it as $(t, 2t+1), t \in \mathbb{R}$. That is, as t changes, the ordered pair traces out the points of the line. In terms of ordered pairs, this line can be written as

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

It is the same in $\mathbb{R}^n$. A parametric line is of the form $\mathbf{x} = \mathbf{a} + t\mathbf{v}$, $t \in \mathbb{R}$. You can see this deserves to be called a line because if you find the vector determined by two points $\mathbf{a}+t_1\mathbf{v}$ and $\mathbf{a}+t_2\mathbf{v}$, this vector is

$$\mathbf{a}+t_2\mathbf{v} - (\mathbf{a}+t_1\mathbf{v}) = (t_2-t_1)\,\mathbf{v}$$

which is parallel to the vector **v**. Thus the vector between any two points on this line is always parallel to **v** which is called the direction vector.

There are two things you need for a line. A point and a direction vector. Here is an example.

Example 2.6.1 *Find a parametric equation for the line between the points* $(1,2,3)$ *and* $(2,-3,1)$.

A direction vector is $(1,-5,-2)$ because this is the vector from the first to the second of these. Then an equation of the line is

$$(x,y,z)=(1,2,3)+t(1,-5,-2), t\in\mathbb{R}$$

The example shows how to do this in general. If you have two points in $\mathbb{R}^n$, $\mathbf{a},\mathbf{b}$, then a parametric equation for the line containing these points is of the form

$$\mathbf{x}=\mathbf{a}+t(\mathbf{b}-\mathbf{a}).$$

Note that when $t=0$ you get the point **a** and when $t=1$, you get the point **b**.

Example 2.6.2 *Find a parametric equation for the line which contains the point* $(1,2,0)$ *and has direction vector* $(1,2,1)$.

From the above this is just

$$(x,y,z)=(1,2,0)+t(1,2,1),\ t\in\mathbb{R}. \tag{2.12}$$

2.7 Exercises

1. Verify all the properties 2.3-2.10.
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) $(-2,3,-2)$

2.8 Vectors And Physics

Suppose you push on something. What is important? There are really two things which are important, how hard you push and the direction you push. This illustrates the concept of force.

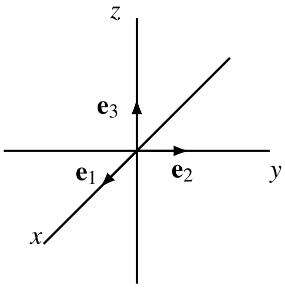
Definition 2.8.1 ***Force*** *is a vector. The magnitude of this vector is a measure of how hard it is pushing. It is measured in units such as Newtons or pounds or tons. Its direction is the direction in which the push is taking place.*

Vectors are used to model force and other physical vectors like velocity. What was just described would be called a force vector. It has two essential ingredients, its magnitude and its direction.

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 $\mathbf{v} = (a_1,\cdots,a_n)$, it follows that

$$\mathbf{v} = a_1\mathbf{e}_1 + \cdots + a_n\mathbf{e}_n = \sum_{k=1}^{n} a_i\mathbf{e}_i.$$

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.

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

Thus the addition of vectors according to the rules of addition in $\mathbb{R}^n$ which were presented earlier, yields the appropriate vector which duplicates the cumulative effect of all the vectors in the sum.

An item of notation should be mentioned here. In the case of $\mathbb{R}^n$ where $n \leq 3$, it is standard notation to use $\mathbf{i}$ for $\mathbf{e}_1$,$\mathbf{j}$ for $\mathbf{e}_2$, and $\mathbf{k}$ for $\mathbf{e}_3$. Now here are some applications of vector addition to some problems.

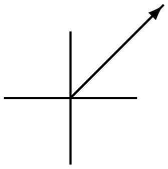
Example 2.8.2 *There are three ropes attached to a car and three people pull on these ropes. The first exerts a force of* $2\mathbf{i}+3\mathbf{j}-2\mathbf{k}$ *Newtons, the second exerts a force of* $3\mathbf{i}+5\mathbf{j}+\mathbf{k}$ *Newtons and the third exerts a force of* $5\mathbf{i}-\mathbf{j}+2\mathbf{k}$. *Newtons. Find the total force in the direction of* $\mathbf{i}$.

To find the total force add the vectors as described above. This gives $10\mathbf{i}+7\mathbf{j}+\mathbf{k}$ Newtons. Therefore, the force in the $\mathbf{i}$ direction is 10 Newtons.

As mentioned earlier, the Newton is a unit of force like pounds.

Example 2.8.3 *An airplane flies North East at 100 miles per hour. Write this as a vector.*

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



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

Definition 2.8.4 *The **speed** of an object is a measure of how fast it is going. It is measured in units of length per unit time. For example, miles per hour, kilometers per minute, feet per second. The **velocity** is a vector having the speed as the magnitude but also specifying the direction.*

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

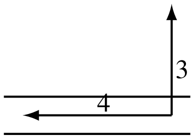
Example 2.8.5 *The velocity of an airplane is* $100\mathbf{i}+\mathbf{j}+\mathbf{k}$ *measured in kilometers per hour and at a certain instant of time its position is* $(1,2,1)$. *Here imagine a Cartesian coordinate system in which the third component is altitude and the first and second components are measured on a line from West to East and a line from South to North. Find the position of this airplane one minute later.*

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

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

Example 2.8.6 *A certain river is one half mile wide with a current flowing at 4 miles per hour from East to West. A man swims directly toward the opposite shore from the South bank of the river at a speed of 3 miles per hour. How far down the river does he find himself when he has swam across? How far does he end up swimming?*

Consider the following picture. You should write these vectors in terms of components. The velocity of the swimmer in still water would be $3\mathbf{j}$ while the velocity of the river would be $-4\mathbf{i}$. Therefore, the velocity of the swimmer is $-4\mathbf{i}+3\mathbf{j}$. Since the component of velocity in the direction across the river is 3, it follows the trip takes $1/6$ hour or 10 minutes. The speed at which he travels



is $\sqrt{4^2+3^2} = 5$ miles per hour and so he travels $5 \times \frac{1}{6} = \frac{5}{6}$ miles. Now to find the distance downstream he finds himself, note that if x is this distance, x and $1/2$ are two legs of a right triangle whose hypotenuse equals $5/6$ miles. Therefore, by the Pythagorean theorem the distance downstream is

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

2.9 Exercises

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

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

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

4. An airplane is flying due north at 150 miles per hour. A wind is pushing the airplane due east at 40 miles per hour. After 1 hour, the plane starts flying 30° East of North. Assuming the plane starts at $(0,0)$, where is it after 2 hours? Let North be the direction of the positive y axis and let East be the direction of the positive x axis.

5. City A is located at the origin while city B is located at $(300,500)$ where distances are in miles. An airplane flies at 250 miles per hour in still air. This airplane wants to fly from city A to city B but the wind is blowing in the direction of the positive y axis at a speed of 50 miles per hour. Find a unit vector such that if the plane heads in this direction, it will end up at city B having flown the shortest possible distance. How long will it take to get there?

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

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

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

9. The total force acting on an object is to be $2\mathbf{i}+\mathbf{j}+\mathbf{k}$ Newtons. A force of $-\mathbf{i}+\mathbf{j}+\mathbf{k}$ Newtons is being applied. What other force should be applied to achieve the desired total force?

10. A bird flies from its nest 5 km. in the direction 60° north of east where it stops to rest on a tree. It then flies 10 km. in the direction due southeast and lands atop a telephone pole. Place an *xy* coordinate system so that the origin is the bird's nest, and the positive *x* axis points east and the positive *y* axis points north. Find the displacement vector from the nest to the telephone pole.

11. A car is stuck in the mud. There is a cable stretched tightly from this car to a tree which is 20 feet long. A person grasps the cable in the middle and pulls with a force of 100 pounds perpendicular to the stretched cable. The center of the cable moves two feet and remains still. What is the tension in the cable? The tension in the cable is the force exerted on this point by the part of the cable nearer the car as well as the force exerted on this point by the part of the cable nearer the tree.

Chapter 3

Vector Products

3.1 The Dot Product

There are two ways of multiplying vectors which are of great importance in applications. The first of these is called the **dot product**, also called the **scalar product** and sometimes the **inner product**.

Definition 3.1.1 *Let* $\mathbf{a},\mathbf{b}$ *be two vectors in* $\mathbb{R}^n$ *define* $\mathbf{a}\cdot\mathbf{b}$ *as*

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

The dot product $\mathbf{a}\cdot\mathbf{b}$ *is sometimes denoted as* $(\mathbf{a},\mathbf{b})$ *of* $\langle\mathbf{a},\mathbf{b}\rangle$ *where a comma replaces* $\cdot$.

With this definition, there are several important properties satisfied by the dot product. In the statement of these properties, α and β will denote scalars and $\mathbf{a},\mathbf{b},\mathbf{c}$ will denote vectors.

Proposition 3.1.2 *The dot product satisfies the following properties.*

$$\mathbf{a}\cdot\mathbf{b}=\mathbf{b}\cdot\mathbf{a} \tag{3.1}$$

$$\mathbf{a}\cdot\mathbf{a}\geq 0 \text{ and equals zero if and only if } \mathbf{a}=\mathbf{0} \tag{3.2}$$

$$(\alpha\mathbf{a}+\beta\mathbf{b})\cdot\mathbf{c}=\alpha(\mathbf{a}\cdot\mathbf{c})+\beta(\mathbf{b}\cdot\mathbf{c}) \tag{3.3}$$

$$\mathbf{c}\cdot(\alpha\mathbf{a}+\beta\mathbf{b})=\alpha(\mathbf{c}\cdot\mathbf{a})+\beta(\mathbf{c}\cdot\mathbf{b}) \tag{3.4}$$

$$|\mathbf{a}|^2=\mathbf{a}\cdot\mathbf{a} \tag{3.5}$$

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

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

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

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

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

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

Theorem 3.1.5 *The dot product satisfies the inequality*

$$|\mathbf{a}\cdot\mathbf{b}| \leq |\mathbf{a}|\,|\mathbf{b}|\,. \tag{3.6}$$

Furthermore equality is obtained if and only if one of **a** *or* **b** *is a scalar multiple of the other.*

Proof: First note that if $\mathbf{b} = \mathbf{0}$ both sides of 3.6 equal zero and so the inequality holds in this case. Therefore, it will be assumed in what follows that $\mathbf{b} \neq \mathbf{0}$.

Define a function of $t \in \mathbb{R}$

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

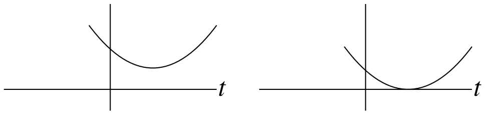
Then by 3.2, $f(t) \geq 0$ for all $t \in \mathbb{R}$. Also from 3.3,3.4,3.1, and 3.5

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

Now this means the graph, $y = f(t)$ is a polynomial which opens up and either its vertex touches the t axis or else the entire graph is above the t axis. In the first case, there exists some t where $f(t) = 0$ and this requires $\mathbf{a} + t\mathbf{b} = \mathbf{0}$ so one vector is a multiple of the other. Then clearly equality holds in 3.6. In the case where **b** is not a multiple of **a**, it follows $f(t) > 0$ for all t which says $f(t)$ has no real zeros and so from the quadratic formula,

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

which is equivalent to $|(\mathbf{a}\cdot\mathbf{b})| < |\mathbf{a}|\,|\mathbf{b}|$. ■



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

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

Theorem 3.1.6 *(Triangle inequality) For* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$

$$|\mathbf{a} + \mathbf{b}| \leq |\mathbf{a}| + |\mathbf{b}| \tag{3.7}$$

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

$$||\mathbf{a}| - |\mathbf{b}|| \leq |\mathbf{a} - \mathbf{b}| \tag{3.8}$$

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

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

Taking square roots of both sides you obtain 3.7.

It remains to consider when equality occurs. If either vector equals zero, then that vector equals zero times the other vector and the claim about when equality occurs is verified. Therefore, it can be assumed both vectors are nonzero. To get equality in the second inequality above, Theorem 3.1.5 implies one of the vectors must be a multiple of the other. Say $\mathbf{b}=\alpha\mathbf{a}$. If $\alpha<0$ then equality cannot occur in the first inequality because in this case

$$(\mathbf{a}\cdot\mathbf{b}) = \alpha|\mathbf{a}|^2 < 0 < |\alpha||\mathbf{a}|^2 = |\mathbf{a}\cdot\mathbf{b}|$$

Therefore, $\alpha \geq 0$.

To get the other form of the triangle inequality,

$$\mathbf{a} = \mathbf{a}-\mathbf{b}+\mathbf{b}$$

so

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

Therefore,

$$|\mathbf{a}|-|\mathbf{b}| \leq |\mathbf{a}-\mathbf{b}| \tag{3.9}$$

Similarly,

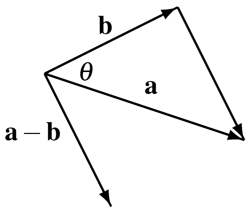
$$|\mathbf{b}|-|\mathbf{a}| \leq |\mathbf{b}-\mathbf{a}| = |\mathbf{a}-\mathbf{b}|. \tag{3.10}$$

It follows from 3.9 and 3.10 that 3.8 holds. This is because $||\mathbf{a}|-|\mathbf{b}||$ equals the left side of either 3.9 or 3.10 and either way, $||\mathbf{a}|-|\mathbf{b}|| \leq |\mathbf{a}-\mathbf{b}|$. ■

3.2 The Geometric Significance Of The Dot Product

3.2.1 The Angle Between Two Vectors

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



By the law of cosines,

$$|\mathbf{a}-\mathbf{b}|^2 + |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2|\mathbf{a}||\mathbf{b}|\cos(\theta).$$

Also from the properties of the dot product,

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

and so comparing the above two formulas,

$$\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}|\,|\mathbf{b}|\cos\theta. \tag{3.11}$$

In words, the dot product of two vectors equals the product of the magnitude of the two vectors multiplied by the cosine of the included angle. Note this gives a geometric description of the dot product which does not depend explicitly on the coordinates of the vectors.

Example 3.2.1 *Find the angle between the vectors* $2\mathbf{i}+\mathbf{j}-\mathbf{k}$ *and* $3\mathbf{i}+4\mathbf{j}+\mathbf{k}$.

The dot product of the two vectors equals $6+4-1=9$. The norms of the two vectors are $\sqrt{4+1+1}=\sqrt{6}$ and $\sqrt{9+16+1}=\sqrt{26}$. Therefore, from 3.11 the cosine of the included angle equals

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

Now the cosine is known, the angle can be determines by solving the equation, $\cos\theta = .72058$. This will involve using a calculator or a table of trigonometric functions. The answer is $\theta = .76616$ radians or in terms of degrees, $\theta = .76616\times\frac{360}{2\pi} = 43.898°$. Recall how this last computation is done. Set up a proportion, $\frac{x}{.76616} = \frac{360}{2\pi}$ because $360°$ corresponds to 2π radians. However, in calculus, you should get used to thinking in terms of radians and not degrees. This is because all the important calculus formulas are defined in terms of radians.

Example 3.2.2 *Let* $\mathbf{u},\mathbf{v}$ *be two vectors whose magnitudes are equal to 3 and 4 respectively and such that if they are placed in standard position with their tails at the origin, the angle between* $\mathbf{u}$ *and the positive x axis equals 30°* *and the angle between* $\mathbf{v}$ *and the positive x axis is* $-30°$. *Find* $\mathbf{u}\cdot\mathbf{v}$.

From the geometric description of the dot product in 3.11

$$\mathbf{u}\cdot\mathbf{v} = 3\times4\times\cos(60°) = 3\times4\times1/2 = 6.$$

Observation 3.2.3 *Two vectors are said to be* ***perpendicular*** *if the included angle is* $\pi/2$ *radians (90°). You can tell if two nonzero vectors are perpendicular by simply taking their dot product. If the answer is zero, this means they are perpendicular because* $\cos\theta = 0$.

Example 3.2.4 *Determine whether the two vectors,* $2\mathbf{i}+\mathbf{j}-\mathbf{k}$ *and* $1\mathbf{i}+3\mathbf{j}+5\mathbf{k}$ *are perpendicular.*

When you take this dot product you get $2+3-5=0$ and so these two are indeed perpendicular.

Definition 3.2.5 *When two lines intersect, the angle between the two lines is the smaller of the two angles determined.*

Example 3.2.6 *Find the angle between the two lines having the parametrizations* $(1,2,0)+t(1,2,3)$ *and* $(0,4,-3)+t(-1,2,-3)$.

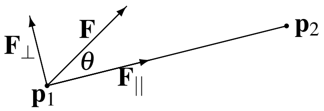
These two lines intersect, when $t=0$ in the first and $t=-1$ in the second. It is only a matter of finding the angle between the direction vectors. One angle determined is given by

$$\cos\theta = \frac{-6}{14} = \frac{-3}{7}. \tag{3.12}$$

We don't want this angle because it is obtuse. The angle desired is the acute angle given by $\cos\theta = \frac{3}{7}$. It is obtained by replacing one of the direction vectors with -1 times it.

3.2.2 Work And Projections

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



In this picture the force, $\mathbf{F}$ is applied to an object which moves on the straight line from $\mathbf{p}_1$ to $\mathbf{p}_2$. There are two vectors shown, $\mathbf{F}_{||}$ and $\mathbf{F}_\perp$ and the picture is intended to indicate that when you add these two vectors you get $\mathbf{F}$ while $\mathbf{F}_{||}$ acts in the direction of motion and $\mathbf{F}_\perp$ acts perpendicular to the direction of motion. Only $\mathbf{F}_{||}$ contributes to the work done by $\mathbf{F}$ on the object as it moves from $\mathbf{p}_1$ to $\mathbf{p}_2$. $\mathbf{F}_{||}$ is called the **component of the force** in the direction of motion. From trigonometry, you see the magnitude of $\mathbf{F}_{||}$ should equal $|\mathbf{F}|\,|\cos\theta|$. Thus, since $\mathbf{F}_{||}$ points in the direction of the vector from $\mathbf{p}_1$ to $\mathbf{p}_2$, the total work done should equal

$$|\mathbf{F}|\left|\overrightarrow{\mathbf{p}_1\mathbf{p}_2}\right|\cos\theta = |\mathbf{F}|\,|\mathbf{p}_2-\mathbf{p}_1|\cos\theta$$

If the included angle had been obtuse, then the work done by the force, $\mathbf{F}$ on the object would have been negative because in this case, the force tends to impede the motion from $\mathbf{p}_1$ to $\mathbf{p}_2$ but in this case, $\cos\theta$ would also be negative and so it is still the case that the work done would be given by the above formula. Thus from the geometric description of the dot product given above, the work equals

$$|\mathbf{F}|\,|\mathbf{p}_2-\mathbf{p}_1|\cos\theta = \mathbf{F}\cdot(\mathbf{p}_2-\mathbf{p}_1)\,.$$

This explains the following definition.

Definition 3.2.7 *Let* $\mathbf{F}$ *be a force acting on an object which moves from the point* $\mathbf{p}_1$ *to the point* $\mathbf{p}_2$*. Then the* ***work*** *done on the object by the given force equals* $\mathbf{F}\cdot(\mathbf{p}_2-\mathbf{p}_1)$*.*

The concept of writing a given vector $\mathbf{F}$ in terms of two vectors, one which is parallel to a given vector $\mathbf{D}$ and the other which is perpendicular can also be explained with no reliance on trigonometry, completely in terms of the algebraic properties of the dot product. As before, this is mathematically more significant than any approach involving geometry or trigonometry because it extends to more interesting situations. This is done next.

Theorem 3.2.8 *Let* $\mathbf{F}$ *and* $\mathbf{D}$ *be nonzero vectors. Then there exist unique vectors* $\mathbf{F}_{||}$ *and* $\mathbf{F}_{\perp}$ *such that*

$$\mathbf{F} = \mathbf{F}_{||} + \mathbf{F}_{\perp} \tag{3.13}$$

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

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

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

Proof: Suppose 3.13 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

$$\begin{aligned}\mathbf{F}_{\perp}\cdot\mathbf{D} &= \mathbf{F}\cdot\mathbf{D} - \frac{\mathbf{F}\cdot\mathbf{D}}{|\mathbf{D}|^2}\mathbf{D}\cdot\mathbf{D}\\ &= \mathbf{F}\cdot\mathbf{D} - \mathbf{F}\cdot\mathbf{D} = 0.\ \blacksquare\end{aligned}$$

Example 3.2.9 *Let* $\mathbf{F} = 2\mathbf{i}+7\mathbf{j}-3\mathbf{k}$ *Newtons. Find the work done by this force in moving from the point* $(1,2,3)$ *to the point* $(-9,-3,4)$ *along the straight line segment joining these points where distances are measured in meters.*

According to the definition, this work is

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

Note that if the force had been given in pounds and the distance had been given in feet, the units on the work would have been foot pounds. In general, work has units equal to units of a force times units of a length. Instead of writing Newton meter, people write joule because a joule is by definition a Newton meter. That word is pronounced "jewel" and it is the unit of work in the metric system of units. Also be sure you observe that the work done by the force can be negative as in the above example. In fact, work can be either positive, negative, or zero. You just have to do the computations to find out.

Example 3.2.10 *Find* $\text{proj}_{\mathbf{u}}(\mathbf{v})$ *if* $\mathbf{u}=2\mathbf{i}+3\mathbf{j}-4\mathbf{k}$ *and* $\mathbf{v}=\mathbf{i}-2\mathbf{j}+\mathbf{k}$.

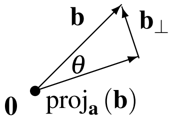
From the above discussion in Theorem 3.2.8, this is just

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

Example 3.2.11 *Suppose* $\mathbf{a}$, *and* $\mathbf{b}$ *are vectors and* $\mathbf{b}_\perp = \mathbf{b}-\text{proj}_{\mathbf{a}}(\mathbf{b})$. *What is the magnitude of* $\mathbf{b}_\perp$ *in terms of the included angle*?

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

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



3.2.3 The Inner Product And Distance In $\mathbb{C}^n$

It is necessary to give a generalization of the dot product for vectors in $\mathbb{C}^n$. This is often called the inner product. It reduces to the definition of the dot product in the case the components of the vector are real.

Definition 3.2.12 *Let* $\mathbf{x},\mathbf{y}\in\mathbb{C}^n$. *Thus* $\mathbf{x}=(x_1,\cdots,x_n)$ *where each* $x_k\in\mathbb{C}$ *and a similar formula holding for* $\mathbf{y}$. *Then the inner product of these two vectors is defined to be*

$$\mathbf{x}\cdot\mathbf{y}\equiv\sum_j x_j\overline{y_j}\equiv x_1\overline{y_1}+\cdots+x_n\overline{y_n}.$$

The inner product is often denoted as $(\mathbf{x},\mathbf{y})$ *or* $\langle\mathbf{x},\mathbf{y}\rangle$.

Notice how you put the conjugate on the entries of the vector **y**. It makes no difference if the vectors happen to be real vectors but with complex vectors you must do it this way. The reason for this is that when you take the inner product of a vector with itself, you want to get the square of the length of the vector, a positive number. Placing the conjugate on the components of **y** in the above definition assures this will take place. Thus

$$\mathbf{x}\cdot\mathbf{x}=\sum_j x_j\overline{x_j}=\sum_j \left|x_j\right|^2\geq 0.$$

If you didn't place a conjugate as in the above definition, things wouldn't work out correctly. For example,

$$(1+i)^2+2^2=4+2i$$

and this is not a positive number.

The following properties of the inner product follow immediately from the definition and you should verify each of them.

Properties of the inner product:

1. $\mathbf{u}\cdot\mathbf{v}=\overline{\mathbf{v}\cdot\mathbf{u}}$.
2. If a,b are numbers and $\mathbf{u},\mathbf{v},\mathbf{z}$ are vectors then $(a\mathbf{u}+b\mathbf{v})\cdot\mathbf{z}=a(\mathbf{u}\cdot\mathbf{z})+b(\mathbf{v}\cdot\mathbf{z})$.
3. $\mathbf{u}\cdot\mathbf{u}\geq 0$ and it equals 0 if and only if $\mathbf{u}=\mathbf{0}$.

Note this implies $(\mathbf{x}\cdot\alpha\mathbf{y})=\overline{\alpha}(\mathbf{x}\cdot\mathbf{y})$ because

$$(\mathbf{x}\cdot\alpha\mathbf{y})=\overline{(\alpha\mathbf{y}\cdot\mathbf{x})}=\overline{\alpha(\mathbf{y}\cdot\mathbf{x})}=\overline{\alpha}(\mathbf{x}\cdot\mathbf{y})$$

The norm is defined in the usual way.

Definition 3.2.13 *For* $\mathbf{x}\in\mathbb{C}^n$,

$$|\mathbf{x}|\equiv\left(\sum_{k=1}^n |x_k|^2\right)^{1/2}=(\mathbf{x}\cdot\mathbf{x})^{1/2}$$

Here is a fundamental inequality called the **Cauchy Schwarz inequality** which is stated here in $\mathbb{C}^n$. First here is a simple lemma.

Lemma 3.2.14 *If* $z\in\mathbb{C}$ *there exists* $\theta\in\mathbb{C}$ *such that* $\theta z=|z|$ *and* $|\theta|=1$.

Proof: Let $\theta=1$ if $z=0$ and otherwise, let $\theta=\dfrac{\overline{z}}{|z|}$. Recall that for $z=x+iy,\overline{z}=x-iy$ and $\overline{z}z=|z|^2$.

I will give a proof of this important inequality which depends only on the above list of properties of the inner product. It will be slightly different than the earlier proof.

Theorem 3.2.15 *(Cauchy Schwarz)The following inequality holds for* **x** *and* $\mathbf{y}\in\mathbb{C}^n$.

$$|(\mathbf{x}\cdot\mathbf{y})|\leq(\mathbf{x}\cdot\mathbf{x})^{1/2}(\mathbf{y}\cdot\mathbf{y})^{1/2}\tag{3.14}$$

Equality holds in this inequality if and only if one vector is a multiple of the other.

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

$$\theta\,(\mathbf{x}\cdot\mathbf{y}) = |(\mathbf{x}\cdot\mathbf{y})|$$

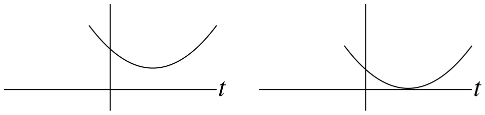
Consider $p(t) \equiv \left(\mathbf{x}+\overline{\theta}t\mathbf{y}, \mathbf{x}+t\overline{\theta}\mathbf{y}\right)$ where $t \in \mathbb{R}$. Then from the above list of properties of the dot product,

$$\begin{aligned}
0 &\leq p(t) = (\mathbf{x}\cdot\mathbf{x}) + t\theta\,(\mathbf{x}\cdot\mathbf{y}) + t\overline{\theta}\,(\mathbf{y}\cdot\mathbf{x}) + t^2\,(\mathbf{y}\cdot\mathbf{y}) \\
&= (\mathbf{x}\cdot\mathbf{x}) + t\theta\,(\mathbf{x}\cdot\mathbf{y}) + t\overline{\theta(\mathbf{x}\cdot\mathbf{y})} + t^2\,(\mathbf{y}\cdot\mathbf{y}) \\
&= (\mathbf{x}\cdot\mathbf{x}) + 2t\,\mathrm{Re}\,(\theta\,(\mathbf{x}\cdot\mathbf{y})) + t^2\,(\mathbf{y}\cdot\mathbf{y}) \\
&= (\mathbf{x}\cdot\mathbf{x}) + 2t\,|(\mathbf{x}\cdot\mathbf{y})| + t^2\,(\mathbf{y}\cdot\mathbf{y}) \qquad (3.15)
\end{aligned}$$

and this must hold for all $t \in \mathbb{R}$. Therefore, if $(\mathbf{y}\cdot\mathbf{y}) = 0$ it must be the case that $|(\mathbf{x}\cdot\mathbf{y})| = 0$ also since otherwise the above inequality would be violated. Therefore, in this case,

$$|(\mathbf{x}\cdot\mathbf{y})| \leq (\mathbf{x}\cdot\mathbf{x})^{1/2}\,(\mathbf{y}\cdot\mathbf{y})^{1/2}.$$

On the other hand, if $(\mathbf{y}\cdot\mathbf{y}) \neq 0$, then $p(t) \geq 0$ for all t means the graph of $y = p(t)$ is a parabola which opens up and it either has exactly one real zero in the case its vertex touches the t axis or it has no real zeros.



From the quadratic formula this happens exactly when

$$4\,|(\mathbf{x}\cdot\mathbf{y})|^2 - 4\,(\mathbf{x}\cdot\mathbf{x})\,(\mathbf{y}\cdot\mathbf{y}) \leq 0$$

which is equivalent to 3.14.

It is clear from a computation that if one vector is a scalar multiple of the other that equality holds in 3.14. Conversely, suppose equality does hold. Then this is equivalent to saying $4\,|(\mathbf{x}\cdot\mathbf{y})|^2 - 4\,(\mathbf{x}\cdot\mathbf{x})\,(\mathbf{y}\cdot\mathbf{y}) = 0$ and so from the quadratic formula, there exists one real zero to $p(t) = 0$. Call it t_0. Then

$$p(t_0) \equiv \left(\left(\mathbf{x}+\overline{\theta}t_0\mathbf{y}\right)\cdot\left(\mathbf{x}+t_0\overline{\theta}\mathbf{y}\right)\right) = \left|\mathbf{x}+\overline{\theta}t\mathbf{y}\right|^2 = 0$$

and so $\mathbf{x} = -\overline{\theta}t_0\mathbf{y}$. ■

Note that I only used part of the above properties of the inner product. It was not necessary to use the one which says that if $(\mathbf{x}\cdot\mathbf{x}) = 0$ then $\mathbf{x} = \mathbf{0}$.

By analogy to the case of $\mathbb{R}^n$, length or magnitude of vectors in $\mathbb{C}^n$ can be defined.

Definition 3.2.16 *Let $\mathbf{z} \in \mathbb{C}^n$. Then $|\mathbf{z}| \equiv (\mathbf{z}\cdot\mathbf{z})^{1/2}$.*

The conclusions of the following theorem are also called the **axioms for a norm.**

Theorem 3.2.17 *For length defined in Definition 3.2.16, the following hold.*

$$|\mathbf{z}| \geq 0 \text{ and } |\mathbf{z}| = 0 \text{ if and only if } \mathbf{z} = \mathbf{0} \qquad (3.16)$$

$$\text{If } \alpha \text{ is a scalar, } |\alpha\mathbf{z}| = |\alpha|\,|\mathbf{z}| \qquad (3.17)$$

$$|\mathbf{z}+\mathbf{w}| \leq |\mathbf{z}| + |\mathbf{w}|. \qquad (3.18)$$

Proof: The first two claims are left as exercises. To establish the third, you use the same argument which was used in $\mathbb{R}^n$.

$$\begin{aligned}
|\mathbf{z}+\mathbf{w}|^2 &= (\mathbf{z}+\mathbf{w},\mathbf{z}+\mathbf{w}) \\
&= \mathbf{z}\cdot\mathbf{z}+\mathbf{w}\cdot\mathbf{w}+\mathbf{w}\cdot\mathbf{z}+\mathbf{z}\cdot\mathbf{w} \\
&= |\mathbf{z}|^2+|\mathbf{w}|^2+2\operatorname{Re}\mathbf{w}\cdot\mathbf{z} \\
&\leq |\mathbf{z}|^2+|\mathbf{w}|^2+2|\mathbf{w}\cdot\mathbf{z}| \\
&\leq |\mathbf{z}|^2+|\mathbf{w}|^2+2|\mathbf{w}||\mathbf{z}| = (|\mathbf{z}|+|\mathbf{w}|)^2. \blacksquare
\end{aligned}$$

Occasionally, I may refer to the inner product in $\mathbb{C}^n$ as the dot product. They are the same thing for $\mathbb{R}^n$. However, it is convenient to draw a distinction when discussing matrix multiplication a little later.

3.3 Exercises

1. Use formula 3.11 to verify the Cauchy Schwarz inequality and to show that equality occurs if and only if one of the vectors is a scalar multiple of the other.

2. For $\mathbf{u},\mathbf{v}$ vectors in $\mathbb{R}^3$, define the product, $\mathbf{u}*\mathbf{v}\equiv u_1v_1+2u_2v_2+3u_3v_3$. Show the axioms for a dot product all hold for this funny product. Prove

   $$|\mathbf{u}*\mathbf{v}|\leq(\mathbf{u}*\mathbf{u})^{1/2}(\mathbf{v}*\mathbf{v})^{1/2}.$$

   Hint: Do not try to do this with methods from trigonometry.

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

4. Find the angle between the vectors $\mathbf{i}-2\mathbf{j}+\mathbf{k}$ and $\mathbf{i}+2\mathbf{j}-7\mathbf{k}$.

5. Find $\operatorname{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{v}=(1,0,-2)$ and $\mathbf{u}=(1,2,3)$.

6. Find $\operatorname{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{v}=(1,2,-2)$ and $\mathbf{u}=(1,0,3)$.

7. Find $\operatorname{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{v}=(1,2,-2,1)$ and $\mathbf{u}=(1,2,3,0)$.

8. Does it make sense to speak of $\operatorname{proj}_{\mathbf{0}}(\mathbf{v})$?

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

10. Prove the Cauchy Schwarz inequality in $\mathbb{R}^n$ as follows. For $\mathbf{u},\mathbf{v}$ vectors, consider

    $$(\mathbf{u}-\operatorname{proj}_{\mathbf{v}}\mathbf{u})\cdot(\mathbf{u}-\operatorname{proj}_{\mathbf{v}}\mathbf{u})\geq 0$$

    Now simplify using the axioms of the dot product and then put in the formula for the projection. Of course this expression equals 0 and you get equality in the Cauchy Schwarz inequality if and only if $\mathbf{u}=\operatorname{proj}_{\mathbf{v}}\mathbf{u}$. What is the geometric meaning of $\mathbf{u}=\operatorname{proj}_{\mathbf{v}}\mathbf{u}$?

11. A boy drags a sled for 100 feet along the ground by pulling on a rope which is 20 degrees from the horizontal with a force of 40 pounds. How much work does this force do?

12. A girl drags a sled for 200 feet along the ground by pulling on a rope which is 30 degrees from the horizontal with a force of 20 pounds. How much work does this force do?

13. A large dog drags a sled for 300 feet along the ground by pulling on a rope which is 45 degrees from the horizontal with a force of 20 pounds. How much work does this force do?

14. How much work in Newton meters does it take to slide a crate 20 meters along a loading dock by pulling on it with a 200 Newton force at an angle of 30° from the horizontal?

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

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

17. An object moves 20 meters in the direction of $\mathbf{k}+\mathbf{j}$. There are two forces acting on this object, $\mathbf{F}_1 = \mathbf{i}+\mathbf{j}+2\mathbf{k}$, and $\mathbf{F}_2 = \mathbf{i}+2\mathbf{j}-6\mathbf{k}$. Find the total work done on the object by the two forces. **Hint:** You can take the work done by the resultant of the two forces or you can add the work done by each force.

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

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

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

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

22. Recall that the open ball having center at $\mathbf{a}$ and radius r is given by

$$B(\mathbf{a},r) \equiv \{\mathbf{x} : |\mathbf{x}-\mathbf{a}| < r\}$$

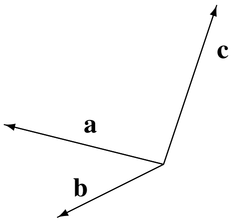
Show that if $\mathbf{y} \in B(\mathbf{a},r)$, then there exists a positive number δ such that $B(\mathbf{y},\delta) \subseteq B(\mathbf{a},r)$. (The symbol $\subseteq$ means that every point in $B(\mathbf{y},\delta)$ is also in $B(\mathbf{a},r)$. In words, it states that $B(\mathbf{y},\delta)$ is contained in $B(\mathbf{a},r)$. The statement $\mathbf{y} \in B(\mathbf{a},r)$ says that $\mathbf{y}$ is one of the points of $B(\mathbf{a},r)$.) When you have done this, you will have shown that an open ball is open. This is a fantastically important observation although its major implications will not be explored very much in this book.

3.4 The Cross Product

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

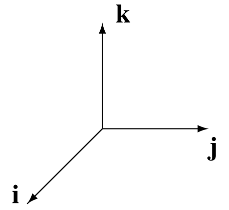
Definition 3.4.1 *Three vectors,* $\mathbf{a},\mathbf{b},\mathbf{c}$ *form a right handed system if when you extend the fingers of your right hand along the vector* $\mathbf{a}$ *and close them in the direction of* $\mathbf{b}$*, the thumb points roughly in the direction of* $\mathbf{c}$*.*

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



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

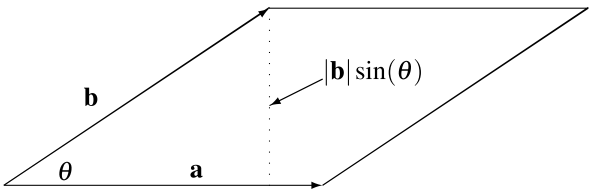
From now on, the vectors, $\mathbf{i},\mathbf{j},\mathbf{k}$ will always form a right handed system. To repeat, if you extend the fingers of your right hand along $\mathbf{i}$ and close them in the direction $\mathbf{j}$, the thumb points in the direction of $\mathbf{k}$. The following is the geometric description of the cross product. It gives both the direction and the magnitude and therefore specifies the vector.



Definition 3.4.2 *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two vectors in* $\mathbb{R}^3$*. Then* $\mathbf{a}\times\mathbf{b}$ *is defined by the following two rules.*

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

Note that $|\mathbf{a}\times\mathbf{b}|$ is the area of the parallelogram determined by $\mathbf{a}$ and $\mathbf{b}$.



The cross product satisfies the following properties.

$$\mathbf{a}\times\mathbf{b}=-\left(\mathbf{b}\times\mathbf{a}\right)\ ,\ \mathbf{a}\times\mathbf{a}=\mathbf{0}, \tag{3.19}$$

For α a scalar,

$$\left(\alpha\mathbf{a}\right)\times\mathbf{b}=\alpha\left(\mathbf{a}\times\mathbf{b}\right)=\mathbf{a}\times\left(\alpha\mathbf{b}\right), \tag{3.20}$$

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

$$\mathbf{a}\times\left(\mathbf{b}+\mathbf{c}\right)=\mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c}, \tag{3.21}$$

$$\left(\mathbf{b}+\mathbf{c}\right)\times\mathbf{a}=\mathbf{b}\times\mathbf{a}+\mathbf{c}\times\mathbf{a}. \tag{3.22}$$

Formula 3.19 follows immediately from the definition. The vectors $\mathbf{a}\times\mathbf{b}$ and $\mathbf{b}\times\mathbf{a}$ have the same magnitude, $|\mathbf{a}|\,|\mathbf{b}|\sin\theta$, and an application of the right hand rule shows they have opposite direction. Formula 3.20 is also fairly clear. If α is a nonnegative scalar, the direction of $\left(\alpha\mathbf{a}\right)\times\mathbf{b}$ is the same as the direction of $\mathbf{a}\times\mathbf{b}$,$\alpha\left(\mathbf{a}\times\mathbf{b}\right)$ and $\mathbf{a}\times\left(\alpha\mathbf{b}\right)$ while the magnitude is just α times the magnitude of $\mathbf{a}\times\mathbf{b}$ which is the same as the magnitude of $\alpha\left(\mathbf{a}\times\mathbf{b}\right)$ and $\mathbf{a}\times\left(\alpha\mathbf{b}\right)$. Using this yields equality in 3.20. In the case where $\alpha<0$, everything works the same way except the vectors are all pointing in the opposite direction and you must multiply by $|\alpha|$ when comparing their magnitudes. The distributive laws are much harder to establish but the second follows from the first quite easily. Thus, assuming the first, and using 3.19,

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

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 3.4.3 *Let* $\mathbf{a}=a_1\mathbf{i}+a_2\mathbf{j}+a_3\mathbf{k}$ *and* $\mathbf{b}=b_1\mathbf{i}+b_2\mathbf{j}+b_3\mathbf{k}$ *be two vectors. Then*

$$\begin{aligned}\mathbf{a}\times\mathbf{b}=&\left(a_2b_3-a_3b_2\right)\mathbf{i}+\left(a_3b_1-a_1b_3\right)\mathbf{j}+\\&+\left(a_1b_2-a_2b_1\right)\mathbf{k}.\end{aligned} \tag{3.23}$$

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

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

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

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

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

$$=\left(a_2b_3-a_3b_2\right)\mathbf{i}+\left(a_3b_1-a_1b_3\right)\mathbf{j}+\left(a_1b_2-a_2b_1\right)\mathbf{k} \tag{3.24}$$

■

It is probably impossible for most people to remember 3.23. Fortunately, there is a somewhat easier way to remember it. Define the determinant of a 2 × 2 matrix as follows

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

Then

$$\mathbf{a}\times\mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} \tag{3.25}$$

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

$$\mathbf{i}(-1)^{1+1}\begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} + \mathbf{j}(-1)^{2+1}\begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} + \mathbf{k}(-1)^{3+1}\begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}$$

$$= \mathbf{i}\begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} - \mathbf{j}\begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} + \mathbf{k}\begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}$$

Note that to get the scalar which multiplies **i** you take the determinant of what is left after deleting the first row and the first column and multiply by $(-1)^{1+1}$ because **i** is in the first row and the first column. Then you do the same thing for the **j** and **k**. In the case of the **j** there is a minus sign because **j** is in the first row and the second column and so$(-1)^{1+2} = -1$ while the **k** is multiplied by $(-1)^{3+1} = 1$. The above equals

$$(a_2b_3 - a_3b_2)\mathbf{i} - (a_1b_3 - a_3b_1)\mathbf{j} + (a_1b_2 - a_2b_1)\mathbf{k} \tag{3.26}$$

which is the same as 3.24. There will be much more presented on determinants later. For now, consider this an introduction if you have not seen this topic.

Example 3.4.4 *Find* $(\mathbf{i} - \mathbf{j} + 2\mathbf{k}) \times (3\mathbf{i} - 2\mathbf{j} + \mathbf{k})$.

Use 3.25 to compute this.

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

Example 3.4.5 *Find the area of the parallelogram determined by the vectors,*

$$(\mathbf{i} - \mathbf{j} + 2\mathbf{k}),\ (3\mathbf{i} - 2\mathbf{j} + \mathbf{k}).$$

These are the same two vectors in Example 3.4.4.

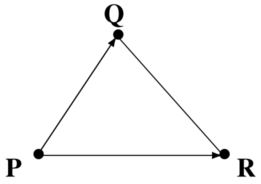
From Example 3.4.4 and the geometric description of the cross product, the area is just the norm of the vector obtained in Example 3.4.4. Thus the area is $\sqrt{9+25+1} = \sqrt{35}$.

Example 3.4.6 *A triangle determined by* $(1,2,3), (0,2,5), (5,1,2)$. *Find its area.*

This triangle is obtained by connecting the three points with lines. Picking $(1,2,3)$ as a starting point, there are two displacement vectors, $(-1,0,2)$ and $(4,-1,-1)$ such that the given vector added to these displacement vectors gives the other two vectors. The area of the triangle is half the area of the parallelogram determined by $(-1,0,2)$ and $(4,-1,-1)$. Thus $(-1,0,2)\times(4,-1,-1)=(2,7,1)$ and so the area of the triangle is $\frac{1}{2}\sqrt{4+49+1}=\frac{3}{2}\sqrt{6}$.

Observation 3.4.7 *In general, if you have three points (vectors) in $\mathbb{R}^3$, $\mathbf{P},\mathbf{Q},\mathbf{R}$ the area of the triangle is given by*

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

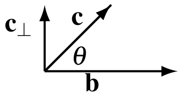


3.4.1 The Distributive Law For The Cross Product

This section gives a proof for 3.21, a fairly difficult topic. It is included here for the interested student. If you are satisfied with taking the distributive law on faith, it is not necessary to read this section. The proof given here is quite clever and follows the one given in [3]. Another approach, based on volumes of parallelepipeds is found in [16] and is discussed a little later.

Lemma 3.4.8 *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,



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

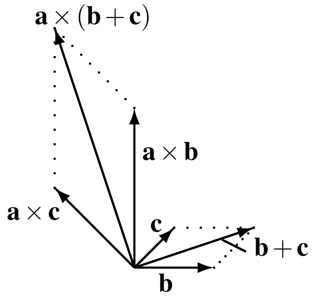
Therefore, $\mathbf{b}\times\mathbf{c}$ and $\mathbf{b}\times\mathbf{c}_{\perp}$ also have the same magnitude and so they are the same vector. ■

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

Theorem 3.4.9 *Let $\mathbf{a},\mathbf{b}$, and $\mathbf{c}$ be vectors in $\mathbb{R}^3$. Then*

$$\mathbf{a}\times(\mathbf{b}+\mathbf{c})=\mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c} \tag{3.27}$$

Proof: Suppose first that $\mathbf{a}\cdot\mathbf{b}=\mathbf{a}\cdot\mathbf{c}=0$. Now imagine $\mathbf{a}$ is a vector coming out of the page and let $\mathbf{b},\mathbf{c}$ and $\mathbf{b}+\mathbf{c}$ be as shown in the following picture. Then $\mathbf{a}\times\mathbf{b},\mathbf{a}\times(\mathbf{b}+\mathbf{c})$, are each vectors in


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

the same plane, and $\mathbf{a}\times\mathbf{c}$ perpendicular to $\mathbf{a}$ as shown. Thus

$$\mathbf{a}\times\mathbf{c}\cdot\mathbf{c}=0,\mathbf{a}\times(\mathbf{b}+\mathbf{c})\cdot(\mathbf{b}+\mathbf{c})=0,$$

and $\mathbf{a}\times\mathbf{b}\cdot\mathbf{b}=0$. This implies that to get $\mathbf{a}\times\mathbf{b}$ you move counterclockwise through an angle of $\pi/2$ radians from the vector $\mathbf{b}$. Similar relationships exist between the vectors $\mathbf{a}\times(\mathbf{b}+\mathbf{c})$ and $\mathbf{b}+\mathbf{c}$ and the vectors $\mathbf{a}\times\mathbf{c}$ and $\mathbf{c}$. Thus the angle between $\mathbf{a}\times\mathbf{b}$ and $\mathbf{a}\times(\mathbf{b}+\mathbf{c})$ is the same as the angle between $\mathbf{b}+\mathbf{c}$ and $\mathbf{b}$ and the angle between $\mathbf{a}\times\mathbf{c}$ and $\mathbf{a}\times(\mathbf{b}+\mathbf{c})$ is the same as the angle between $\mathbf{c}$ and $\mathbf{b}+\mathbf{c}$. In addition to this, since $\mathbf{a}$ is perpendicular to these vectors,

$$|\mathbf{a}\times\mathbf{b}|=|\mathbf{a}|\,|\mathbf{b}|\,,|\mathbf{a}\times(\mathbf{b}+\mathbf{c})|=|\mathbf{a}|\,|\mathbf{b}+\mathbf{c}|\,,\text{ and}$$

$$|\mathbf{a}\times\mathbf{c}|=|\mathbf{a}|\,|\mathbf{c}|\,.$$

Therefore,

$$\frac{|\mathbf{a}\times(\mathbf{b}+\mathbf{c})|}{|\mathbf{b}+\mathbf{c}|}=\frac{|\mathbf{a}\times\mathbf{c}|}{|\mathbf{c}|}=\frac{|\mathbf{a}\times\mathbf{b}|}{|\mathbf{b}|}=|\mathbf{a}|$$

and so

$$\frac{|\mathbf{a}\times(\mathbf{b}+\mathbf{c})|}{|\mathbf{a}\times\mathbf{c}|}=\frac{|\mathbf{b}+\mathbf{c}|}{|\mathbf{c}|},\frac{|\mathbf{a}\times(\mathbf{b}+\mathbf{c})|}{|\mathbf{a}\times\mathbf{b}|}=\frac{|\mathbf{b}+\mathbf{c}|}{|\mathbf{b}|}$$

showing the triangles making up the parallelogram on the right and the four sided figure on the left in the above picture are similar. It follows the four sided figure on the left is in fact a parallelogram and this implies the diagonal is the vector sum of the vectors on the sides, yielding 3.27.

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

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

The result of Problem 18 of the exercises 3.3 is used to go from the first to the second line.

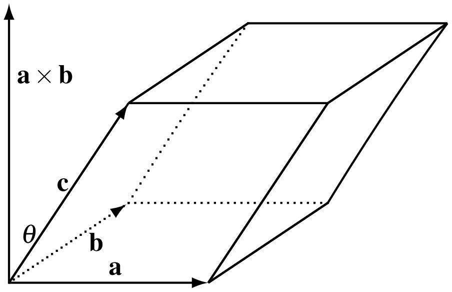
3.4.2 The Box Product

Definition 3.4.10 *A parallelepiped determined by the three vectors,* $\mathbf{a},\mathbf{b}$, *and* $\mathbf{c}$ *consists of*

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

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

The following is a picture of such a thing.You notice the area of the base of the parallelepiped,



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

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

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

Example 3.4.11 *Find the volume of the parallelepiped determined by the vectors,* $\mathbf{i}+2\mathbf{j}-5\mathbf{k},\mathbf{i}+3\mathbf{j}-6\mathbf{k},3\mathbf{i}+2\mathbf{j}+3\mathbf{k}$.

According to the above discussion, pick any two of these, take the cross product and then take the dot product of this with the third of these vectors. The result will be either the desired volume or minus the desired volume.

$$(\mathbf{i}+2\mathbf{j}-5\mathbf{k})\times(\mathbf{i}+3\mathbf{j}-6\mathbf{k}) = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 2 & -5 \\ 1 & 3 & -6 \end{vmatrix} = 3\mathbf{i}+\mathbf{j}+\mathbf{k}$$

Now take the dot product of this vector with the third which yields

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

This shows the volume of this parallelepiped is 14 cubic units.

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

Lemma 3.4.12 *Let* $\mathbf{a},\mathbf{b}$, *and* $\mathbf{c}$ *be vectors. Then* $(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c}=\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})$.

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

Notation 3.4.13 *The box product* $\mathbf{a}\times\mathbf{b}\cdot\mathbf{c}=\mathbf{a}\cdot\mathbf{b}\times\mathbf{c}$ *is denoted more compactly as* $[\mathbf{a},\mathbf{b},\mathbf{c}]$.

3.4.3 Another Proof Of The Distributive Law

Here is another proof of the distributive law for the cross product. Let $\mathbf{x}$ be a vector. From the above observation,

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

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

Therefore,

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

for all $\mathbf{x}$. In particular, this holds for $\mathbf{x}=\mathbf{a}\times(\mathbf{b}+\mathbf{c})-(\mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c})$ showing that

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

and this proves the distributive law for the cross product another way.

Observation 3.4.14 *Suppose you have three vectors,* $\mathbf{u}=(a,b,c)$, $\mathbf{v}=(d,e,f)$, *and* $\mathbf{w}=(g,h,i)$. *Then* $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$ *is given by the following.*

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

$$=a\begin{vmatrix}e&f\\ h&i\end{vmatrix}-b\begin{vmatrix}d&f\\ g&i\end{vmatrix}+c\begin{vmatrix}d&e\\ g&h\end{vmatrix}\equiv\det\begin{pmatrix}a&b&c\\ d&e&f\\ g&h&i\end{pmatrix}.$$

The message is that to take the box product, you can simply take the determinant of the matrix which results by letting the rows be the rectangular components of the given vectors in the order in which they occur in the box product. More will be presented on determinants later.

3.5 The Vector Identity Machine

In practice, you often have to deal with combinations of several cross products mixed in with dot products. It is extremely useful to have a technique which will allow you to discover vector identities and simplify expressions involving cross and dot products in three dimensions. This involves two special symbols, δ_{ij} and ε_{ijk} which are very useful in dealing with vector identities. To begin with, here is the definition of these symbols.

Definition 3.5.1 *The symbol δ_{ij}, called the Kronecker 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 Kronecker symbol i and j can equal any integer in $\{1, 2, \cdots, n\}$ for any $n \in \mathbb{N}$.

Definition 3.5.2 *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 $\sum_j \delta_{ij} x_j = x_i$. Thus $\delta_{ij} x_j = x_i$, $\delta_{ii} = 3$, $\delta_{ij} x_{jkl} = x_{ikl}$. When you use this convention, there is one very important thing to never forget. It is this: **Never have an index be repeated more than once**. Thus $a_i b_i$ is all right but $a_{ii} b_i$ is not. The reason for this is that you end up getting confused about what is meant. If you want to write $\sum_i a_i b_i c_i$ it is best to simply use the summation notation. There is a very important reduction identity connecting these two symbols.

Lemma 3.5.3 *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 3.5.4 *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 that 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

$$(\mathbf{u}\times\mathbf{v})_1=(u_2v_3-u_3v_2).$$

From the above formula in the proposition,

$$\varepsilon_{1jk}u_jv_k\equiv u_2v_3-u_3v_2,$$

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

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

Example 3.5.5 *Discover a formula which simplifies* $(\mathbf{u}\times\mathbf{v})\cdot(\mathbf{z}\times\mathbf{w})$, $\mathbf{u},\mathbf{v}\in\mathbb{R}^3$.

From the above description of the cross product and dot product, along with the reduction identity,

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

$$\begin{aligned}\varepsilon_{ijk}u_jv_k\varepsilon_{irs}z_rw_s &= (\delta_{jr}\delta_{ks}-\delta_{js}\delta_{kr})u_jv_kz_rw_s\\ &= u_jv_kz_jw_k-u_jv_kz_kw_j\\ &= (\mathbf{u}\cdot\mathbf{z})(\mathbf{v}\cdot\mathbf{w})-(\mathbf{u}\cdot\mathbf{w})(\mathbf{v}\cdot\mathbf{z})\end{aligned}$$

Example 3.5.6 *Simplify* $\mathbf{u}\times(\mathbf{u}\times\mathbf{v})$.

The i^{th} component is

$$\begin{aligned}\varepsilon_{ijk}u_j(\mathbf{u}\times\mathbf{v})_k &= \varepsilon_{ijk}u_j\varepsilon_{krs}u_rv_s=\varepsilon_{kij}\varepsilon_{krs}u_ju_rv_s\\ &= (\delta_{ir}\delta_{js}-\delta_{jr}\delta_{is})u_ju_rv_s\\ &= u_ju_iv_j-u_ju_jv_i\\ &= (\mathbf{u}\cdot\mathbf{v})u_i-|\mathbf{u}|^2v_i\end{aligned}$$

Hence

$$\mathbf{u}\times(\mathbf{u}\times\mathbf{v})=(\mathbf{u}\cdot\mathbf{v})\mathbf{u}-|\mathbf{u}|^2\mathbf{v}$$

because the i^{th} components of the two sides are equal for any i.

3.6 Exercises

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

2. Find the area of the triangle determined by the three points, $(1,2,3),(4,2,0)$ and $(-3,2,1)$.

3. Find the area of the triangle determined by the three points, $(1,0,3),(4,1,0)$ and $(-3,1,1)$.

4. Find the area of the triangle determined by the three points, $(1,2,3),(2,3,4)$ and $(3,4,5)$. Did something interesting happen here? What does it mean geometrically?

5. Find the area of the parallelogram determined by the vectors, $(1,2,3)$, $(3,-2,1)$.

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

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

8. 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?

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

10. Using Problem 9, find an equation of a plane containing the two position vectors, $\mathbf{a}$ and $\mathbf{b}$ and the point $\mathbf{0}$. **Hint:** If (x,y,z) is a point on this plane the volume of the parallelepiped determined by (x,y,z) and the vectors $\mathbf{a},\mathbf{b}$ equals 0.

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

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

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

12. Is $\mathbf{a}\times(\mathbf{b}\times\mathbf{c})=(\mathbf{a}\times\mathbf{b})\times\mathbf{c}$? What is the meaning of $\mathbf{a}\times\mathbf{b}\times\mathbf{c}$? Explain. **Hint:** Try $(\mathbf{i}\times\mathbf{j})\times\mathbf{j}$.

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

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

15. Simplify $(\mathbf{u}\times\mathbf{v})\cdot(\mathbf{v}\times\mathbf{w})\times(\mathbf{w}\times\mathbf{z})$.

16. Simplify $|\mathbf{u}\times\mathbf{v}|^2+(\mathbf{u}\cdot\mathbf{v})^2-|\mathbf{u}|^2|\mathbf{v}|^2$.

17. For $\mathbf{u},\mathbf{v},\mathbf{w}$ functions of t, $\mathbf{u}'(t)$ is defined as the limit of the difference quotient as in calculus, $(\lim_{h\to 0}\mathbf{w}(h))_i\equiv\lim_{h\to 0}w_i(h)$. Show the following

$$(\mathbf{u}\times\mathbf{v})'=\mathbf{u}'\times\mathbf{v}+\mathbf{u}\times\mathbf{v}',\ (\mathbf{u}\cdot\mathbf{v})'=\mathbf{u}'\cdot\mathbf{v}+\mathbf{u}\cdot\mathbf{v}'$$

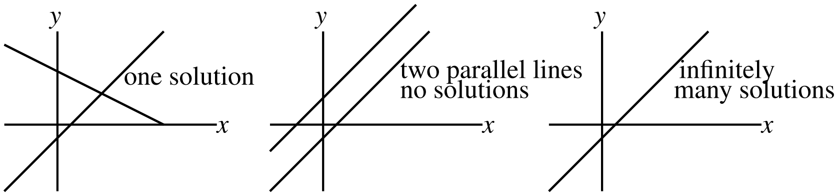
18. If $\mathbf{u}$ is a function of t, and the magnitude $|\mathbf{u}(t)|$ is a constant, show from the above problem that the velocity $\mathbf{u}'$ is perpendicular to $\mathbf{u}$.

19. When you have a rotating rigid body with angular velocity vector $\blacksquare$, then the velocity vector $\mathbf{v} \equiv \mathbf{u}'$ is given by
$$\mathbf{v} = \blacksquare \times \mathbf{u}$$
where $\mathbf{u}$ is a position vector. The acceleration is the derivative of the velocity. Show that if $\blacksquare$ is a constant vector, then the acceleration vector $\mathbf{a} = \mathbf{v}'$ is given by the formula
$$\mathbf{a} = \blacksquare \times (\blacksquare \times \mathbf{u}).$$
Now simplify the expression. It turns out this is centripetal acceleration.

20. Verify directly that the coordinate description of the cross product, $\mathbf{a} \times \mathbf{b}$ has the property that it is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$. Then show by direct computation that this coordinate description satisfies
$$\begin{aligned} |\mathbf{a} \times \mathbf{b}|^2 &= |\mathbf{a}|^2 |\mathbf{b}|^2 - (\mathbf{a} \cdot \mathbf{b})^2 \\ &= |\mathbf{a}|^2 |\mathbf{b}|^2 \left(1 - \cos^2(\theta)\right) \end{aligned}$$
where θ is the angle included between the two vectors. Explain why $|\mathbf{a} \times \mathbf{b}|$ has the correct magnitude. All that is missing is the material about the right hand rule. Verify directly from the coordinate description of the cross product that the right thing happens with regards to the vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$. Next verify that the distributive law holds for the coordinate description of the cross product. This gives another way to approach the cross product. First define it in terms of coordinates and then get the geometric properties from this. However, this approach does not yield the right hand rule property very easily.

3.7 Systems Of Equations, Geometry

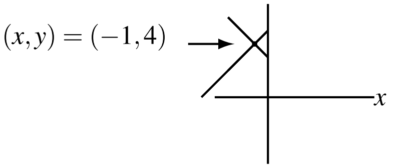
As you know, equations like $2x + 3y = 6$ can be graphed as straight lines in $\mathbb{R}^2$. To find the solution to two such equations, you could graph the two straight lines and the ordered pairs identifying the point (or points) of intersection would give the x and y values of the solution to the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



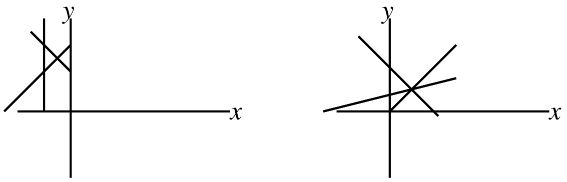
In the first example of the above picture, there is a unique point of intersection. In the second, there are no points of intersection. The other thing which can occur is that the two lines are really the same line. For example, $x + y = 1$ and $2x + 2y = 2$ are relations which when graphed yield the same line. In this case there are infinitely many points in the simultaneous solution of these two equations, every ordered pair which is on the graph of the line. It is always this way when considering linear systems of equations. There is either no solution, exactly one or infinitely many although the reasons for this are not completely comprehended by considering a simple picture in two dimensions, $\mathbb{R}^2$.

Example 3.7.1 *Find the solution to the system $x+y=3$, $y-x=5$.*

You can verify the solution is $(x,y)=(-1,4)$. You can see this geometrically by graphing the equations of the two lines. If you do so correctly, you should obtain a graph which looks something like the following in which the point of intersection represents the solution of the two equations.



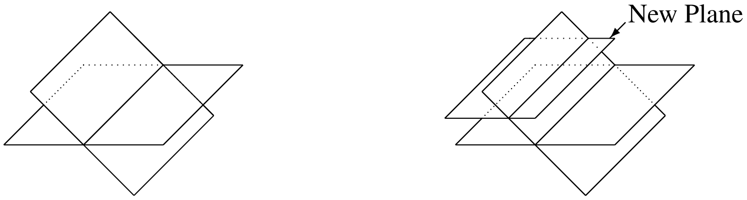
Example 3.7.2 *You can also imagine other situations such as the case of three intersecting lines having no common point of intersection or three intersecting lines which do intersect at a single point as illustrated in the following picture.*



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

The points, (x,y,z) satisfying an equation in three variables like $2x+4y-5z=8$ form a plane [1] and geometrically, when you solve systems of equations involving three variables, you are taking intersections of planes.

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



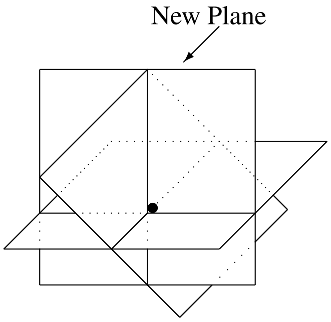
Now imagine a third plane. One thing that could happen is this third plane could have an intersection with one of the first planes which results in a line which fails to intersect the first line as illustrated in the following picture.

Thus there is no point which lies in all three planes. The picture illustrates the situation in which the line of intersection of the new plane with one of the original planes forms a

[1] Don't worry about why this is at this time. It is not important. The discussion is intended to show you that geometric considerations like this don't take you anywhere. It is the algebraic procedures which are important and lead to important applications.

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

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



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

Relations like $x+y-2z+4w=8$ are often called **hyperplanes**.[2] However, it is impossible to draw pictures of such things. The only rational and useful way to deal with this subject is through the use of algebra not art. Mathematics exists partly to free us from having to always draw pictures in order to draw conclusions. The next chapter gives useful procedures which do not depend on pictures for finding solutions to systems of equations.

[2] The evocative semi word, "hyper" conveys absolutely no meaning but is traditional usage which makes the terminology sound more impressive than something like long wide flat thing.Later we will discuss some terms which are not just evocative but yield real understanding.

Chapter 4

Systems Of Equations

4.1 Systems Of Equations, Algebraic Procedures

4.1.1 Elementary Operations

Consider the following example.

Example 4.1.1 *Find x and y such that*

$$x+y=7 \text{ and } 2x-y=8. \tag{4.1}$$

The set of ordered pairs, (x,y) which solve both equations is called the ***solution set****.*

You can verify that $(x,y)=(5,2)$ is a solution to the above system. The interesting question is this: If you were not given this information to verify, how could you determine the solution? You can do this by using the following basic operations on the equations, none of which change the set of solutions of the system of equations.

Definition 4.1.2 ***Elementary operations*** *are those operations consisting of the following.*

1. *Interchange the order in which the equations are listed.*
2. *Multiply any equation by a nonzero number.*
3. *Replace any equation with itself added to a multiple of another equation.*

Example 4.1.3 *To illustrate the third of these operations on this particular system, consider the following.*

$$\begin{aligned} x+y&=7 \\ 2x-y&=8 \end{aligned}$$

The system has the same solution set as the system

$$\begin{aligned} x+y&=7 \\ -3y&=-6 \end{aligned}.$$

To obtain the second system, take the second equation of the first system and add -2 times the first equation to obtain

$$-3y = -6.$$

Now, this clearly shows that $y = 2$ and so it follows from the other equation that $x+2 = 7$ and so $x = 5$.

Of course a linear system may involve many equations and many variables. The solution set is still the collection of solutions to the equations. In every case, the above operations of Definition 4.1.2 do not change the set of solutions to the system of linear equations.

Theorem 4.1.4 *Suppose you have two equations, involving the variables,*

$$(x_1, \cdots, x_n)$$

$$E_1 = f_1, E_2 = f_2 \tag{4.2}$$

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

$$E_1 = f_1,\ E_2 + aE_1 = f_2 + af_1. \tag{4.3}$$

Also the system $E_1 = f_1, E_2 = f_2$ has the same solutions as the system, $E_2 = f_2, E_1 = f_1$. The system $E_1 = f_1, E_2 = f_2$ has the same solution as the system $E_1 = f_1, aE_2 = af_2$ provided $a \neq 0$.

Proof: If $(x_1, \cdots, x_n)$ solves $E_1 = f_1, E_2 = f_2$ then it solves the first equation in $E_1 = f_1$, $E_2 + aE_1 = f_2 + af_1$. Also, it satisfies $aE_1 = af_1$ and so, since it also solves $E_2 = f_2$ it must solve $E_2 + aE_1 = f_2 + af_1$. Therefore, if $(x_1, \cdots, x_n)$ solves $E_1 = f_1, E_2 = f_2$ it must also solve $E_2 + aE_1 = f_2 + af_1$. On the other hand, if it solves the system $E_1 = f_1$ and $E_2 + aE_1 = f_2 + af_1$, then $aE_1 = af_1$ and so you can subtract these equal quantities from both sides of $E_2 + aE_1 = f_2 + af_1$ to obtain $E_2 = f_2$ showing that it satisfies $E_1 = f_1, E_2 = f_2$.

The second assertion of the theorem which says that the system $E_1 = f_1, E_2 = f_2$ has the same solution as the system, $E_2 = f_2, E_1 = f_1$ is seen to be true because it involves nothing more than listing the two equations in a different order. They are the same equations.

The third assertion of the theorem which says $E_1 = f_1, E_2 = f_2$ has the same solution as the system $E_1 = f_1, aE_2 = af_2$ provided $a \neq 0$ is verified as follows: If $(x_1, \cdots, x_n)$ is a solution of $E_1 = f_1, E_2 = f_2$, then it is a solution to $E_1 = f_1, aE_2 = af_2$ because the second system only involves multiplying the equation, $E_2 = f_2$ by a. If $(x_1, \cdots, x_n)$ is a solution of $E_1 = f_1, aE_2 = af_2$, then upon multiplying $aE_2 = af_2$ by the number $1/a$, you find that $E_2 = f_2$. ■

Stated simply, the above theorem shows that the elementary operations do not change the solution set of a system of equations.

Here is an example in which there are three equations and three variables. You want to find values for x, y, z such that each of the given equations are satisfied when these values are plugged in to the equations.

Example 4.1.5 *Find the solutions to the system,*

$$\begin{aligned} x + 3y + 6z &= 25 \\ 2x + 7y + 14z &= 58 \\ 2y + 5z &= 19 \end{aligned} \tag{4.4}$$

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

$$\begin{aligned} x+3y+6z &= 25 \\ y+2z &= 8 \\ 2y+5z &= 19 \end{aligned} \tag{4.5}$$

Now take (-2) times the second and add to the third. More precisely, replace the third equation with (-2) times the second added to the third. This yields the system

$$\begin{aligned} x+3y+6z &= 25 \\ y+2z &= 8 \\ z &= 3 \end{aligned} \tag{4.6}$$

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

Alternatively, in 4.6 you could have continued as follows. Add (-2) times the bottom equation to the middle and then add (-6) times the bottom to the top. This yields

$$x+3y=7,\ y=2,\ z=3$$

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

$$x=1,\ y=2,\ z=3,$$

a system which has the same solution set as the original system. This avoided back substitution and led to the same solution set.

4.1.2 Gauss Elimination

A less cumbersome way to represent a linear system is to write it as an **augmented matrix.** For example the linear system, 4.4 can be written as

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

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 4.4 would be to take (-2) times the first row of the augmented matrix above and add it to the second row,

$$\left(\begin{array}{ccc|c} 1 & 3 & 6 & 25 \\ 0 & 1 & 2 & 8 \\ 0 & 2 & 5 & 19 \end{array}\right).$$

Note how this corresponds to 4.5. Next take (-2) times the second row and add to the third,

$$\left(\begin{array}{ccc|c} 1 & 3 & 6 & 25 \\ 0 & 1 & 2 & 8 \\ 0 & 0 & 1 & 3 \end{array}\right)$$

This augmented matrix corresponds to the system

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

which is the same as 4.6. By back substitution you obtain the solution $x = 1, y = 6$, and $z = 3$.

In general a linear system is of the form

$$\begin{array}{c} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \end{array}, \tag{4.7}$$

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

$$\left(\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right). \tag{4.8}$$

Changes to the system of equations in 4.7 as a result of an elementary operation translate into changes of the augmented matrix resulting from a row operation. Note that Theorem 4.1.4 implies that the row operations deliver an augmented matrix for a system of equations which has the same solution set as the original system.

Definition 4.1.6 *The* ***row operations*** *consist of the following*

1. *Switch two rows.*
2. *Multiply a row by a nonzero number.*
3. *Replace a row by a multiple of another row added to it.*

Gauss elimination is a systematic procedure to simplify an augmented matrix to a reduced form. In the following definition, the term "**leading entry**" refers to the first nonzero entry of a row when scanning the row from left to right.

Definition 4.1.7 *An augmented matrix is in* ***echelon form*** *if*

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

How do you know when to stop doing row operations? You might stop when you have obtained an echelon form as described above, but you certainly should stop doing row operations if you have gotten a matrix in row reduced echelon form described next.

Definition 4.1.8 *An augmented matrix is in* ***row reduced echelon form*** *if*

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

Example 4.1.9 *Here are some matrices which are in row reduced echelon form.*

$$\begin{pmatrix} 1 & 0 & 0 & 5 & 8 & 0 \\ 0 & 0 & 1 & 2 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Example 4.1.10 *Here are matrices in echelon form which are not in row reduced echelon form but which are in echelon form.*

$$\begin{pmatrix} 1 & 0 & 6 & 5 & 8 & 2 \\ 0 & 0 & 2 & 2 & 7 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 3 & 5 & 4 \\ 0 & 2 & 0 & 7 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Example 4.1.11 *Here are some matrices which are not in echelon form.*

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 2 & 3 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & -6 \\ 4 & 0 & 7 \end{pmatrix}, \begin{pmatrix} 0 & 2 & 3 & 3 \\ 1 & 5 & 0 & 2 \\ 7 & 5 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Definition 4.1.12 *A* ***pivot position*** *in a matrix is the location of a leading entry in an echelon form resulting from the application of row operations to the matrix. A* ***pivot column*** *is a column that contains a pivot position.*

For example consider the following.

Example 4.1.13 *Suppose*

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 6 \\ 4 & 4 & 4 & 10 \end{pmatrix}$$

Where are the pivot positions and pivot columns?

Replace the second row by -3 times the first added to the second. This yields

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -6 \\ 4 & 4 & 4 & 10 \end{pmatrix}.$$

This is not in reduced echelon form so replace the bottom row by -4 times the top row added to the bottom. This yields

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -6 \\ 0 & -4 & -8 & -6 \end{pmatrix}.$$

This is still not in reduced echelon form. Replace the bottom row by -1 times the middle row added to the bottom. This yields

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -6 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

which is in echelon form, although not in reduced echelon form. Therefore, the pivot positions in the original matrix are the locations corresponding to the first row and first column and the second row and second columns as shown in the following:

$$\begin{pmatrix} \boxed{1} & 2 & 3 & 4 \\ 3 & \boxed{2} & 1 & 6 \\ 4 & 4 & 4 & 10 \end{pmatrix}$$

Thus the pivot columns in the matrix are the first two columns.

The following is the algorithm for obtaining a matrix which is in row reduced echelon form.

Algorithm 4.1.14

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

1. Find the first nonzero column from the left. This is the first pivot column. The position at the top of the first pivot column is the first pivot position. Switch rows if necessary to place a nonzero number in the first pivot position.

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

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

Example 4.1.15 *Here is a matrix.*

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

Do row reductions till you obtain a matrix in echelon form. Then complete the process by producing one in row reduced echelon form.

The pivot column is the second. Hence the pivot position is the one in the first row and second column. Switch the first two rows to obtain a nonzero entry in this pivot position.

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

Step two is not necessary because all the entries below the first pivot position in the resulting matrix are zero. Now ignore the top row and the columns to the left of this first pivot position. Thus you apply the same operations to the smaller matrix

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

The next pivot column is the third corresponding to the first in this smaller matrix and the second pivot position is therefore, the one which is in the second row and third column.

In this case it is not necessary to switch any rows to place a nonzero entry in this position because there is already a nonzero entry there. Multiply the third row of the original matrix by -2 and then add the second row to it. This yields

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

The next matrix the steps in the algorithm are applied to is

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

The first pivot column is the first column in this case and no switching of rows is necessary because there is a nonzero entry in the first pivot position. Therefore, the algorithm yields for the next step

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

Now the algorithm will be applied to the matrix

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

There is only one column and it is nonzero so this single column is the pivot column. Therefore, the algorithm yields the following matrix for the echelon form.

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

To complete placing the matrix in reduced echelon form, multiply the third row by 3 and add -2 times the fourth row to it. This yields

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

Next multiply the second row by 3 and take 2 times the fourth row and add to it. Then add the fourth row to the first.

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

Next work on the fourth column in the same way.

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

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

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

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

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

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

Example 4.1.16 *Give the complete solution to the system of equations,* $5x+10y-7z=-2$, $2x+4y-3z=-1$, *and* $3x+6y+5z=9$.

The augmented matrix for this system is

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

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

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

Switch the last two rows to get

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

then take -3 times the top added to 2 times the middle to get

$$\begin{pmatrix} -6 & -12 & 9 & 3 \\ 0 & 0 & 19 & 21 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

Multiply bottom by 19 and take the second row times -1 added to the bottom. This gives

$$\begin{pmatrix} -6 & -12 & 9 & 3 \\ 0 & 0 & 19 & 21 \\ 0 & 0 & 0 & -2 \end{pmatrix}$$

a row of zeros with -2 at the right end, representing the equation $0x+0y+0z=-2$ which has no solution so there is no solution to this system of equations. When this happens, the system is called **inconsistent**. In this case it is very easy to describe the solution set. The system has no solution.

Here is another example based on the use of row operations.

Example 4.1.17 *Give the complete solution to the system of equations,* $3x-y-5z=9$, $y-10z=0$, *and* $-2x+y=-6$.

The augmented matrix of this system is

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

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

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

The entry, 3 in this sequence of row operations is called the **pivot**. It is used to create zeros in the other places of the column. Next take -1 times the middle row and add to the bottom. Here the 1 in the second row is the pivot.

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

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

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

This is in reduced echelon form. The equations corresponding to this reduced echelon form are $y = 10z$ and $x = 3+5z$. Apparently z can equal any number. Lets call this number t. [1]Therefore, the solution set of this system is $x = 3+5t, y = 10t$, and $z = t$ where t is completely arbitrary. The system has an infinite set of solutions which are given in the above simple way. This is what it is all about, finding the solutions to the system.

There is some terminology connected to this which is useful. Recall how each column corresponds to a variable in the original system of equations. The variables corresponding to a pivot column are called **basic variables**. The other variables are called **free variables.** In Example 4.1.17 there was one free variable, z, and two basic variables, x and y. In describing the solution to the system of equations, the free variables are assigned a parameter. In Example 4.1.17 this parameter was t. Sometimes there are many free variables and in these cases, you need to use many parameters. Here is another example.

Example 4.1.18 *Find the solution to the system*

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

The augmented matrix is

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

Take -1 times the first row and add to the second. Then take -1 times the first row and add to the third. This yields

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

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

Now add the second row to the bottom row

$$\left(\begin{array}{ccccc} 1 & 2 & -1 & 1 & 3 \\ 0 & -1 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right) \tag{4.9}$$

This matrix is in echelon form and you see the basic variables are x and y while the free variables are z and w. Assign s to z and t to w. Then the second row yields the equation, $y = 2$ while the top equation yields the equation, $x + 2y - s + t = 3$ and so since $y = 2$, this gives $x + 4 - s + t = 3$ showing that $x = -1 + s - t, y = 2, z = s$, and $w = t$. One can write this in the form

$$\left(\begin{array}{c} x \\ y \\ z \\ w \end{array}\right) = \left(\begin{array}{c} -1 + s - t \\ 2 \\ s \\ t \end{array}\right). \tag{4.10}$$

This is another example of a system which has an infinite solution set but this time the solution set depends on two parameters, not one. Most people find it less confusing in the case of an infinite solution set to first place the augmented matrix in row reduced echelon form rather than just echelon form before seeking to write down the description of the solution. In the above, this means we don't stop with the echelon form 4.9. Instead we first place it in reduced echelon form as follows.

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

Then the solution is $y = 2$ from the second row and $x = -1 + z - w$ from the first. Thus letting $z = s$ and $w = t$, the solution is given in 4.10.

The number of free variables is always equal to the number of **different** parameters used to describe the solution. If there are no free variables, then either there is no solution as in the case where row operations yield an echelon form like

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

or there is a unique solution as in the case where row operations yield an echelon form like

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

Also, sometimes there are free variables and no solution as in the following:

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

There are a lot of cases to consider but it is not necessary to make a major production of this. Do row operations till you obtain a matrix in echelon form or reduced echelon form and determine whether there is a solution. If there is, see if there are free variables. In this case, there will be infinitely many solutions. Find them by assigning different parameters to the free variables and obtain the solution. If there are no free variables, then there will be a unique solution which is easily determined once the augmented matrix is in echelon or row reduced echelon form. In every case, the process yields a straightforward way to describe the solutions to the linear system. As indicated above, you are probably less likely to become confused if you place the augmented matrix in row reduced echelon form rather than just echelon form.

In summary,

Definition 4.1.19 *A **system of linear equations** is a list of equations,*

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

where a_{ij} are numbers, and b_j is a number. The above is a system of m equations in the n variables, $x_1, x_2 \cdots, x_n$. Nothing is said about the relative size of m and n. Written more simply in terms of summation notation, the above can be written in the form

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

It is desired to find $(x_1,\cdots,x_n)$ solving each of the equations listed.

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

Definition 4.1.20 *A system of linear equations is called **consistent** if there exists a solution. It is called **inconsistent** if there is no solution.*

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

4.1.3 Balancing Chemical Reactions

Consider the chemical reaction

$$SnO_2 + H_2 \to Sn + H_2O$$

Here the elements involved are tin Sn oxygen O and Hydrogen H. Some chemical reaction happens and you end up with some tin and some water. The question is, how much do you start with and how much do you end up with.

The balance of mass requires that you have the same number of oxygen, tin, and hydrogen on both sides of the reaction. However, this does not happen in the above. For example, there are two oxygen atoms on the left and only one on the right. The problem is to find numbers x, y, z, w such that

$$xSnO_2 + yH_2 \to zSn + wH_2O$$

and both sides have the same number of atoms of the various substances. You can do this in a systematic way by setting up a system of equations which will require that this take place. Thus you need

$$\begin{aligned} Sn: &\quad x = z \\ O: &\quad 2x = w \\ H: &\quad 2y = 2w \end{aligned}$$

The augmented matrix for this system of equations is then

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

Row reducing this yields

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

Thus you could let $w = 2$ and this would yield $x = 1, y = 2$, and $z = 1$. Hence, the description of the reaction which has the same numbers of atoms on both sides would be

$$SnO_2 + 2H_2 \to Sn + 2H_2O$$

You see that this preserves the total number of atoms and so the chemical equation is balanced. Here is another example

Example 4.1.21 *Potassium is denoted by K, oxygen by O, phosphorus by P and hydrogen by H. The reaction is*

$$KOH + H_3PO_4 \to K_3PO_4 + H_2O$$

balance this equation.

You need to have

$$xKOH + yH_3PO_4 \to zK_3PO_4 + wH_2O$$

Equations which preserve the total number of atoms of each element on both sides of the equation are

$$\begin{array}{rc} K: & x = 3z \\ O: & x + 4y = 4z + w \\ H: & x + 3y = 2w \\ P: & y = z \end{array}$$

The augmented matrix for this system is

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

Then the row reduced echelon form is

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

You could let $w = 3$ and this yields $x = 3, y = 1, z = 1$. Then the balanced equation is

$$3KOH + 1H_3PO_4 \to 1K_3PO_4 + 3H_2O$$

Note that this results in the same number of atoms on both sides.

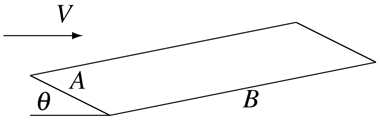
Of course these numbers you are finding would typically be the number of moles of the molecules on each side. Thus three moles of KOH added to one mole of H_3PO_4 yields one mole of K_3PO_4 and three moles of H_2O, water.

Note that in this example, you have a row of zeros. This means that some of the information in computing the appropriate numbers was redundant. If this can happen with a single reaction, think how much more it could happen if you were dealing with hundreds of reactions. This aspect of the problem can be understood later in terms of the rank of a matrix.

For an introduction to the chemical considerations mentioned here, there is a nice site on the web http://chemistry.about.com/od/chemicalreactions/a/reactiontypes.htm where there is a sample test and examples of chemical reactions. For names of the various elements symbolized by the various letters, you can go to the site
http://chemistry.about.com/od/elementfacts/a/elementlist.htm. Of course these things are in standard chemistry books, but if you have not seen much chemistry, these sites give a nice introduction to these concepts.

4.1.4 Dimensionless Variables*

This section shows how solving systems of equations can be used to determine appropriate dimensionless variables. It is only an introduction to this topic. I got this example from [7]. This considers a specific example of a simple airplane wing shown below. We assume for simplicity that it is just a flat plane at an angle to the wind which is blowing against it with speed V as shown.



The angle is called the angle of incidence, B is the span of the wing and A is called the chord. Denote by l the lift. Then this should depend on various quantities like θ, V, B, A and so forth. Here is a table which indicates various quantities on which it is reasonable to expect l to depend.

Variable	Symbol	Units
chord	A	m
span	B	m
angle incidence	θ	$m^0 kg^0 \sec^0$
speed of wind	V	$m \sec^{-1}$
speed of sound	V_0	$m \sec^{-1}$
density of air	ρ	kgm^{-3}
viscosity	μ	$kg \sec^{-1} m^{-1}$
lift	l	$kg \sec^{-2} m$

Here m denotes meters, sec refers to seconds and kg refers to kilograms. All of these are likely familiar except for μ. One can simply decree that these are the dimensions of something called viscosity but it might be better to consider this a little more.

Viscosity is a measure of how much internal friction is experienced when the fluid moves. It is roughly a measure of how “sticky” the fluid is. Consider a piece of area parallel to the direction of motion of the fluid. To say that the viscosity is large is to say that the tangential force applied to this area must be large in order to achieve a given change in speed of the fluid in a direction normal to the tangential force. Thus

$$\mu\,(\text{area})\,(\text{velocity gradient}) = \text{ tangential force.}$$

Hence

$$(\text{units on } \mu)\, m^2 \left(\frac{m}{\sec m}\right) = kg \sec^{-2} m$$

Thus the units on μ are $kg \sec^{-1} m^{-1}$ as claimed above.

Then one would think that you would want

$$l = f\,(A, B, \theta, V, V_0, \rho, \mu)$$

However, this is very cumbersome because it depends on seven variables. Also, it doesn't make very good sense. It is likely that without much care, a change in the units such as going from meters to feet would result in an incorrect value for l. The way to get around this problem is to look for l as a function of dimensionless variables multiplied by something which has units of force. It is helpful because first of all, you will likely have fewer independent variables and secondly, you could expect the formula to hold independent of the way of specifying length, mass and so forth. One looks for

$$l = f\,(g_1, \cdots, g_k)\, \rho V^2 AB$$

where the units on $\rho V^2 AB$ are

$$\frac{kg}{m^3}\left(\frac{m}{\sec}\right)^2 m^2 = \frac{kg\times m}{\sec^2}$$

which are the units of force. Each of these g_i is of the form

$$A^{x_1}B^{x_2}\theta^{x_3}V^{x_4}V_0^{x_5}\rho^{x_6}\mu^{x_7} \tag{4.11}$$

and each g_i is independent of the dimensions. That is, this expression must not depend on meters, kilograms, seconds, etc. Thus, placing in the units for each of these quantities, one needs

$$m^{x_1}m^{x_2}\left(m^{x_4}\sec^{-x_4}\right)\left(m^{x_5}\sec^{-x_5}\right)\left(kgm^{-3}\right)^{x_6}\left(kg\sec^{-1}m^{-1}\right)^{x_7} = m^0 kg^0 \sec^0$$

Notice that there are no units on θ because it is just the radian measure of an angle. Hence its dimensions consist of length divided by length, thus it is dimensionless. Then this leads to the following equations for the x_i.

$$\begin{array}{rc} m: & x_1+x_2+x_4+x_5-3x_6-x_7=0 \\ \sec: & -x_4-x_5-x_7=0 \\ kg: & x_6+x_7=0 \end{array}$$

Then the augmented matrix for this system of equations is

$$\left(\begin{array}{cccccccc} 1 & 1 & 0 & 1 & 1 & -3 & -1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{array}\right)$$

The row reduced echelon form is then

$$\left(\begin{array}{cccccccc} 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{array}\right)$$

and so the solutions are of the form

$$x_1 = -x_2 - x_7,\ x_3 = x_3, x_4 = -x_5 - x_7, x_6 = -x_7$$

Thus, in terms of vectors, the solution is

$$\begin{array}{rl} & \left(\begin{array}{ccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \end{array}\right) \\ = & \left(\begin{array}{ccccccc} -x_2-x_7 & x_2 & x_3 & -x_5-x_7 & x_5 & -x_7 & x_7 \end{array}\right) \end{array}$$

Thus the free variables are x_2, x_3, x_5, x_7. By assigning values to these, we can obtain dimensionless variables by placing the values obtained for the x_i in the formula 4.11. For example, let $x_2 = 1$ and all the rest of the free variables are 0. This yields

$$x_1 = -1, x_2 = 1, x_3 = 0, x_4 = 0, x_5 = 0, x_6 = 0, x_7 = 0.$$

The dimensionless variable is then $A^{-1}B^{1}$. This is the ratio between the span and the chord. It is called the aspect ratio, denoted as AR. Next let $x_3 = 1$ and all others equal zero. This gives for a dimensionless quantity the angle θ. Next let $x_5 = 1$ and all others equal zero. This gives

$$x_1 = 0, x_2 = 0, x_3 = 0, x_4 = -1, x_5 = 1, x_6 = 0, x_7 = 0.$$

Then the dimensionless variable is $V^{-1}V_0^1$. However, it is written as V/V_0. This is called the Mach number $\mathscr{M}$. Finally, let $x_7 = 1$ and all the other free variables equal 0. Then

$$x_1 = -1, x_2 = 0, x_3 = 0, x_4 = -1, x_5 = 0, x_6 = -1, x_7 = 1$$

then the dimensionless variable which results from this is $A^{-1}V^{-1}\rho^{-1}\mu$. It is customary to write it as $\mathrm{Re} = (AV\rho)/\mu$. This one is called the Reynolds number. It is the one which involves viscosity. Thus we would look for

$$l = f(\mathrm{Re}, AR, \theta, \mathscr{M})\, kg \times m/\sec^2$$

This is quite interesting because it is easy to vary Re by simply adusting the velocity or A but it is hard to vary things like μ or ρ. Note that all the quantities are easy to adjust. Now this could be used, along with wind tunnel experiments to get a formula for the lift which would be reasonable. Obviously, you could consider more variables and more complicated situations in the same way.

4.2 MATLAB And Row Reduced Echelon Form

MATLAB will find the row reduced echelon form of a matrix and save you the trouble of tedious computations. You open matlab. You will see >>. Then next to it you type the following:

```
rref([1,2,3,4;2,5,6,10;3,2,0,-5])
```

Then press enter on your keyboard. It will give the following.

```
ans =
1  0 0 -3
0 1 0 2
0 0 1 1
```

In usual notation, this is the row reduced echelon form of the matrix

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 5 & 6 & 10 \\ 3 & 2 & 0 & -5 \end{pmatrix}$$

Notice how you enter a row by placing commas between entries and then when you start a new row, you put a ; to indicate it is a new row. You do something similar for another matrix. You can also simply leave a space between the entries of a row and it will know what to do, but you indicate a new row by using ;. The semicolon ; is also used to defer an operation. MATLAB will know about it but won't do anything.

In using MATLAB, you press shift enter to go to a new line. One thing might be helpful to mention about MATLAB. It is very good at manipulating matrices and vectors and there is distinctive notation used to accomplish this. For example say you type

```
x=[1,2,3]; y=[2,3,4]; x.*y
```

and then press “enter”. You will get 2,6,12. You would get an error if you wrote x*y. Similarly, type

```
[2,4,6,8]./[1,2,3,4]
```

and press “enter”. This yields $2,2,2,2$. The expression [2,4,6,8]/[1,2,3,4] doesn’t make any sense.

4.3 Exercises

1. Find the point (x_1, y_1) which lies on both lines, $x+3y=1$ and $4x-y=3$.
2. Solve Problem 1 graphically. That is, graph each line and see where they intersect.
3. Find the point of intersection of the two lines $3x+y=3$ and $x+2y=1$.
4. Solve Problem 3 graphically. That is, graph each line and see where they intersect.
5. Do the three lines, $x+2y=1, 2x-y=1$, and $4x+3y=3$ have a common point of intersection? If so, find the point and if not, tell why they don’t have such a common point of intersection.
6. Do the three planes, $x+y-3z=2$, $2x+y+z=1$, and $3x+2y-2z=0$ have a common point of intersection? If so, find one and if not, tell why there is no such point.
7. You have a system of k equations in two variables, $k \geq 2$. Explain the geometric significance of

   (a) No solution.

   (b) A unique solution.

   (c) An infinite number of solutions.

8. Here is an augmented matrix in which $*$ denotes an arbitrary number and $\blacksquare$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

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

9. Here is an augmented matrix in which $*$ denotes an arbitrary number and $\blacksquare$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

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

10. Here is an augmented matrix in which $*$ denotes an arbitrary number and $\blacksquare$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

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

11. Here is an augmented matrix in which $*$ denotes an arbitrary number and $\blacksquare$ denotes a nonzero number. Determine whether the given augmented matrix is consistent. If consistent, is the solution unique?

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

12. Suppose a system of equations has fewer equations than variables. Must such a system be consistent? If so, explain why and if not, give an example which is not consistent.

13. If a system of equations has more equations than variables, can it have a solution? If so, give an example and if not, tell why not.

14. Find h such that

$$\left(\begin{array}{cc|c} 2 & h & 4 \\ 3 & 6 & 7 \end{array}\right)$$

is the augmented matrix of an inconsistent matrix.

15. Find h such that

$$\left(\begin{array}{cc|c} 1 & h & 3 \\ 2 & 4 & 6 \end{array}\right)$$

is the augmented matrix of a consistent matrix.

16. Find h such that

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

is the augmented matrix of a consistent matrix.

17. Choose h and k such that the augmented matrix shown has one solution. Then choose h and k such that the system has no solutions. Finally, choose h and k such that the system has infinitely many solutions.

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

18. Choose h and k such that the augmented matrix shown has one solution. Then choose h and k such that the system has no solutions. Finally, choose h and k such that the system has infinitely many solutions.

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

19. Determine if the system is consistent. If so, is the solution unique?

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

20. Determine if the system is consistent. If so, is the solution unique?

$$\begin{gathered} x+2y+z-w=2 \\ x-y+z+w=0 \\ 2x+y-z=1 \\ 4x+2y+z=3 \end{gathered}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

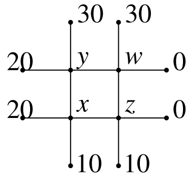
33. Give the complete solution to the system of equations, $-9x+15y=66, -11x+18y=79$,$-x+y=4$, and $z=3$.

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

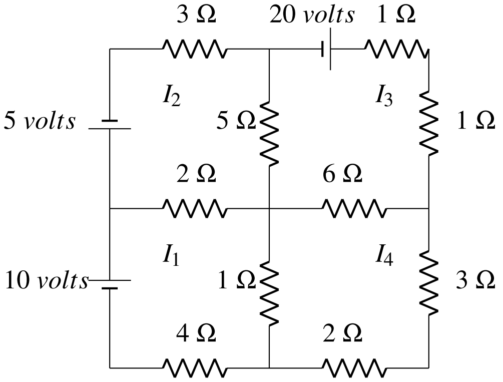
35. Consider the system $-5x+2y-z=0$ and $-5x-2y-z=0$. Both equations equal zero and so $-5x+2y-z=-5x-2y-z$ which is equivalent to $y=0$. Thus x and z can equal anything. But when $x=1$, $z=-4$, and $y=0$ are plugged in to the equations, it doesn't work. Why?

36. 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 sisters.

37. The steady state temperature, u in a plate solves Laplace's equation, $\Delta u=0$. One way to approximate the solution which is often used is to divide the plate into a square mesh and require the temperature at each node to equal the average of the temperature at the four adjacent nodes. This procedure is justified by the mean value property of harmonic functions. In the following picture, the numbers represent the observed temperature at the indicated nodes. Your task is to find the temperature at the interior nodes, indicated by $x,y,z,$ and w. One of the equations is $z=\frac{1}{4}(10+0+w+x)$.



38. Consider the following diagram of four circuits.



Those jagged places denote resistors and the numbers next to them give their resistance in ohms, written as Ω. The breaks in the lines having one short line and one long line denote a voltage source which causes the current to flow in the direction which goes from the longer of the two lines toward the shorter along the unbroken part of the circuit. The current in amps in the four circuits is denoted by I_1, I_2, I_3, I_4 and it is understood that the motion is in the counter clockwise direction. If I_k ends up being negative, then it just means the current flows in the clockwise direction. Then Kirchhoff's law states that

The sum of the resistance times the amps in the counter clockwise direction around a loop equals the sum of the voltage sources in the same direction around the loop.

In the above diagram, the top left circuit should give the equation

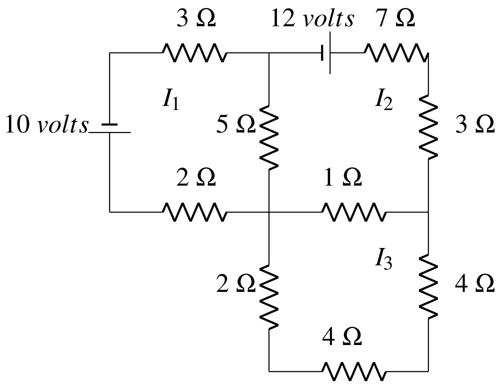
$$2I_2 - 2I_1 + 5I_2 - 5I_3 + 3I_2 = 5$$

For the circuit on the lower left, you should have

$$4I_1 + I_1 - I_4 + 2I_1 - 2I_2 = -10$$

Write equations for each of the other two circuits and then give a solution to the resulting system of equations. You might use a computer algebra system to find the solution. It might be more convenient than doing it by hand.

39. Consider the following diagram of three circuits.



Those jagged places denote resistors and the numbers next to them give their resistance in ohms, written as Ω. The breaks in the lines having one short line and one long line denote a voltage source which causes the current to flow in the direction which goes from the longer of the two lines toward the shorter along the unbroken part of the circuit. The current in amps in the four circuits is denoted by I_1, I_2, I_3 and it is understood that the motion is in the counter clockwise direction. If I_k ends up being negative, then it just means the current flows in the clockwise direction. Then Kirchhoff's law states that

The sum of the resistance times the amps in the counter clockwise direction around a loop equals the sum of the voltage sources in the same direction around the loop. Find I_1, I_2, I_3.

40. Here are some chemical reactions. Balance them.

    (a) $KNO_3 + H_2CO_3 \to K_2CO_3 + HNO_3$

    (b) $Ba_3N_2 + H_2O \to Ba(OH)_2 + NH_3$

    (c) $CaCl_2 + Na_3PO_4 \to Ca_3(PO_4)_2 + NaCl$

41. In the section on dimensionless variables 71 it was observed that $\rho V^2 AB$ has the units of force. Describe a systematic way to obtain such combinations of the variables which will yield something which has the units of force.

42. A system of equations for x, y, z, w has augmented matrix

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

Such systems are called homogenous systems because the column on the right is all zeros. Find all solutions to the system of equations having this as augmented matrix. Using the row reduced echelon form, show that the solutions to the underlying system of equations are of the form

$$\begin{pmatrix} -2z \\ 3z-4w \\ z \\ w \end{pmatrix} = z\begin{pmatrix} -2 \\ 3 \\ 1 \\ 0 \end{pmatrix} + w\begin{pmatrix} 0 \\ -4 \\ 0 \\ 1 \end{pmatrix}$$

if we define the operations in the obvious way adding corresponding entries as done in calculus or in the earlier chapters of this book. Here the free variables are z, w and these can take any value to yield a solution. The two vectors in the above are called a basis for the solution set. The idea of a basis will be described carefully later in the book.

43. Using the same approach as the above problem, find a basis for the solution set to the system of equations having

$$\begin{pmatrix} 1 & -1 & 5 & 1 & 0 \\ -1 & 2 & -8 & 1 & 0 \\ -1 & 1 & -5 & 0 & 0 \\ 2 & -3 & 13 & 1 & 0 \end{pmatrix}$$

as its augmented matrix. Give an explicit description of the solution set using these vector(s). In this case there will be only one vector in the basis.

44. Find a basis for the solution set to the system of equations having

$$\begin{pmatrix} 1 & 1 & -5 & 5 & -1 & 0 \\ 1 & 2 & -8 & 9 & -3 & 0 \\ 1 & 0 & -2 & 1 & 2 & 0 \\ 1 & 2 & -8 & 9 & -4 & 0 \end{pmatrix}$$

as its augmented matrix. Give an explicit description of the solution set using these vector(s).

45. You can find the row reduced echelon form for any matrix and it is uniquely determined. This is shown in Chapter 8. The rank will be defined later as the number of pivot columns, equivalently number of nonzero rows in the row reduced echelon form. Here is a matrix:

$$\begin{pmatrix} -1 & 1 & 4 & -4 & -2 \\ 2 & 3 & 3 & -13 & 0 \\ -2 & -2 & -1 & 9 & -1 \\ 0 & 1 & 3 & -5 & 0 \end{pmatrix}$$

Find the rank of this matrix by finding the row reduced echelon form and counting the number of pivot columns.

46. Here is a matrix:

$$\begin{pmatrix} -1 & 1 & -1 & -4 & -2 \\ 2 & 3 & 7 & -13 & 0 \\ -2 & -2 & -6 & 9 & -1 \\ 0 & 1 & 1 & -5 & 0 \end{pmatrix}$$

Find its row reduced echelon form and determine its rank.

47. Here is a matrix:

$$A = \begin{pmatrix} -1 & 1 & 1 & -4 & -2 \\ 2 & 3 & 13 & -13 & 0 \\ -2 & -2 & -10 & 9 & -1 \\ 0 & 1 & 3 & -5 & 0 \end{pmatrix}$$

Its row reduced echelon form is

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

The last column is 1 times the first column added to -5 times the second column added to -1 times the fourth column. In adding columns we just add the corresponding entries to get the result and in multiplying by a scalar, we simply multiply all entries by the scalar. Does the same hold for the columns of the original matrix A? Is the last column 1 times the first column added to (-5) times the second column added to (-1) times the fourth column? This equation is called a linear relation. Can you obtain a linear relation with the first three columns of the original matrix using this row reduced echelon form in the same way?

Chapter 5

Matrices

5.1 Matrix Arithmetic

5.1.1 Addition And Scalar Multiplication Of Matrices

You have now solved systems of equations by writing them in terms of an augmented matrix and then doing row operations on this augmented matrix. It turns out such rectangular arrays of numbers are important from many other different points of view. Numbers are also called **scalars**. In this book, numbers will generally be either real or complex numbers. I will refer to the set of numbers as $\mathbb{F}$ sometimes when it is not important to worry about whether the number is real or complex. Thus $\mathbb{F}$ can be either the real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$. However, most of the algebraic considerations hold for more general fields of scalars.

A **matrix** is a rectangular array of numbers. Several of them are referred to as **matrices**. For example, here is a matrix.

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

The size or dimension of a matrix is defined as $m \times n$ where m is the number of rows and n is the number of columns. The above matrix is a 3×4 matrix because there are three rows and four columns. The first row is $(1\ 2\ 3\ 4)$, the second row is $(5\ 2\ 8\ 7)$ and so forth. The first column is $\begin{pmatrix} 1 \\ 5 \\ 6 \end{pmatrix}$. When specifying the size of a matrix, you always list the number of rows before the number of columns. Also, you can remember the columns are like columns in a Greek temple. They stand upright while the rows just lie there like rows made by a tractor in a plowed field. Elements of the matrix are identified according to position in the matrix. For example, 8 is in position $2,3$ because it is in the second row and the third column. You might remember that you always list the rows before the columns by using the phrase **Row**man **C**atholic. The symbol, (a_{ij}) refers to a matrix. The entry in the i^{th} row and the j^{th} column of this matrix is denoted by a_{ij}. Using this notation on the above matrix, $a_{23} = 8, a_{32} = -9, a_{12} = 2$, etc.

There are various operations which are done on matrices. Matrices can be added multiplied by a scalar, and multiplied by other matrices. To illustrate scalar multiplication, consider the following example in which a matrix is being multiplied by the scalar 3.

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

The new matrix is obtained by multiplying every entry of the original matrix by the given scalar. If A is an $m \times n$ matrix, $-A$ is defined to equal $(-1)A$.

Two matrices must be the same size to be added. The sum of two matrices is a matrix which is obtained by adding the corresponding entries. Thus

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 2 \end{pmatrix} + \begin{pmatrix} -1 & 4 \\ 2 & 8 \\ 6 & -4 \end{pmatrix} = \begin{pmatrix} 0 & 6 \\ 5 & 12 \\ 11 & -2 \end{pmatrix}.$$

Two matrices are equal exactly when they are the same size and the corresponding entries 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 *(Scalar Multiplication) If $A = (a_{ij})$ and k is a scalar, then $kA = (ka_{ij})$.*

Example 5.1.2 $7\begin{pmatrix} 2 & 0 \\ 1 & -4 \end{pmatrix} = \begin{pmatrix} 14 & 0 \\ 7 & -28 \end{pmatrix}.$

Definition 5.1.3 *(Addition) If $A = (a_{ij})$ and $B = (b_{ij})$ are two $m \times n$ matrices. Then $A + B = C$ where*

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

for $c_{ij} = a_{ij} + b_{ij}$.

Example 5.1.4

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

To save on notation, we will often use A_{ij} to refer to the ij^{th} entry of the matrix A.

Definition 5.1.5 *(The zero matrix) The $m \times n$ zero matrix is the $m \times n$ matrix having every entry equal to zero. It is denoted by* 0.

Example 5.1.6 *The* 2×3 *zero matrix is* $\left(\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right)$.

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

Definition 5.1.7 *(Equality of matrices) Let A and B be two matrices. Then $A = B$ means that the two matrices are of the same size and for $A = (a_{ij})$ and $B = (b_{ij})$, $a_{ij} = b_{ij}$ for all $1 \leq i \leq m$ and $1 \leq j \leq n$.*

The following properties of matrices can be easily verified. You should do so. These properties are called the vector space axioms.

- Commutative Law Of Addition.

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

- Associative Law for Addition.

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

- Existence of an Additive Identity

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

- Existence of an Additive Inverse

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

Also for α, β scalars, the following additional properties hold.

- Distributive law over Matrix Addition.

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

- Distributive law over Scalar Addition

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

- Associative law for Scalar Multiplication

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

- Rule for Multiplication by 1.

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

As an example, consider the Commutative Law of Addition. Let $A+B=C$ and $B+A=D$. Why is $D=C$?

$$C_{ij}=A_{ij}+B_{ij}=B_{ij}+A_{ij}=D_{ij}.$$

Therefore, $C=D$ because the ij^{th} entries are the same. Note that the conclusion follows from the commutative law of addition of numbers.

5.1.2 Multiplication Of Matrices

This is where things get interesting. Matrices can be thought of as a rule for making new vectors from old vectors.

Definition 5.1.8 *Matrices which are $n\times 1$ or $1\times n$ are called* ***vectors*** *and are often denoted by a bold letter. Thus the $n\times 1$ matrix*

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

is also called a ***column vector.*** *The $1\times n$ matrix*

$$\begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}$$

is called a ***row vector.***

Although the following description of matrix multiplication may seem strange, it is in fact the most important and useful of the matrix operations. To begin with consider the case where a matrix is multiplied by a column vector. First consider a special case.

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

By definition, this equals

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

In more general terms,

$$\begin{aligned}\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} &= x_1\begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}+x_2\begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix}+x_3\begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix} \\ &= \begin{pmatrix} a_{11}x_1+a_{12}x_2+a_{13}x_3 \\ a_{21}x_1+a_{22}x_2+a_{23}x_3 \end{pmatrix}.\end{aligned}$$

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. The above sum is called a **linear combination** of the given column vectors. These will be discussed more later. In general, a linear combination of vectors is just a sum consisting of scalars times vectors. When you multiply a matrix on the left by a vector on the right, the numbers making up the vector are just the scalars to be used in the linear combination of the columns as illustrated above.

More generally, here is the definition of how to multiply an $(m\times n)$ matrix times a $(n\times 1)$ matrix (column vector).

Definition 5.1.9 *Let $A = A_{ij}$ be an $m \times n$ matrix and let $\mathbf{v}$ be an $n \times 1$ matrix,*

$$\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}, \, A = (\mathbf{a}_1, \cdots, \mathbf{a}_n)$$

where $\mathbf{a}_i$ is an $m \times 1$ column vector. Then $A\mathbf{v}$, written as

$$\begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix},$$

is the $m \times 1$ column vector which equals the following linear combination of the columns.

$$v_1\mathbf{a}_1 + v_2\mathbf{a}_2 + \cdots + v_n\mathbf{a}_n \equiv \sum_{j=1}^{n} v_j \mathbf{a}_j \tag{5.9}$$

If the j^{th} column of A is

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

then 5.9 takes the form

$$v_1 \begin{pmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{m1} \end{pmatrix} + v_2 \begin{pmatrix} A_{12} \\ A_{22} \\ \vdots \\ A_{m2} \end{pmatrix} + \cdots + v_n \begin{pmatrix} A_{1n} \\ A_{2n} \\ \vdots \\ A_{mn} \end{pmatrix}$$

Thus the i^{th} entry of $A\mathbf{v}$ is $\sum_{j=1}^{n} A_{ij} v_j$. Note that multiplication by an $m \times n$ matrix takes an $n \times 1$ matrix, and produces an $m \times 1$ matrix (vector).

Here is another example.

Example 5.1.10 *Compute*

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

First of all this is of the form $(3 \times 4)(4 \times 1)$ and so the result should be a (3×1). Note how the inside numbers cancel. To get the element in the second row and first and only column, compute

$$\begin{aligned} \sum_{k=1}^{4} a_{2k} v_k &= a_{21}v_1 + a_{22}v_2 + a_{23}v_3 + a_{24}v_4 \\ &= 0 \times 1 + 2 \times 2 + 1 \times 0 + (-2) \times 1 = 2. \end{aligned}$$

You should do the rest of the problem and verify

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

The next task is to multiply an $m \times n$ matrix times an $n \times p$ matrix. Before doing so, the following may be helpful.

For A and B matrices, in order to form the product, AB the number of columns of A must equal the number of rows of B. Thus the form of the product must be

$$(m \times n)(n \times p) = m \times p$$

Note the two outside numbers give the size of the product. Remember:

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

Definition 5.1.11 *When the number of columns of A equals the number of rows of B the two matrices are said to be* ***conformable*** *and the product AB is obtained as follows. Let A be an $m \times n$ matrix and let B be an $n \times p$ matrix. Then B is of the form*

$$B = (\mathbf{b}_1, \cdots, \mathbf{b}_p)$$

where $\mathbf{b}_k$ is an $n \times 1$ matrix or column vector. Then the $m \times p$ matrix AB is defined as follows:

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

where $A\mathbf{b}_k$ is an $m \times 1$ matrix or column vector which gives the k^{th} column of AB.

Example 5.1.12 *Multiply the following.*

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

The first thing you need to check before doing anything else is whether it is possible to do the multiplication. The first matrix on left is a 2 × 3 and the second matrix on right is a 3 × 3. Therefore, is it possible to multiply these matrices. According to the above discussion it should be a 2 × 3 matrix of the form

$$\left(\overbrace{\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}}^{\text{First column}}, \overbrace{\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}}^{\text{Second column}}, \overbrace{\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}}^{\text{Third column}} \right)$$

You know how to multiply a matrix times a vector and so you do so to obtain each of the three columns. Thus

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

Example 5.1.13 *Multiply the following.*

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

First check if it is possible. This is of the form $(3\times 3)(2\times 3)$. The inside numbers do not match and so you can't do this multiplication. This means that anything you write will be absolute nonsense because it is impossible to multiply these matrices in this order. Aren't they the same two matrices considered in the previous example? Yes they are. It is just that here they are in a different order. This shows something you must always remember about matrix multiplication.

Order Matters!

Matrix Multiplication Is Not Commutative!

This is very different than multiplication of numbers!

5.1.3 The ij^{th} Entry Of A Product

It is important to describe matrix multiplication in terms of entries of the matrices. What is the ij^{th} entry of AB? It would be the i^{th} entry of the j^{th} column of AB. Thus it would be the i^{th} entry of $A\mathbf{b}_j$. Now

$$\mathbf{b}_j = \begin{pmatrix} B_{1j} \\ \vdots \\ B_{nj} \end{pmatrix}$$

and from the above definition, the i^{th} entry is

$$\sum_{k=1}^{n} A_{ik}B_{kj} \equiv A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{in}B_{nj} \tag{5.11}$$

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 5.1.11.

Definition 5.1.14 *Let $A = (A_{ij})$ be an $m \times n$ matrix and let $B = (B_{ij})$ be an $n \times p$ matrix. Then AB is an $m \times p$ matrix and*

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj} \equiv A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{in}B_{nj}. \tag{5.12}$$

Another way to write this is

$$(AB)_{ij} = \overset{1\times n}{\begin{pmatrix} A_{i1} & A_{i2} & \cdots & A_{in} \end{pmatrix}} \overset{n\times 1}{\begin{pmatrix} B_{1j} \\ B_{2j} \\ \vdots \\ B_{nj} \end{pmatrix}}$$

Note that to get $(AB)_{ij}$ you multiply the i^{th} row of A and the j^{th} column of B. In terms of the dot product from calculus or earlier in this book, the ij^{th} entry of AB is the dot product of the i^{th} row of A with the j^{th} column of B.

I will summarize the above discussion in the following proposition which shows that the above definition delivers the earlier one about $AB = \begin{pmatrix} A\mathbf{b}_1 & \cdots & A\mathbf{b}_p \end{pmatrix}$. It is important to realize these two definitions are equivalent.

Proposition 5.1.15 *Let A be an $m \times n$ matrix. Let $B = \begin{pmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_p \end{pmatrix}$ where each $\mathbf{b}_k$ is a column vector or $n \times 1$ matrix so B is an $n \times p$ matrix. Then AB is an $m \times p$ matrix and*

$$AB = \begin{pmatrix} A\mathbf{b}_1 & \cdots & A\mathbf{b}_p \end{pmatrix}$$

so the k^{th} column of AB is just $A\mathbf{b}_k$.

Proof: From the definition of multiplication of matrices, $(AB)_{ik} = \sum_r A_{ir} B_{rk}$. However,

$$\mathbf{b}_k = \begin{pmatrix} B_{1k} \\ \vdots \\ B_{nk} \end{pmatrix}$$

and so, from the way we multiply a matrix times a vector,

$$(A\mathbf{b}_k)_i = \sum_r A_{ir} (\mathbf{b}_k)_r = \sum_r A_{ir} B_{rk}$$

Thus, the i^{th} entry from the top of $A\mathbf{b}_k$ is the i^{th} entry in the k^{th} column of AB showing that indeed the claim is true. ■

Example 5.1.16 *Multiply if possible* $\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \end{pmatrix}$.

First check to see if this is possible. It is of the form $(3 \times 2)(2 \times 3)$ and since the inside numbers match, the two matrices are conformable and it is possible to do the multiplication. The result should be a 3×3 matrix. The answer is of the form

$$\left(\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 \\ 7 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 3 \\ 6 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right)$$

where the commas separate the columns in the resulting product. Thus the above product equals

$$\begin{pmatrix} 16 & 15 & 5 \\ 13 & 15 & 5 \\ 46 & 42 & 14 \end{pmatrix},$$

a 3×3 matrix as desired. In terms of the ij^{th} entries and the above definition, the entry in the third row and second column of the product should equal

$$\sum_j a_{3k} b_{k2} = a_{31} b_{12} + a_{32} b_{22} = 2 \times 3 + 6 \times 6 = 42.$$

You should try a few more such examples to verify the above definition in terms of the ij^{th} entries works for other entries.

Example 5.1.17 *Multiply if possible* $\begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \\ 0 & 0 & 0 \end{pmatrix}$.

This is not possible because it is of the form $(3 \times 2)(3 \times 3)$ and the middle numbers don't match. In other words the two matrices are not conformable in the indicated order.

Example 5.1.18 *Multiply if possible* $\begin{pmatrix} 2 & 3 & 1 \\ 7 & 6 & 2 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}$.

This is possible because in this case it is of the form $(3 \times 3)(3 \times 2)$ and the middle numbers do match so the matrices are conformable. When the multiplication is done it equals

$$\begin{pmatrix} 13 & 13 \\ 29 & 32 \\ 0 & 0 \end{pmatrix}.$$

Check this and be sure you come up with the same answer.

Example 5.1.19 *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

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

5.1.4 Properties Of Matrix Multiplication

As pointed out above, sometimes it is possible to multiply matrices in one order but not in the other order. What if it makes sense to multiply them in either order? Will the two products be equal then?

Example 5.1.20 *Compare* $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ *and* $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$.

The first product is

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

The second product is

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

You see these are not equal. Again you cannot conclude that $AB = BA$ for matrix multiplication even when multiplication is defined in both orders. However, there are some properties which do hold.

Proposition 5.1.21 *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 Definition 5.1.14,

$$(A\,(aB+bC))_{ij} = \sum_k A_{ik}\,(aB+bC)_{kj} = \sum_k A_{ik}\left(aB_{kj} + bC_{kj}\right)$$

$$= a\sum_k A_{ik}B_{kj} + b\sum_k A_{ik}C_{kj} = a\,(AB)_{ij} + b\,(AC)_{ij}$$

$$= (a\,(AB) + b\,(AC))_{ij}\,.$$

Thus $A\,(B+C) = AB + AC$ as claimed. Formula 5.14 is entirely similar.

Formula 5.15 is the associative law of multiplication. Using Definition 5.1.14,

$$(A\,(BC))_{ij} = \sum_k A_{ik}\,(BC)_{kj} = \sum_k A_{ik} \sum_l B_{kl}C_{lj}$$

$$= \sum_l (AB)_{il}\,C_{lj} = ((AB)\,C)_{ij}\,.$$

This proves 5.15. ■

5.1.5 The Transpose

Another important operation on matrices is that of taking the **transpose**. The following example shows what is meant by this operation, denoted by placing a T as an exponent on the matrix.

$$\begin{pmatrix} 1 & 4 \\ 3 & 1 \\ 2 & 6 \end{pmatrix}^T = \begin{pmatrix} 1 & 3 & 2 \\ 4 & 1 & 6 \end{pmatrix}$$

What happened? The first column became the first row and the second column became the second row. Thus the 3×2 matrix became a 2×3 matrix. The number 3 was in the second row and the first column and it ended up in the first row and second column. Here is the definition.

Definition 5.1.22 *Let A be an $m \times n$ matrix. Then A^T denotes the $n \times m$ matrix which is defined as follows.*

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

Example 5.1.23

$$\left(\begin{array}{ccc} 1 & 2 & -6 \\ 3 & 5 & 4 \end{array}\right)^T = \left(\begin{array}{cc} 1 & 3 \\ 2 & 5 \\ -6 & 4 \end{array}\right).$$

The transpose of a matrix has the following important properties.

Lemma 5.1.24 *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,

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

The proof of Formula 5.17 is left as an exercise. ■

Definition 5.1.25 *An $n \times n$ matrix A is said to be* ***symmetric*** *if $A = A^T$. It is said to be* ***skew symmetric*** *if $A = -A^T$.*

Example 5.1.26 *Let*

$$A = \left(\begin{array}{ccc} 2 & 1 & 3 \\ 1 & 5 & -3 \\ 3 & -3 & 7 \end{array}\right).$$

Then A is symmetric.

Example 5.1.27 *Let*

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

Then A is skew symmetric.

5.1.6 The Identity And Inverses

There is a special matrix called I and referred to as the identity matrix. It is always a square matrix, meaning the number of rows equals the number of columns and it has the property that there are ones down the main diagonal and zeroes elsewhere. Here are some identity matrices of various sizes.

$$(1), \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right), \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right), \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right).$$

The first is the 1×1 identity matrix, the second is the 2×2 identity matrix, the third is the 3×3 identity matrix, and the fourth is the 4×4 identity matrix. By extension, you can likely see what the $n\times n$ identity matrix would be. It is so important that there is a special symbol to denote the ij^{th} entry of the identity matrix $I_{ij}=\delta_{ij}$ where δ_{ij} is the **Kronecker 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 5.1.28 *Suppose A is an $m\times n$ matrix and I_n is the $n\times n$ identity matrix. Then $AI_n=A$. If I_m is the $m\times m$ identity matrix, it also follows that $I_mA=A$.*

Proof:

$$(AI_n)_{ij}=\sum_k A_{ik}\delta_{kj}=A_{ij}$$

and so $AI_n=A$. The other case is left as an exercise for you. ■

Definition 5.1.29 *An $n\times n$ matrix A has an **inverse**, A^{-1} if and only if $AA^{-1}=A^{-1}A=I$. Such a matrix is called **invertible**.*

It is very important to observe that the inverse of a matrix, if it exists, is unique. Another way to think of this is that if it acts like the inverse, then it is the inverse.

Theorem 5.1.30 *Suppose A^{-1} exists and $AB=BA=I$. Then $B=A^{-1}$.*

Proof:

$$A^{-1}=A^{-1}I=A^{-1}(AB)=\left(A^{-1}A\right)B=IB=B. \blacksquare$$

Unlike ordinary multiplication of numbers, it can happen that $A\neq 0$ but A may fail to have an inverse. This is illustrated in the following example.

Example 5.1.31 *Let $A=\begin{pmatrix}1&1\\1&1\end{pmatrix}$. Does A have an inverse?*

One might think A would have an inverse because it does not equal zero. However,

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

and if A^{-1} existed, this could not happen because you could write

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

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

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

Example 5.1.32 *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.

5.1.7 Finding The Inverse Of A Matrix

In the last example, how would you find A^{-1}? You wish to find a matrix $\begin{pmatrix} x & z \\ y & w \end{pmatrix}$ such that

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x & z \\ y & w \end{pmatrix}=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This requires the solution of the systems of equations,

$$x+y=1, x+2y=0$$

and

$$z+w=0, z+2w=1.$$

Writing the augmented matrix for these two systems gives

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

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

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

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

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

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

$$\left(\begin{array}{cc} 2 & -1 \\ -1 & 1 \end{array}\right).$$

Didn't the above seem rather repetitive? Note that exactly the same row operations were used in both systems. In each case, the end result was something of the form $(I|\mathbf{v})$ where I is the identity and $\mathbf{v}$ gave a column of the inverse. In the above, $\left(\begin{array}{c} x \\ y \end{array}\right)$, the first column of the inverse was obtained first and then the second column $\left(\begin{array}{c} z \\ w \end{array}\right)$.

To simplify this procedure, you could have written

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

and row reduced till you obtained

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

and read off the inverse as the 2 × 2 matrix on the right side.

This is the reason for the following simple procedure for finding the inverse of a matrix. This procedure is called the **Gauss-Jordan procedure**.

Procedure 5.1.33 *Suppose A is an $n \times n$ matrix. To find A^{-1} if it exists, form the augmented $n \times 2n$ matrix*

$$(A|I)$$

and then, if possible do row operations until you obtain an $n \times 2n$ matrix of the form

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

When this has been done, $B = A^{-1}$. If it is impossible to row reduce to a matrix of the form $(I|B)$, then A has no inverse.

Actually, all this shows is how to find a right inverse if it exists. What has been shown from the above discussion is that $AB = I$. Later, I will show that this right inverse is **the** inverse. See Corollary 7.1.15 or Theorem 8.2.11 presented later. However, it is not hard to see that this should be the case as follows.

The row operations are all reversible. If the row operation involves switching two rows, the reverse row operation involves switching them again to get back to where you started. If the row operation involves multiplying a row by $a \neq 0$, then you would get back to where you began by multiplying the row by $1/a$. The third row operation involving addition of c times row i to row j can be reversed by adding $-c$ times row i to row j.

In the above procedure, a sequence of row operations applied to I yields B while the same sequence of operations applied to A yields I. Therefore, the sequence of reverse row operations in the opposite order applied to B will yield I and applied to I will yield A. That is, there are row operations which provide

$$(B|I) \to (I|A)$$

and as just explained, A must be a right inverse for B. Therefore, $BA = I$. Hence B is both a right and a left inverse for A because $AB = BA = I$.

If it is impossible to row reduce $(A|I)$ to get $(I|B)$, then in particular, it is impossible to row reduce A to I and consequently impossible to do a sequence of row operations to I and get A. The only way this can happen is that it is possible to row reduce A to a matrix of the form $\begin{pmatrix} C \\ \mathbf{0} \end{pmatrix}$ where $\mathbf{0}$ is a row of zeros. Then there will be no solution to the system of equations represented by the augmented matrix

$$\begin{pmatrix} C & | & \mathbf{0} \\ \mathbf{0} & | & 1 \end{pmatrix}$$

Using the reverse row operations in the opposite order on both matrices in the above, it follows that there must exist $\mathbf{a}$ such that there is no solution to the system of equations represented by $(A|\mathbf{a})$. Hence A fails to have an inverse, because if it did, then there would be a solution $\mathbf{x}$ to the equation $A\mathbf{x} = \mathbf{a}$ given by $A^{-1}\mathbf{a}$.

Example 5.1.34 *Let* $A = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 3 & 1 & -1 \end{pmatrix}$. *Find* A^{-1} *if it exists.*

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

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

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

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

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

$$\left(\begin{array}{ccccccc} 1 & 2 & 2 & | & 1 & 0 & 0 \\ 0 & -10 & 0 & | & -5 & 5 & 0 \\ 0 & 0 & 14 & | & 1 & 5 & -2 \end{array}\right)$$

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

$$\left(\begin{array}{ccccccc} -7 & -14 & 0 & | & -6 & 5 & -2 \\ 0 & -10 & 0 & | & -5 & 5 & 0 \\ 0 & 0 & 14 & | & 1 & 5 & -2 \end{array}\right).$$

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

$$\left(\begin{array}{ccccccc} -7 & 0 & 0 & | & 1 & -2 & -2 \\ 0 & -10 & 0 & | & -5 & 5 & 0 \\ 0 & 0 & 14 & | & 1 & 5 & -2 \end{array}\right).$$

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

$$\left(\begin{array}{ccccccc} 1 & 0 & 0 & | & -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ 0 & 1 & 0 & | & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & | & \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{array}\right).$$

Therefore, the inverse is

$$\left(\begin{array}{ccc} -\frac{1}{7} & \frac{2}{7} & \frac{2}{7} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{14} & \frac{5}{14} & -\frac{1}{7} \end{array}\right)$$

Example 5.1.35 *Let* $A = \left(\begin{array}{ccc} 1 & 2 & 2 \\ 1 & 0 & 2 \\ 2 & 2 & 4 \end{array}\right)$. *Find* A^{-1} *if it exists.*

Write the augmented matrix $(A|I)$

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

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

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

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

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

At this point, you can see there will be no inverse because you have obtained a row of zeros in the left half of the augmented matrix $(A|I)$. Thus there will be no way to obtain I on the left.

Example 5.1.36 *Let* $A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}$. *Find* A^{-1} *if it exists.*

►►

Form the augmented matrix

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

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

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

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

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

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

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

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

Example 5.1.37 *In this example, it is shown how to use the inverse of a matrix to find the solution to a system of equations. Consider the following system of equations. Use the inverse of a suitable matrix to give the solutions to this system.*

$$\begin{pmatrix} x+z=1 \\ x-y+z=3 \\ x+y-z=2 \end{pmatrix}.$$

The system of equations can be written in terms of matrices as

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}. \tag{5.21}$$

More simply, this is of the form $A\mathbf{x}=\mathbf{b}$. Suppose you find the inverse of the matrix A^{-1}. Then you could multiply both sides of this equation by A^{-1} to obtain

$$\mathbf{x} = \left(A^{-1}A\right)\mathbf{x} = A^{-1}\left(A\mathbf{x}\right) = A^{-1}\mathbf{b}.$$

This gives the solution as $\mathbf{x}=A^{-1}\mathbf{b}$. Note that once you have found the inverse, you can easily get the solution for different right hand sides without any effort. It is always just $A^{-1}\mathbf{b}$. In the given example, the inverse of the matrix is

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

This was shown in Example 5.1.36. Therefore, from what was just explained, the solution to the given system is

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & -1 & 0 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} = \begin{pmatrix} \frac{5}{2} \\ -2 \\ -\frac{3}{2} \end{pmatrix}.$$

What if the right side of 5.21 had been $\begin{pmatrix} 0 & 1 & 3 \end{pmatrix}^T$? What would be the solution to

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix}?$$

By the above discussion, it is just

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

This illustrates why once you have found the inverse of a given matrix, you can use it to solve many different systems easily.

5.2 MATLAB And Matrix Arithmetic

To find the inverse of a square matrix in matlab, you open it and type the following. The >> will already be there.

>>inv([1,2,3;5,2,7;8,2,1]) Then press enter and it will give the following:

```
ans =
-0.1667  0.0556   0.1111
0.7083  -0.3194   0.1111
-0.0833  0.1944 -0.1111
```

Note how it computed the inverse in decimals. If you want the answer in terms of fractions, you do the following:

>>inv(sym([1,2,3;5,2,7;8,2,1])) Then press enter and it will give the following:

```
ans =
[ -1/6, 1/18, 1/9]
[ 17/24, -23/72, 1/9]
[ -1/12, 7/36, -1/9]
```

You can do other things as well. Say you have

```
>>A=[1,2,3;5,2,7;8,2,1];B=[3,2,-5;3,11,2;-3,-1,5];
C=[1,2;4,-3;7,3];D=[1,2,3;-3,2,1];
```

This defines some matrices. Then suppose you wanted to find $\left(A^{-1}D^T+BC\right)^T$. You would then type

```
transpose(inv(sym(A))*transpose(D)+B*C) or (inv(sym(A))*D'+B*C)'
```

and press enter. This gives

```
ans =
[ -427/18, 4421/72, 1007/36]
[ -257/18, -1703/72, 451/36]
```

In matlab, A' means $\bar{A}^T$ the conjugate transpose of A. Since everything is real here, this reduces to the transpose.

To get to a new line in matlab, you need to press shift enter. You can adapt this to other situations. Notice how a ; was placed after the definition of A,B,C,D. This tells matlab that you have defined something but not to say anything about it. If you don't do this, then when you press return, it will list the matrices and you don't want to see that. You just want the answer. When you have done a computation in matlab, you ought to go to >> and type "clear all" and then enter. That way, you can use the symbols again with different definition. If you don't do the "clear all" thing, it will go on thinking that A is what you defined earlier.

5.3 Exercises

1. Here are some matrices:

$$\begin{aligned} A &= \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 7 \end{pmatrix}, B = \begin{pmatrix} 3 & -1 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\ C &= \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}, D = \begin{pmatrix} -1 & 2 \\ 2 & -3 \end{pmatrix}, E = \begin{pmatrix} 2 \\ 3 \end{pmatrix}. \end{aligned}$$

Find if possible $-3A, 3B-A, AC, CB, AE, EA$. If it is not possible explain why.

2. Here are some matrices:

$$\begin{aligned} A &= \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\ C &= \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 & 1 \\ 4 & -3 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}. \end{aligned}$$

   Find if possible $-3A, 3B - A, AC, CA, AE, EA, BE, DE$. If it is not possible explain why.

3. Here are some matrices:

$$\begin{aligned} A &= \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\ C &= \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 & 1 \\ 4 & -3 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}. \end{aligned}$$

   Find if possible $-3A^T, 3B - A^T, AC, CA, AE, E^T B, BE, DE, EE^T, E^T E$. If it is not possible explain why.

4. Here are some matrices:

$$\begin{aligned} A &= \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & -1 \end{pmatrix}, B = \begin{pmatrix} 2 & -5 & 2 \\ -3 & 2 & 1 \end{pmatrix}, \\ C &= \begin{pmatrix} 1 & 2 \\ 5 & 0 \end{pmatrix}, D = \begin{pmatrix} -1 \\ 4 \end{pmatrix}, E = \begin{pmatrix} 1 \\ 3 \end{pmatrix}. \end{aligned}$$

   Find the following if possible and explain why it is not possible if this is the case.

$$AD, DA, D^T B, D^T BE, E^T D, DE^T.$$

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

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

6. 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) $(A+B)(A-B) = A^2 - B^2$

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

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

9. Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 1 & 2 \\ 3 & k \end{pmatrix}$. Is it possible to choose k such that $AB = BA$? If so, what should k equal?

10. Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 1 & 2 \\ 1 & k \end{pmatrix}$. Is it possible to choose k such that $AB = BA$? If so, what should k equal?

11. In 5.1 - 5.8 describe $-A$ and 0.

12. Let A be an $n\times n$ matrix. Show A equals the sum of a symmetric and a skew symmetric matrix. (M is skew symmetric if $M = -M^T$. M is symmetric if $M^T = M$.) **Hint:** Show that $\frac{1}{2}\left(A^T + A\right)$ is symmetric and then consider using this as one of the matrices.

13. Show every skew symmetric matrix has all zeros down the main diagonal. The main diagonal consists of every entry of the matrix which is of the form a_{ii}. It runs from the upper left down to the lower right.

14. Suppose M is a 3×3 skew symmetric matrix. Show there exists a vector $\blacksquare$ such that for all $\mathbf{u} \in \mathbb{R}^3$
$$M\mathbf{u} = \blacksquare \times \mathbf{u}$$
Hint: Explain why, since M is skew symmetric it is of the form
$$M = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}$$
where the ω_i are numbers. Then consider $\omega_1\mathbf{i} + \omega_2\mathbf{j} + \omega_3\mathbf{k}$.

15. Using only the properties 5.1 - 5.8 show $-A$ is unique.

16. Using only the properties 5.1 - 5.8 show 0 is unique.

17. 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.

18. Using only the properties 5.1 - 5.8 and previous problems show $(-1)A = -A$.

19. Prove 5.17.

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

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

22. Suppose $AB = AC$ and A is an invertible $n \times n$ matrix. Does it follow that $B = C$? Explain why or why not. What if A were a non invertible $n \times n$ matrix?

23. Find your own examples:

    (a) 2×2 matrices, A and B such that $A \neq 0, B \neq 0$ with $AB \neq BA$.

    (b) 2×2 matrices, A and B such that $A \neq 0, B \neq 0$, but $AB = 0$.

    (c) 2×2 matrices, A, D, and C such that $A \neq 0, C \neq D$, but $AC = AD$.

24. Explain why if $AB = AC$ and A^{-1} exists, then $B = C$.

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

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

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

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

29. Let
$$A = \begin{pmatrix} 0 & 1 \\ 5 & 3 \end{pmatrix}.$$
Find A^{-1} if possible. If A^{-1} does not exist, determine why.

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

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

32. Let A be a 2×2 matrix which has an inverse. Say $A=\left(\begin{array}{cc} a & b \\ c & d \end{array}\right)$. Find a formula for A^{-1} in terms of a,b,c,d.

33. Let
$$A=\left(\begin{array}{ccc} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 1 & 0 & 2 \end{array}\right).$$
Find A^{-1} if possible. If A^{-1} does not exist, determine why.

34. Let
$$A=\left(\begin{array}{ccc} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{array}\right).$$
Find A^{-1} if possible. If A^{-1} does not exist, determine why.

35. Let
$$A=\left(\begin{array}{ccc} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 4 & 5 & 10 \end{array}\right).$$
Find A^{-1} if possible. If A^{-1} does not exist, determine why.

36. Let
$$A=\left(\begin{array}{cccc} 1 & 2 & 0 & 2 \\ 1 & 1 & 2 & 0 \\ 2 & 1 & -3 & 2 \\ 1 & 2 & 1 & 2 \end{array}\right)$$
Find A^{-1} if possible. If A^{-1} does not exist, determine why.

37. Write $\left(\begin{array}{c} x_1-x_2+2x_3 \\ 2x_3+x_1 \\ 3x_3 \\ 3x_4+3x_2+x_1 \end{array}\right)$ in the form $A\left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array}\right)$ where A is an appropriate matrix.

38. Write $\left(\begin{array}{c} x_1+3x_2+2x_3 \\ 2x_3+x_1 \\ 6x_3 \\ x_4+3x_2+x_1 \end{array}\right)$ in the form $A\left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array}\right)$ where A is an appropriate matrix.

39. Write $\left(\begin{array}{c} x_1+x_2+x_3 \\ 2x_3+x_1+x_2 \\ x_3-x_1 \\ 3x_4+x_1 \end{array}\right)$ in the form $A\left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array}\right)$ where A is an appropriate matrix.

40. Using the inverse of the matrix, find the solution to the systems

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$$
$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \\ -2 \end{pmatrix}.$$

Now give the solution in terms of a, b, and c to

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

41. Using the inverse of the matrix, find the solution to the systems

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$$
$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \\ -2 \end{pmatrix}.$$

Now give the solution in terms of a, b, and c to

$$\begin{pmatrix} 1 & 0 & 3 \\ 2 & 3 & 4 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

42. Using the inverse of the matrix, find the solution to the system

$$\begin{pmatrix} -1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 3 & \frac{1}{2} & -\frac{1}{2} & -\frac{5}{2} \\ -1 & 0 & 0 & 1 \\ -2 & -\frac{3}{4} & \frac{1}{4} & \frac{9}{4} \end{pmatrix}\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}.$$

43. 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 $\mathbf{x} = A^{-1}\mathbf{b}$.

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

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

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

47. Show $(AB)^{-1} = B^{-1}A^{-1}$ by verifying that $AB\left(B^{-1}A^{-1}\right) = I$ and
$$B^{-1}A^{-1}(AB) = I.$$
Hint: Use Problem 45.

48. Show that $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ by verifying that
$$(ABC)\left(C^{-1}B^{-1}A^{-1}\right) = I$$
and $\left(C^{-1}B^{-1}A^{-1}\right)(ABC) = I$. **Hint:** Use Problem 45.

49. If A is invertible, show $\left(A^2\right)^{-1} = \left(A^{-1}\right)^2$. **Hint:** Use Problem 45.

50. If A is invertible, show $\left(A^{-1}\right)^{-1} = A$. **Hint:** Use Problem 45.

51. Let A and be a real $m \times n$ matrix and let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Show $(A\mathbf{x}, \mathbf{y})_{\mathbb{R}^m} = \left(\mathbf{x}, A^T\mathbf{y}\right)_{\mathbb{R}^n}$ where $(\cdot,\cdot)_{\mathbb{R}^k}$ denotes the dot product in $\mathbb{R}^k$. In the notation above,
$$A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A^T\mathbf{y}.$$
Use the definition of matrix multiplication to do this. Recall $\mathbf{x} \cdot \mathbf{y} \equiv \sum_j x_j y_j$.

52. Use the result of Problem 51 to verify directly that $(AB)^T = B^T A^T$ without making any reference to subscripts.

53. Suppose A is an $n \times n$ matrix and for each j,
$$\sum_{i=1}^{n} \left|A_{ij}\right| < 1$$
Show that the infinite series $\sum_{k=0}^{\infty} A^k$ converges in the sense that the ij^{th} entry of the partial sums converge for each ij. **Hint:** Let $R \equiv \max_j \sum_{i=1}^{n} \left|A_{ij}\right|$. Thus $R < 1$. Show that $\left|\left(A^2\right)_{ij}\right| \leq R^2$. Then generalize to show that $\left|\left(A^m\right)_{ij}\right| \leq R^m$. Use this to show that the ij^{th} entry of the partial sums is a Cauchy sequence. From calculus, these converge by completeness of the real or complex numbers. Next show that $(I-A)^{-1} = \sum_{k=0}^{\infty} A^k$. The Leontief model in economics involves solving an equation for $\mathbf{x}$ of the form
$$\mathbf{x} = A\mathbf{x} + \mathbf{b}, \text{ or } (I-A)\mathbf{x} = \mathbf{b}$$
The vector $A\mathbf{x}$ is called the intermediate demand and the vectors $A^k\mathbf{x}$ have economic meaning. From the above,
$$\mathbf{x} = I\mathbf{b} + A\mathbf{b} + A^2\mathbf{b} + \cdots$$
The series is also called the Neuman series. It is important in functional analysis.

54. An elementary matrix is one which results from doing a row operation to the identity matrix. Thus the elementary matrix E which results from adding a times the i^{th} row to the j^{th} row would have $a\delta_{ik} + \delta_{jk}$ as the jk^{th} entry and all other rows would be unchanged. That is δ_{rs} provided $r \neq j$. Show that multiplying this matrix on the left of an appropriate sized matrix A results in doing the row operation to the matrix A. You might also want to verify that the other elementary matrices have the same effect, doing the row operation which resulted in the elementary matrix to A.

Chapter 6

Determinants

6.1 Basic Techniques And Properties

6.1.1 Cofactors And 2×2 Determinants

Let A be an $n \times n$ matrix. The **determinant** of A, denoted as $\det(A)$ is a number. If the matrix is a 2×2 matrix, this number is very easy to find.

Definition 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?

Definition 6.1.3 *Suppose A is a* 3×3 *matrix. The* ij^{th} ***minor,*** *denoted as* $minor(A)_{ij}$, *is the determinant of the* 2×2 *matrix which results from deleting the* i^{th} *row and the* j^{th} *column.*

Example 6.1.4 *Consider the matrix*

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

The $(1,2)$ *minor is the determinant of the* 2×2 *matrix which results when you delete the first row and the second column. This minor is therefore*

$$\det \begin{pmatrix} 4 & 2 \\ 3 & 1 \end{pmatrix} = -2.$$

The $(2,3)$ minor is the determinant of the 2×2 matrix which results when you delete the second row and the third column. This minor is therefore

$$\det\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = -4.$$

Definition 6.1.5 *Suppose A is a 3×3 matrix. The ij^{th}* ***cofactor*** *is defined to be $(-1)^{i+j}\times\left(ij^{th}\text{ minor}\right)$. In words, you multiply $(-1)^{i+j}$ times the ij^{th} minor to get the ij^{th} cofactor. The cofactors of a matrix are so important that special notation is appropriate when referring to them. The ij^{th} cofactor of a matrix A will be denoted by $\operatorname{cof}(A)_{ij}$. It is also convenient to refer to the cofactor of an entry of a matrix as follows. For a_{ij} an entry of the matrix, its cofactor is just $\operatorname{cof}(A)_{ij}$. Thus the cofactor of the ij^{th} entry is just the ij^{th} cofactor.*

Example 6.1.6 *Consider the matrix*

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

The $(1,2)$ minor is the determinant of the 2×2 matrix which results when you delete the first row and the second column. This minor is therefore

$$\det\begin{pmatrix} 4 & 2 \\ 3 & 1 \end{pmatrix} = -2.$$

It follows

$$\operatorname{cof}(A)_{12} = (-1)^{1+2}\det\begin{pmatrix} 4 & 2 \\ 3 & 1 \end{pmatrix} = (-1)^{1+2}(-2) = 2$$

The $(2,3)$ minor is the determinant of the 2×2 matrix which results when you delete the second row and the third column. This minor is therefore

$$\det\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = -4.$$

Therefore,

$$\operatorname{cof}(A)_{23} = (-1)^{2+3}\det\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix} = (-1)^{2+3}(-4) = 4.$$

Similarly,

$$\operatorname{cof}(A)_{22} = (-1)^{2+2}\det\begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix} = -8.$$

Definition 6.1.7 *The determinant of a 3×3 matrix A, is obtained by picking a row (column) and taking the product of each entry in that row (column) with its cofactor and adding these up. This process when applied to the i^{th} row (column) is known as expanding the determinant along the i^{th} row (column).*

Example 6.1.8 *Find the determinant of*

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

Here is how it is done by "**expanding along the first column**".

$$\overbrace{1(-1)^{1+1}\det\begin{pmatrix} 3 & 2 \\ 2 & 1 \end{pmatrix}}^{\operatorname{cof}(A)_{11}} + \overbrace{4(-1)^{2+1}\det\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}}^{\operatorname{cof}(A)_{21}} + \overbrace{3(-1)^{3+1}\det\begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}}^{\operatorname{cof}(A)_{31}} = 0.$$

You see, we just followed the rule in the above definition. We took the 1 in the first column and multiplied it by its cofactor, the 4 in the first column and multiplied it by its cofactor, and the 3 in the first column and multiplied it by its cofactor. Then we added these numbers together.

You could also expand the determinant along the second row as follows.

$$\overbrace{4(-1)^{2+1}\det\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}}^{\operatorname{cof}(A)_{21}} + \overbrace{3(-1)^{2+2}\det\begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}}^{\operatorname{cof}(A)_{22}} + \overbrace{2(-1)^{2+3}\det\begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}}^{\operatorname{cof}(A)_{23}} = 0.$$

Observe this gives the same number. You should try expanding along other rows and columns. If you don't make any mistakes, you will always get the same answer.

What about a 4×4 matrix? You know now how to find the determinant of a 3×3 matrix. The pattern is the same.

Definition 6.1.9 *Suppose A is a 4×4 matrix. The ij^{th} **minor** is the determinant of the 3×3 matrix you obtain when you delete the i^{th} row and the j^{th} column. The ij^{th} **cofactor,** $\operatorname{cof}(A)_{ij}$ is defined to be $(-1)^{i+j} \times (ij^{th}\ minor)$. In words, you multiply $(-1)^{i+j}$ times the ij^{th} minor to get the ij^{th} cofactor.*

Definition 6.1.10 *The determinant of a 4×4 matrix A, is obtained by picking a row (column) and taking the product of each entry in that row (column) with its cofactor and adding these together. This process when applied to the i^{th} row (column) is known as expanding the determinant along the i^{th} row (column).*

Example 6.1.11 *Find* $\det(A)$ *where*

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

As in the case of a 3×3 matrix, you can expand this along any row or column. Lets pick the third column. $\det(A) =$

$$3(-1)^{1+3}\det\begin{pmatrix} 5 & 4 & 3 \\ 1 & 3 & 5 \\ 3 & 4 & 2 \end{pmatrix} + 2(-1)^{2+3}\det\begin{pmatrix} 1 & 2 & 4 \\ 1 & 3 & 5 \\ 3 & 4 & 2 \end{pmatrix}$$

$$+4(-1)^{3+3}\det\begin{pmatrix} 1 & 2 & 4 \\ 5 & 4 & 3 \\ 3 & 4 & 2 \end{pmatrix} + 3(-1)^{4+3}\det\begin{pmatrix} 1 & 2 & 4 \\ 5 & 4 & 3 \\ 1 & 3 & 5 \end{pmatrix}.$$

Now you know how to expand each of these 3×3 matrices along a row or a column. If you do so, you will get -12 assuming you make no mistakes. You could expand this matrix along any row or any column and assuming you make no mistakes, you will always get the same thing which is defined to be the determinant of the matrix A. This method of evaluating a determinant by expanding along a row or a column is called the **method of Laplace expansion**.

Note that each of the four terms above involves three terms consisting of determinants of 2×2 matrices and each of these will need 2 terms. Therefore, there will be $4 \times 3 \times 2 = 24$ terms to evaluate in order to find the determinant using the method of Laplace expansion. Suppose now you have a 10×10 matrix and you follow the above pattern for evaluating determinants. By analogy to the above, there will be $10! = 3,628,800$ terms involved in the evaluation of such a determinant by Laplace expansion along a row or column. This is a lot of terms.

In addition to the difficulties just discussed, you should regard the above claim that you always get the same answer by picking any row or column with considerable skepticism. It is incredible and not at all obvious. However, it requires a little effort to establish it. This is done in the section on the theory of the determinant.

Definition 6.1.12 *Let $A = (a_{ij})$ be an $n \times n$ matrix and suppose the determinant of a $(n-1) \times (n-1)$ matrix has been defined. Then a new matrix called the* ***cofactor matrix****, $\operatorname{cof}(A)$ is defined by $\operatorname{cof}(A) = (c_{ij})$ where to obtain c_{ij} delete the i^{th} row and the j^{th} column of A, take the determinant of the $(n-1) \times (n-1)$ matrix which results, (This is called the ij^{th}* ***minor*** *of A.) and then multiply this number by $(-1)^{i+j}$. Thus $(-1)^{i+j} \times$ (the ij^{th} minor) equals the ij^{th} cofactor. To make the formulas easier to remember, $\operatorname{cof}(A)_{ij}$ will denote the ij^{th} entry of the cofactor matrix.*

With this definition of the cofactor matrix, here is how to define the determinant of an $n \times n$ matrix.

Definition 6.1.13 *Let A be an $n \times n$ matrix where $n \geq 2$ and suppose the determinant of an $(n-1) \times (n-1)$ has been defined. Then*

$$\det(A) = \sum_{j=1}^{n} a_{ij} \operatorname{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij} \operatorname{cof}(A)_{ij}. \tag{6.1}$$

The first formula consists of expanding the determinant along the i^{th} row and the second expands the determinant along the j^{th} column.

Theorem 6.1.14 *Expanding the $n \times n$ matrix along any row or column always gives the same answer so the above definition is a good definition.*

6.1.2 The Determinant Of A Triangular Matrix

Notwithstanding the difficulties involved in using the method of Laplace expansion, certain types of matrices are very easy to deal with.

Definition 6.1.15 *A matrix M, is upper triangular if $M_{ij} = 0$ whenever $i > j$. Thus such a matrix equals zero below the main diagonal, the entries of the form M_{ii}, as shown.*

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

A lower triangular matrix is defined similarly as a matrix for which all entries above the main diagonal are equal to zero.

You should verify the following using the above theorem on Laplace expansion.

Corollary 6.1.16 *Let M be an upper (lower) triangular matrix. Then $\det(M)$ is obtained by taking the product of the entries on the main diagonal.*

Example 6.1.17 *Let*

$$A = \begin{pmatrix} 1 & 2 & 3 & 77 \\ 0 & 2 & 6 & 7 \\ 0 & 0 & 3 & 33.7 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

Find $\det(A)$.

From the above corollary, it suffices to take the product of the diagonal elements. Thus $\det(A) = 1 \times 2 \times 3 \times (-1) = -6$. Without using the corollary, you could expand along the first column. This gives

$$1\det\begin{pmatrix} 2 & 6 & 7 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{pmatrix} + 0(-1)^{2+1}\det\begin{pmatrix} 2 & 3 & 77 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{pmatrix}$$
$$+0(-1)^{3+1}\det\begin{pmatrix} 2 & 3 & 77 \\ 2 & 6 & 7 \\ 0 & 0 & -1 \end{pmatrix} + 0(-1)^{4+1}\det\begin{pmatrix} 2 & 3 & 77 \\ 2 & 6 & 7 \\ 0 & 3 & 33.7 \end{pmatrix}$$

and the only nonzero term in the expansion is

$$1\det\begin{pmatrix} 2 & 6 & 7 \\ 0 & 3 & 33.7 \\ 0 & 0 & -1 \end{pmatrix}.$$

Now expand this along the first column to obtain

$$\left(\begin{array}{c} 2\times\det\left(\begin{array}{cc} 3 & 33.7 \\ 0 & -1 \end{array}\right)+0\left(-1\right)^{2+1}\det\left(\begin{array}{cc} 6 & 7 \\ 0 & -1 \end{array}\right) \\ +0\left(-1\right)^{3+1}\det\left(\begin{array}{cc} 6 & 7 \\ 3 & 33.7 \end{array}\right) \end{array}\right)=1\times 2\times\det\left(\begin{array}{cc} 3 & 33.7 \\ 0 & -1 \end{array}\right)$$

Next expand this last determinant along the first column to obtain the above equals

$$1\times 2\times 3\times(-1)=-6$$

which is just the product of the entries down the main diagonal of the original matrix. It works this way in general.

6.1.3 Properties Of Determinants

There are many properties satisfied by determinants. Some of these properties have to do with row operations. Recall the row operations.

Definition 6.1.18 *The row operations consist of the following*

1. *Switch two rows.*
2. *Multiply a row by a nonzero number.*
3. *Replace a row by a multiple of another row added to itself.*

Theorem 6.1.19 *Let A be an $n\times n$ matrix and let A_1 be a matrix which results from multiplying some row of A by a scalar c. Then $c\det(A)=\det(A_1)$.*

Example 6.1.20 *Let* $A=\left(\begin{array}{cc} 1 & 2 \\ 3 & 4 \end{array}\right), A_1=\left(\begin{array}{cc} 2 & 4 \\ 3 & 4 \end{array}\right)$. $\det(A)=-2$, $\det(A_1)=-4$.

Theorem 6.1.21 *Let A be an $n\times n$ matrix and let A_1 be a matrix which results from switching two rows of A. Then $\det(A)=-\det(A_1)$. Also, if one row of A is a multiple of another row of A, then $\det(A)=0$.*

Example 6.1.22 *Let* $A=\left(\begin{array}{cc} 1 & 2 \\ 3 & 4 \end{array}\right)$ *and let* $A_1=\left(\begin{array}{cc} 3 & 4 \\ 1 & 2 \end{array}\right)$. $\det A=-2$, $\det(A_1)=2$.

Theorem 6.1.23 *Let A be an $n\times n$ matrix and let A_1 be a matrix which results from applying row operation 3. That is you replace some row by a multiple of another row added to itself. Then $\det(A)=\det(A_1)$.*

Example 6.1.24 *Let* $A=\left(\begin{array}{cc} 1 & 2 \\ 3 & 4 \end{array}\right)$ *and let* $A_1=\left(\begin{array}{cc} 1 & 2 \\ 4 & 6 \end{array}\right)$. *Thus the second row of A_1 is one times the first row added to the second row.* $\det(A)=-2$ *and* $\det(A_1)=-2$.

Theorem 6.1.25 *In Theorems 6.1.19 - 6.1.23 you can replace the word, "row" with the word "column".*

There are two other major properties of determinants which do not involve row operations.

Theorem 6.1.26 *Let A and B be two $n \times n$ matrices. Then*

$$\boxed{\det(AB) = \det(A)\det(B).}$$

Also,

$$\boxed{\det(A) = \det\left(A^T\right).}$$

Example 6.1.27 *Compare* $\det(AB)$ *and* $\det(A)\det(B)$ *for*

$$A = \begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix}, B = \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix}.$$

First

$$AB = \begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix}\begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 11 & 4 \\ -1 & -4 \end{pmatrix}$$

and so

$$\det(AB) = \det\begin{pmatrix} 11 & 4 \\ -1 & -4 \end{pmatrix} = -40.$$

Now

$$\det(A) = \det\begin{pmatrix} 1 & 2 \\ -3 & 2 \end{pmatrix} = 8,\ \det(B) = \det\begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix} = -5.$$

Thus $\det(A)\det(B) = 8 \times (-5) = -40$.

6.1.4 Finding Determinants Using Row Operations

Theorems 6.1.23 - 6.1.25 can be used to find determinants using row operations. As pointed out above, the method of Laplace expansion will not be practical for any matrix of large size. Here is an example in which all the row operations are used.

Example 6.1.28 *Find the determinant of the matrix*

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 1 & 2 & 3 \\ 4 & 5 & 4 & 3 \\ 2 & 2 & -4 & 5 \end{pmatrix}$$

Replace the second row by (-5) times the first row added to it. Then replace the third row by (-4) times the first row added to it. Finally, replace the fourth row by (-2) times the first row added to it. This yields the matrix

$$B = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -9 & -13 & -17 \\ 0 & -3 & -8 & -13 \\ 0 & -2 & -10 & -3 \end{pmatrix}$$

and from Theorem 6.1.23, it has the same determinant as A. Now using other row operations, $\det(B) = \left(\frac{-1}{3}\right)\det(C)$ where

$$C = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 11 & 22 \\ 0 & -3 & -8 & -13 \\ 0 & 6 & 30 & 9 \end{pmatrix}.$$

The second row was replaced by (-3) times the third row added to the second row. By Theorem 6.1.23 this didn't change the value of the determinant. Then the last row was multiplied by (-3). By Theorem 6.1.19 the resulting matrix has a determinant which is (-3) times the determinant of the un-multiplied matrix. Therefore, we multiplied by $-1/3$ to retain the correct value. Now replace the last row with 2 times the third added to it. This does not change the value of the determinant by Theorem 6.1.23. Finally switch the third and second rows. This causes the determinant to be multiplied by (-1). Thus $\det(C) = -\det(D)$ where

$$D = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -3 & -8 & -13 \\ 0 & 0 & 11 & 22 \\ 0 & 0 & 14 & -17 \end{pmatrix}$$

You could do more row operations or you could note that this can be easily expanded along the first column followed by expanding the 3×3 matrix which results along its first column. Thus

$$\det(D) = 1(-3)\det\begin{pmatrix} 11 & 22 \\ 14 & -17 \end{pmatrix} = 1485$$

and so $\det(C) = -1485$ and $\det(A) = \det(B) = \left(\frac{-1}{3}\right)(-1485) = 495$.

Example 6.1.29 *Find the determinant of the matrix*

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

Replace the second row by (-1) times the first row added to it. Next take -2 times the first row and add to the third and finally take -3 times the first row and add to the last row. This yields

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -5 & -1 & -1 \\ 0 & -3 & -4 & 1 \\ 0 & -10 & -8 & -4 \end{pmatrix}.$$

By Theorem 6.1.23 this matrix has the same determinant as the original matrix. Remember you can work with the columns also. Take -5 times the last column and add to the second column. This yields

$$\begin{pmatrix} 1 & -8 & 3 & 2 \\ 0 & 0 & -1 & -1 \\ 0 & -8 & -4 & 1 \\ 0 & 10 & -8 & -4 \end{pmatrix}$$

By Theorem 6.1.25 this matrix has the same determinant as the original matrix. Now take (-1) times the third row and add to the top row. This gives.

$$\begin{pmatrix} 1 & 0 & 7 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & -8 & -4 & 1 \\ 0 & 10 & -8 & -4 \end{pmatrix}$$

which by Theorem 6.1.23 has the same determinant as the original matrix. Lets expand it now along the first column. This yields the following for the determinant of the original matrix.

$$\det\begin{pmatrix} 0 & -1 & -1 \\ -8 & -4 & 1 \\ 10 & -8 & -4 \end{pmatrix}$$

which equals

$$8\det\begin{pmatrix} -1 & -1 \\ -8 & -4 \end{pmatrix} + 10\det\begin{pmatrix} -1 & -1 \\ -4 & 1 \end{pmatrix} = -82$$

We suggest you do not try to be fancy in using row operations. That is, stick mostly to the one which replaces a row or column with a multiple of another row or column added to it. Also note there is no way to check your answer other than working the problem more than one way. To be sure you have gotten it right you must do this.

6.2 Applications

6.2.1 A Formula For The Inverse

The definition of the determinant in terms of Laplace expansion along a row or column also provides a way to give a formula for the inverse of a matrix. Recall the definition of the inverse of a matrix in Definition 5.1.29 on Page 95. Also recall the definition of the cofactor matrix given in Definition 6.1.12 on Page 112. This cofactor matrix was just the matrix which results from replacing the ij^{th} entry of the matrix with the ij^{th} cofactor.

The following theorem says that to find the inverse, take the transpose of the cofactor matrix and divide by the determinant. The transpose of the cofactor matrix is called the **adjugate** or sometimes the **classical adjoint** of the matrix A. In other words, A^{-1} is equal to one divided by the determinant of A times the adjugate matrix of A. This is what the following theorem says with more precision.

Theorem 6.2.1 *A^{-1} exists if and only if $\det(A) \neq 0$. If $\det(A) \neq 0$, then $A^{-1} = \left(a_{ij}^{-1}\right)$ where*

$$a_{ij}^{-1} = \det(A)^{-1} \operatorname{cof}(A)_{ji}$$

for $\operatorname{cof}(A)_{ij}$ *the* ij^{th} *cofactor of* A.

Example 6.2.2 *Find the inverse of the matrix*

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

First find the determinant of this matrix. Using Theorems 6.1.23 - 6.1.25 on Page 114, 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 the cofactor associated with the position of the entry. Therefore, from the above theorem, the inverse of A should equal

$$\frac{1}{12}\begin{pmatrix} -2 & -2 & 6 \\ 4 & -2 & 0 \\ 2 & 8 & -6 \end{pmatrix}^T = \begin{pmatrix} -1/6 & 1/3 & 1/6 \\ -1/6 & -1/6 & 2/3 \\ 1/2 & 0 & -1/2 \end{pmatrix}.$$

Does it work? You should check to see if it does. When the matrices are multiplied

$$\begin{pmatrix} -1/6 & 1/3 & 1/6 \\ -1/6 & -1/6 & 2/3 \\ 1/2 & 0 & -1/2 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 3 & 0 & 1 \\ 1 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and so it is correct.

Example 6.2.3 *Find the inverse of the matrix*

$$A = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{6} & \frac{1}{3} & -\frac{1}{2} \\ -\frac{5}{6} & \frac{2}{3} & -\frac{1}{2} \end{pmatrix}$$

First find its determinant. This determinant is $\frac{1}{6}$. Now replace each entry by the cofactor associated with the position of the entry. Thus the cofactor associated with the $-\frac{1}{6}$ in the first column is $-\det\begin{pmatrix} 0 & 1/2 \\ 2/3 & -1/2 \end{pmatrix}$. After this, take the transpose of what results and multiply by 6 which is $1/(\det(A))$. Thus, the inverse is

$$6\begin{pmatrix} \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{6} & -\frac{1}{3} \\ -\frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}^T.$$

Then

$$6\begin{pmatrix} 1/6 & 1/3 & 1/6 \\ 1/3 & 1/6 & -1/3 \\ -1/6 & 1/6 & 1/6 \end{pmatrix}^T = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix}$$

which should be the inverse. Always check your work.

$$\begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix}\begin{pmatrix} 1/2 & 0 & 1/2 \\ -1/6 & 1/3 & -1/2 \\ -5/6 & 2/3 & -1/2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and so we got it right. If the result of multiplying these matrices had been something other than the identity matrix, you would know there was an error. When this happens, you need to search for the mistake if you are interested in getting the right answer. A common mistake is to forget to take the transpose of the cofactor matrix.

Proof of Theorem 6.2.1: From the definition of the determinant in terms of expansion along a column, and letting $(a_{ir}) = A$, if $\det(A) \neq 0$,

$$\sum_{i=1}^{n} a_{ir}\operatorname{cof}(A)_{ir}\det(A)^{-1} = \det(A)\det(A)^{-1} = 1.$$

Now consider

$$\sum_{i=1}^{n} a_{ir}\operatorname{cof}(A)_{ik}\det(A)^{-1}$$

when $k \neq r$. Replace the k^{th} column with the r^{th} column to obtain a matrix B_k whose determinant equals zero by Theorem 6.1.21. However, expanding this matrix B_k along the k^{th} column yields

$$0 = \det(B_k)\det(A)^{-1} = \sum_{i=1}^{n} a_{ir}\operatorname{cof}(A)_{ik}\det(A)^{-1}$$

Summarizing,

$$\sum_{i=1}^{n} a_{ir}\operatorname{cof}(A)_{ik}\det(A)^{-1} = \delta_{rk} \equiv \begin{cases} 1 \text{ if } r = k \\ 0 \text{ if } r \neq k \end{cases}.$$

Now

$$\sum_{i=1}^{n} a_{ir}\operatorname{cof}(A)_{ik} = \sum_{i=1}^{n} a_{ir}\operatorname{cof}(A)_{ki}^{T}$$

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 Definition 6.1.13, and similar reasoning,

$$\sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)_{kj} \det(A)^{-1} = \delta_{rk}$$

Now

$$\sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)_{kj} = \sum_{j=1}^{n} a_{rj} \operatorname{cof}(A)_{jk}^T$$

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 6.1.26,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A)\det\left(A^{-1}\right)$$

so $\det(A) \neq 0$. ■

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. It also has enormous theoretical significance in more advanced mathematics.

Example 6.2.4 *Suppose*

$$A(t) = \begin{pmatrix} e^t & 0 & 0 \\ 0 & \cos t & \sin t \\ 0 & -\sin t & \cos t \end{pmatrix}$$

Show that $A(t)^{-1}$ exists and then find it.

First note $\det(A(t)) = e^t \neq 0$ so $A(t)^{-1}$ exists. The cofactor matrix is

$$C(t) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{pmatrix}$$

and so the inverse is

$$\frac{1}{e^t}\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t\cos t & e^t\sin t \\ 0 & -e^t\sin t & e^t\cos t \end{pmatrix}^T = \begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}.$$

6.2.2 Cramer's Rule

This formula for the inverse also implies a famous procedure known as **Cramer's rule**. Cramer's rule gives a formula for the solutions, $\mathbf{x}$, to a system of equations, $A\mathbf{x} = \mathbf{y}$ in the special case that A is a square matrix. Note this rule does not apply if you have a system of equations in which there is a different number of equations than variables.

In case you are solving a system of equations, $A\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$, it follows that if A^{-1} exists,

$$\mathbf{x} = \left(A^{-1}A\right)\mathbf{x} = A^{-1}\left(A\mathbf{x}\right) = A^{-1}\mathbf{y}$$

thus solving the system. Now in the case that A^{-1} exists, there is a formula for A^{-1} given above. Using this formula,

$$x_i = \sum_{j=1}^{n} a_{ij}^{-1} y_j = \sum_{j=1}^{n} \frac{1}{\det(A)} \operatorname{cof}(A)_{ji} y_j.$$

By the formula for the expansion of a determinant along a column,

$$x_i = \frac{1}{\det(A)} \det \begin{pmatrix} * & \cdots & y_1 & \cdots & * \\ \vdots & & \vdots & & \vdots \\ * & \cdots & y_n & \cdots & * \end{pmatrix},$$

where here the i^{th} column of A is replaced with the column vector $(y_1 \cdots\cdot, y_n)^T$, and the determinant of this modified matrix is taken and divided by $\det(A)$. This formula is known as Cramer's rule.

Procedure 6.2.5 *Suppose A is an $n \times n$ matrix and it is desired to solve the system*

$$A\mathbf{x} = \mathbf{y}, \mathbf{y} = (y_1, \cdots, y_n)^T$$

for $\mathbf{x} = (x_1, \cdots, x_n)^T$. Then Cramer's rule says

$$x_i = \frac{\det A_i}{\det A}$$

where A_i is obtained from A by replacing the i^{th} column of A with the column

$$(y_1, \cdots, y_n)^T.$$

Find x, y if

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

The determinant of the matrix of coefficients, $\begin{pmatrix} 1 & 2 & 1 \\ 3 & 2 & 1 \\ 2 & -3 & 2 \end{pmatrix}$ is -14. From Cramer's rule, to get x, you replace the first column of A with the right side of the equation and take its determinant and divide by the determinant of A. Thus

$$x = \frac{\det \begin{pmatrix} 1 & 2 & 1 \\ 2 & 2 & 1 \\ 3 & -3 & 2 \end{pmatrix}}{-14} = \frac{1}{2}$$

Now to find y, z, you do something similar.

$$y = \frac{\det \begin{pmatrix} 1 & 1 & 1 \\ 3 & 2 & 1 \\ 2 & 3 & 2 \end{pmatrix}}{-14} = -\frac{1}{7},\ z = \frac{\det \begin{pmatrix} 1 & 2 & 1 \\ 3 & 2 & 2 \\ 2 & -3 & 3 \end{pmatrix}}{-14} = \frac{11}{14}$$

You see the pattern. For large systems Cramer's rule is less than useful if you want to find an answer. This is because to use it you must evaluate determinants. However, you have no practical way to evaluate determinants for large matrices other than row operations and if you are using row operations, you might just as well use them to solve the system to begin with. It will be a lot less trouble. Nevertheless, there are situations in which Cramer's rule is useful.

Example 6.2.6 *Solve for z if*

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ t \\ t^2 \end{pmatrix}$$

You could do it by row operations but it might be easier in this case to use Cramer's rule because the matrix of coefficients does not consist of numbers but of functions. Thus

$$z = \frac{\det \begin{pmatrix} 1 & 0 & 1 \\ 0 & e^t \cos t & t \\ 0 & -e^t \sin t & t^2 \end{pmatrix}}{\det \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t \cos t & e^t \sin t \\ 0 & -e^t \sin t & e^t \cos t \end{pmatrix}} = t\left((\cos t)\, t + \sin t\right) e^{-t}.$$

You end up doing this sort of thing sometimes in ordinary differential equations in the method of variation of parameters.

6.3 MATLAB And Determinants

MATLAB can find determinants. Here is an example.

```
>> A=[1,3,2,4;-5,7,2,3;2,3,7,11;1,2,3,4]; det(A)
```

Then press enter and you get

```
ans =
-102.0000
```

To enter a complex number $1+2i$ for example, you type: complex(1,2). However, when matlab gives the answer, it will write it in the usual form $1+2i$. If you have matrices in which there are complex entries, you can go ahead and let matlab do the tedious computations for you.

6.4 Exercises

1. Find the determinants of the following matrices.

   (a) $\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 2 \\ 0 & 9 & 8 \end{pmatrix}$ (The answer is 31.)

   (b) $\begin{pmatrix} 4 & 3 & 2 \\ 1 & 7 & 8 \\ 3 & -9 & 3 \end{pmatrix}$ (The answer is 375.)

   (c) $\begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 3 & 2 & 3 \\ 4 & 1 & 5 & 0 \\ 1 & 2 & 1 & 2 \end{pmatrix}$, (The answer is -2.)

2. Find the following determinant by expanding along the first row and second column.

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

3. Find the following determinant by expanding along the first column and third row.

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

4. Find the following determinant by expanding along the second row and first column.

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

5. Compute the determinant by cofactor expansion. Pick the easiest row or column to use.

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

6. Find the determinant using row operations.
$$\det\begin{pmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ -4 & 1 & 2 \end{pmatrix}$$

7. Find the determinant using row operations.
$$\det\begin{pmatrix} 2 & 1 & 3 \\ 2 & 4 & 2 \\ 1 & 4 & -5 \end{pmatrix}$$

8. Find the determinant using row operations.
$$\det\begin{pmatrix} 1 & 2 & 1 & 2 \\ 3 & 1 & -2 & 3 \\ -1 & 0 & 3 & 1 \\ 2 & 3 & 2 & -2 \end{pmatrix}$$

9. Find the determinant using row operations.
$$\det\begin{pmatrix} 1 & 4 & 1 & 2 \\ 3 & 2 & -2 & 3 \\ -1 & 0 & 3 & 3 \\ 2 & 1 & 2 & -2 \end{pmatrix}$$

10. Verify an example of each property of determinants found in Theorems 6.1.23 - 6.1.25 for 2×2 matrices.

11. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

12. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} c & d \\ a & b \end{pmatrix}$$

13. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & b \\ a+c & b+d \end{pmatrix}$$

14. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} a & b \\ 2c & 2d \end{pmatrix}$$

15. An operation is done to get from the first matrix to the second. Identify what was done and tell how it will affect the value of the determinant.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} b & a \\ d & c \end{pmatrix}$$

16. Let A be an $r \times r$ matrix and suppose there are $r-1$ rows (columns) such that all rows (columns) are linear combinations of these $r-1$ rows (columns). Show $\det(A) = 0$.

17. Show $\det(aA) = a^n \det(A)$ where here A is an $n \times n$ matrix and a is a scalar.

18. Illustrate with an example of 2×2 matrices that the determinant of a product equals the product of the determinants.

19. 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.

20. An $n \times n$ matrix is called **nilpotent** if for some positive integer, k it follows $A^k = 0$. If A is a nilpotent matrix and k is the smallest possible integer such that $A^k = 0$, what are the possible values of $\det(A)$?

21. 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?

22. Fill in the missing entries to make the matrix orthogonal as in Problem 21.

$$\begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{\sqrt{12}}{6} \\ \frac{1}{\sqrt{2}} & \_ & \_ \\ \_ & \frac{\sqrt{6}}{3} & \_ \end{pmatrix}.$$

23. 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 $A \sim C$.

24. In the context of Problem 23 show that if $A \sim B$, then $\det(A) = \det(B)$.

25. Two $n \times n$ matrices, A and B, are similar if $B = S^{-1}AS$ for some invertible $n \times n$ matrix S. Show that if two matrices are similar, they have the same characteristic polynomials. The characteristic polynomial of an $n \times n$ matrix M is the polynomial, $\det(\lambda I - M)$.

26. Tell whether the statement is true or false.

(a) If A is a 3×3 matrix with a zero determinant, then one column must be a multiple of some other column.

(b) If any two columns of a square matrix are equal, then the determinant of the matrix equals zero.

(c) For A and B two $n\times n$ matrices, $\det(A+B)=\det(A)+\det(B)$.

(d) For A an $n\times n$ matrix, $\det(3A)=3\det(A)$

(e) If A^{-1} exists then $\det\left(A^{-1}\right)=\det(A)^{-1}$.

(f) If B is obtained by multiplying a single row of A by 4 then $\det(B)=4\det(A)$.

(g) For A an $n\times n$ matrix, $\det(-A)=(-1)^n\det(A)$.

(h) If A is a real $n\times n$ matrix, then $\det\left(A^TA\right)\geq 0$.

(i) Cramer's rule is useful for finding solutions to systems of linear equations in which there is an infinite set of solutions.

(j) If $A^k=0$ for some positive integer, k, then $\det(A)=0$.

(k) If $A\mathbf{x}=\mathbf{0}$ for some $\mathbf{x}\neq\mathbf{0}$, then $\det(A)=0$.

27. Use Cramer's rule to find the solution to $x+2y=1, 2x-y=2$.

28. Use Cramer's rule to find the solution to $x+2y+z=1, 2x-y-z=2,\ x+z=1$.

29. Here is a matrix,
$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 2 & 1 \\ 3 & 1 & 0 \end{pmatrix}$$
Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

30. Here is a matrix,
$$\begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 1 \\ 3 & 1 & 1 \end{pmatrix}$$
Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

31. Here is a matrix,
$$\begin{pmatrix} 1 & 3 & 3 \\ 2 & 4 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$
Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

32. Here is a matrix,

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 2 & 1 \\ 2 & 6 & 7 \end{pmatrix}$$

Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

33. Here is a matrix,

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

Determine whether the matrix has an inverse by finding whether the determinant is non zero. If the determinant is nonzero, find the inverse using the formula for the inverse which involves the cofactor matrix.

34. Use the formula for the inverse in terms of the cofactor matrix to find if possible the inverses of the matrices

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

If the inverse does not exist, explain why.

35. Here is a matrix,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}$$

Does there exist a value of t for which this matrix fails to have an inverse? Explain.

36. Here is a matrix,

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

Does there exist a value of t for which this matrix fails to have an inverse? Explain.

37. Here is a matrix,

$$\begin{pmatrix} e^t & \cosh t & \sinh t \\ e^t & \sinh t & \cosh t \\ e^t & \cosh t & \sinh t \end{pmatrix}$$

Does there exist a value of t for which this matrix fails to have an inverse? Explain.

38. Show that if $\det(A) \neq 0$ for A an $n \times n$ matrix, it follows that if $A\mathbf{x} = \mathbf{0}$, then $\mathbf{x} = \mathbf{0}$.

39. Suppose A,B are $n\times n$ matrices and that $AB=I$. Show that then $BA=I$. **Hint:** You might do something like this: First explain why $\det(A),\det(B)$ are both nonzero. Then $(AB)A=A$ and then show $BA(BA-I)=0$. From this use what is given to conclude $A(BA-I)=0$. Then use Problem 38.

40. Use the formula for the inverse in terms of the cofactor matrix to find the inverse of the matrix
$$A=\begin{pmatrix} e^t & 0 & 0 \\ 0 & e^t\cos t & e^t\sin t \\ 0 & e^t\cos t-e^t\sin t & e^t\cos t+e^t\sin t \end{pmatrix}.$$

41. Find the inverse if it exists of the matrix
$$\begin{pmatrix} e^t & \cos t & \sin t \\ e^t & -\sin t & \cos t \\ e^t & -\cos t & -\sin t \end{pmatrix}.$$

42. Here is a matrix,
$$\begin{pmatrix} e^t & e^{-t}\cos t & e^{-t}\sin t \\ e^t & -e^{-t}\cos t-e^{-t}\sin t & -e^{-t}\sin t+e^{-t}\cos t \\ e^t & 2e^{-t}\sin t & -2e^{-t}\cos t \end{pmatrix}$$
Does there exist a value of t for which this matrix fails to have an inverse? Explain.

43. 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?

44. If A,B, and C are each $n\times n$ matrices and ABC is invertible, why are each of A,B, and C invertible.

45. 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?

46. Let $Ly = y^{(n)} + a_{n-1}(x)\, y^{(n-1)} + \cdots + a_1(x)\, y' + a_0(x)\, y$ where the a_i are given continuous functions defined on a closed interval, (a, b) and y is some function which has n derivatives so it makes sense to write Ly. Suppose $Ly_k = 0$ for $k = 1, 2, \cdots, n$. The **Wronskian** of these functions, y_i is defined as

$$W(y_1, \cdots, y_n)(x) \equiv \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n-1)}(x) & \cdots & y_n^{(n-1)}(x) \end{pmatrix}$$

Show that for $W(x) = W(y_1, \cdots, y_n)(x)$ to save space,

$$W'(x) = \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n)}(x) & \cdots & y_n^{(n)}(x) \end{pmatrix}.$$

Now use the differential equation, $Ly = 0$ which is satisfied by each of these functions, y_i and properties of determinants presented above to verify that

$$W' + a_{n-1}(x)\, W = 0.$$

Give an explicit solution of this linear differential equation, **Abel's formula,** and use your answer to verify that the Wronskian of these solutions to the equation, $Ly = 0$ either vanishes identically on (a, b) or never. **Hint:** To solve the differential equation, let $A'(x) = a_{n-1}(x)$ and multiply both sides of the differential equation by $e^{A(x)}$ and then argue the left side is the derivative of something.

47. Find the following determinants and the inverses of the given matrices. You might use MATLAB to do this with no trouble.

(a) $\det \begin{pmatrix} 2 & 2+2i & 3-3i \\ 2-2i & 5 & 1-7i \\ 3+3i & 1+7i & 16 \end{pmatrix}$ (b) $\det \begin{pmatrix} 10 & 2+6i & 8-6i \\ 2-6i & 9 & 1-7i \\ 8+6i & 1+7i & 17 \end{pmatrix}$

Chapter 7

Determinants Mathematical Theory*

7.0.1 The Function sgn

The following Lemma will be essential in the definition of the determinant.

Lemma 7.0.1 *There exists a function,* sgn_n *which maps each ordered list of numbers from* $\{1, \cdots, n\}$ *to one of the three numbers,* $0, 1,$ *or* -1 *which also has the following properties.*

$$\text{sgn}_n(1, \cdots, n) = 1 \tag{7.1}$$

$$\text{sgn}_n(i_1, \cdots, p, \cdots, q, \cdots, i_n) = -\text{sgn}_n(i_1, \cdots, q, \cdots, p, \cdots, i_n) \tag{7.2}$$

In words, the second property states that if two of the numbers are switched, the value of the function is multiplied by -1*. Also, in the case where* $n > 1$ *and* $\{i_1, \cdots, i_n\} = \{1, \cdots, n\}$ *so that every number from* $\{1, \cdots, n\}$ *appears in the ordered list,* $(i_1, \cdots, i_n)$*,*

$$\begin{gathered}\text{sgn}_n(i_1, \cdots, i_{\theta-1}, n, i_{\theta+1}, \cdots, i_n) \equiv \\ (-1)^{n-\theta} \text{sgn}_{n-1}(i_1, \cdots, i_{\theta-1}, i_{\theta+1}, \cdots, i_n)\end{gathered} \tag{7.3}$$

where $n = i_\theta$ *in the ordered list,* $(i_1, \cdots, i_n)$*.*

Proof: Define $\text{sign}(x) = 1$ if $x > 0, -1$ if $x < 0$ and 0 if $x = 0$. If $n = 1$, there is only one list and it is just the number 1. Thus one can define $\text{sgn}_1(1) \equiv 1$. For the general case where $n > 1$, simply define

$$\text{sgn}_n(i_1, \cdots, i_n) \equiv \text{sign}\left(\prod_{r<s}(i_s - i_r)\right)$$

This delivers either $-1, 1,$ or 0 by definition. What about the other claims? Suppose you switch i_p with i_q where $p < q$ so two numbers in the ordered list $(i_1, \cdots, i_n)$ are switched. Denote the new ordered list of numbers as $(j_1, \cdots, j_n)$. Thus $j_p = i_q$ and $j_q = i_p$ and if $r \notin \{p, q\}$, $j_r = i_r$. See the following illustration.

i_1	i_2	$\cdots$	i_p	$\cdots$	i_q	$\cdots$	i_n
1	2		p		q		n
i_1	i_2	$\cdots$	i_q	$\cdots$	i_p	$\cdots$	i_n
1	2		p		q		n
j_1	j_2	$\cdots$	j_p	$\cdots$	j_q	$\cdots$	j_n
1	2		p		q		n

Then

$$\operatorname{sgn}_n(j_1,\cdots,j_n) \equiv \operatorname{sign}\left(\prod_{r<s}(j_s-j_r)\right)$$

$$= \operatorname{sign}\left(\overset{\text{both } p,q}{(i_p-i_q)}\overbrace{\prod_{p<j<q}(i_j-i_q)\prod_{p<j<q}(i_p-i_j)}^{\text{one of } p,q}\overset{\text{neither } p \text{ nor } q}{\prod_{r<s,r,s\notin\{p,q\}}}(i_s-i_r)\right)$$

The last product consists of the product of terms which were in the un-switched product

$$\prod_{r<s}(i_s-i_r)$$

so produces no change in sign, while the two products in the middle both introduce $q-p-1$ minus signs. Thus their product produces no change in sign. The first factor is of opposite sign to the i_q-i_p which occured in $\operatorname{sgn}_n(i_1,\cdots,i_n)$. Therefore, this switch introduced a minus sign and

$$\operatorname{sgn}_n(j_1,\cdots,j_n) = -\operatorname{sgn}_n(i_1,\cdots,i_n)$$

Now consider the last claim. In computing $\operatorname{sgn}_n(i_1,\cdots,i_{\theta-1},n,i_{\theta+1},\cdots,i_n)$ there will be the product of $n-\theta$ negative terms

$$(i_{\theta+1}-n)\cdots(i_n-n)$$

and the other terms in the product for computing $\operatorname{sgn}_n(i_1,\cdots,i_{\theta-1},n,i_{\theta+1},\cdots,i_n)$ are those which are required to compute $\operatorname{sgn}_{n-1}(i_1,\cdots,i_{\theta-1},i_{\theta+1},\cdots,i_n)$ multiplied by terms of the form $(n-i_j)$ which are nonnegative. It follows that

$$\operatorname{sgn}_n(i_1,\cdots,i_{\theta-1},n,i_{\theta+1},\cdots,i_n) = (-1)^{n-\theta}\operatorname{sgn}_{n-1}(i_1,\cdots,i_{\theta-1},i_{\theta+1},\cdots,i_n)$$

It is obvious that if there are repeats in the list the function gives 0. ■

Lemma 7.0.2 *Every ordered list of distinct numbers from $\{1,2,\cdots,n\}$ can be obtained from every other such ordered list by a finite number of switches. Also,* sgn_n *is unique.*

Proof: This is obvious if $n=1$ or 2. Suppose then that it is true for sets of $n-1$ elements. Take two ordered lists of numbers, P_1,P_2. Make one switch in both to place n at the end. Call the result P_1^n and P_2^n. Then using induction, there are finitely many switches in P_1^n so that it will coincide with P_2^n. Now switch the n in what results to where it was in P_2.

To see sgn_n is unique, if there exist two functions, f and g both satisfying 7.1 and 7.2, you could start with $f(1,\cdots,n) = g(1,\cdots,n) = 1$ 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 7.2 gives both functions are equal to zero for that ordered list. ■

Definition 7.0.3 *An ordered list of distinct numbers from* $\{1,2,\cdots,n\}$, *say*

$$(i_1,\cdots,i_n),$$

is called a permutation. The symbol for all such permutations is S_n. *The number* $\operatorname{sgn}_n(i_1,\cdots,i_n)$ *is called the sign of the permutation.*

A permutation can also be considered as a function from the set

$$\{1,2,\cdots,n\} \text{ to } \{1,2,\cdots,n\}$$

as follows. Let $f(k) = i_k$. Permutations are of fundamental importance in certain areas of math. For example, it was by considering permutations that Galois was able to give a criterion for solution of polynomial equations by radicals, but this is a different direction than what is being attempted here.

In what follows sgn will often be used rather than sgn_n because the context supplies the appropriate n.

7.1 The Determinant

Definition 7.1.1 *Let* f *be a 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,*

$$\sum_{(k_1,k_2)} f(k_1,k_2) = f(1,2)+f(2,1)+f(1,1)+f(2,2).$$

7.1.1 The Definition

Definition 7.1.2 *Let* $(a_{ij}) = A$ *denote an* $n\times n$ *matrix. The determinant of* A, *denoted by* $\det(A)$ *is defined by*

$$\det(A) \equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots a_{nk_n}$$

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.*

7.1.2 Permuting Rows Or Columns

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

$$\det(A(r_1,\cdots,r_n)) = \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n} \tag{7.4}$$

and

$$A(1,\cdots,n) = A.$$

Proposition 7.1.3 *Let*

$$(r_1,\cdots,r_n)$$

be an ordered list of numbers from $\{1,\cdots,n\}$*. Then*

$$\operatorname{sgn}(r_1,\cdots,r_n)\det(A)$$

$$= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n} \tag{7.5}$$

$$= \det(A(r_1,\cdots,r_n)). \tag{7.6}$$

Proof: Let $(1,\cdots,n) = (1,\cdots,r,\cdots s,\cdots,n)$ so $r < s$.

$$\det(A(1,\cdots,r,\cdots,s,\cdots,n)) = \tag{7.7}$$

$$\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_r,\cdots,k_s,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_r}\cdots a_{sk_s}\cdots a_{nk_n},$$

and renaming the variables, calling k_s, k_r and k_r, k_s, this equals

$$= \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_s,\cdots,k_r,\cdots,k_n)\, a_{1k_1}\cdots a_{rk_s}\cdots a_{sk_r}\cdots a_{nk_n}$$

$$= \sum_{(k_1,\cdots,k_n)} -\operatorname{sgn}\left(k_1,\cdots,\overbrace{k_r,\cdots,k_s}^{\text{These got switched}},\cdots,k_n\right) a_{1k_1}\cdots a_{sk_r}\cdots a_{rk_s}\cdots a_{nk_n}$$

$$= -\det(A(1,\cdots,s,\cdots,r,\cdots,n)). \tag{7.8}$$

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 7.0.1, 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 7.7 -7.8 shows that $\det A(r_1,\cdots,r_n) = 0$ and also $\operatorname{sgn}(r_1,\cdots,r_n) = 0$ so the formula holds in this case also. ■

Observation 7.1.4 *There are* $n!$ *ordered lists of distinct numbers from* $\{1,\cdots,n\}$.

To see this, consider n slots placed in order. There are n choices for the first slot. For each of these choices, there are $n-1$ choices for the second. Thus there are $n(n-1)$ 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.

7.1.3 A Symmetric Definition

With the above, it is possible to give a more symmetric description of the determinant from which it will follow that $\det(A) = \det\left(A^T\right)$.

Corollary 7.1.5 *The following formula for* $\det(A)$ *is valid.*

$$\det(A) = \frac{1}{n!}\cdot$$

$$\sum_{(r_1,\cdots,r_n)}\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(r_1,\cdots,r_n)\operatorname{sgn}(k_1,\cdots,k_n)\, a_{r_1k_1}\cdots a_{r_nk_n}. \tag{7.9}$$

And also $\det\left(A^T\right) = \det(A)$ *where* A^T *is the transpose of* A*. (Recall that* $\left(A^T\right)_{ij} = A_{ji}$*.)*

Proof: From Proposition 7.1.3, 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. ■

7.1.4 The Alternating Property Of The Determinant

Corollary 7.1.6 *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 7.1.3 when two rows are switched, the determinant of the resulting matrix is (-1) times the determinant of the original matrix. By Corollary 7.1.5 the same holds for columns because the columns of the matrix equal the rows of the transposed matrix. Thus if A_1 is the matrix obtained from A by switching two columns,

$$\det(A) = \det\left(A^T\right) = -\det\left(A_1^T\right) = -\det(A_1).$$

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.

$$\begin{aligned}\det(A) &\equiv \sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots\left(xa_{k_i}+yb_{k_i}\right)\cdots a_{nk_n} \\ &= x\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots a_{k_i}\cdots a_{nk_n} \\ &+y\sum_{(k_1,\cdots,k_n)} \operatorname{sgn}(k_1,\cdots,k_n)\, a_{1k_1}\cdots b_{k_i}\cdots a_{nk_n} \\ &\equiv x\det(A_1)+y\det(A_2).\end{aligned}$$

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. ■

7.1.5 Linear Combinations And Determinants

Linear combinations have been discussed already. However, here is a review and some new terminology.

Definition 7.1.7 *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 7.1.8 *Suppose* A *is an* $n\times n$ *matrix and some column (row) is a linear combination of* r *other columns (rows). Then* $\det(A)=0$.

Proof: Let $A=\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{pmatrix}$ 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 7.1.6 the determinant of A is zero if and only if the determinant of the matrix B, which has this special column placed in the last position, equals zero. Thus $\mathbf{a}_n=\sum_{k=1}^{r} c_k\mathbf{a}_k$ and so

$$\det(B)=\det\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \sum_{k=1}^{r} c_k\mathbf{a}_k\end{pmatrix}.$$

By Corollary 7.1.6

$$\det(B)=\sum_{k=1}^{r} c_k\det\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_r & \cdots & \mathbf{a}_{n-1} & \mathbf{a}_k\end{pmatrix}=0.$$

because there are two equal columns. The case for rows follows from the fact that $\det(A)=\det\left(A^T\right)$. ■

7.1.6 The Determinant Of A Product

Recall the following definition of matrix multiplication.

Definition 7.1.9 *If A and B are $n\times n$ matrices, $A=(a_{ij})$ and $B=(b_{ij})$, $AB=(c_{ij})$ where*

$$c_{ij}\equiv\sum_{k=1}^{n}a_{ik}b_{kj}.$$

One of the most important rules about determinants is that the determinant of a product equals the product of the determinants.

Theorem 7.1.10 *Let A and B be $n\times n$ matrices. Then*

$$\det(AB)=\det(A)\det(B).$$

Proof: Let c_{ij} be the ij^{th} entry of AB. Then by Proposition 7.1.3,

$$\det(AB)=$$

$$\begin{aligned}&\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_n)\,c_{1k_1}\cdots c_{nk_n}\\ =&\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_n)\left(\sum_{r_1}a_{1r_1}b_{r_1k_1}\right)\cdots\left(\sum_{r_n}a_{nr_n}b_{r_nk_n}\right)\\ =&\sum_{(r_1\cdots,r_n)}\sum_{(k_1,\cdots,k_n)}\operatorname{sgn}(k_1,\cdots,k_n)\,b_{r_1k_1}\cdots b_{r_nk_n}\left(a_{1r_1}\cdots a_{nr_n}\right)\\ =&\sum_{(r_1\cdots,r_n)}\operatorname{sgn}(r_1\cdots r_n)\,a_{1r_1}\cdots a_{nr_n}\det(B)=\det(A)\det(B).\ \blacksquare\end{aligned}$$

7.1.7 Cofactor Expansions

Lemma 7.1.11 *Suppose a matrix is of the form*

$$M=\begin{pmatrix}A&*\\ \mathbf{0}&a\end{pmatrix}\tag{7.10}$$

or

$$M=\begin{pmatrix}A&\mathbf{0}\\ *&a\end{pmatrix}\tag{7.11}$$

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 Lemma 7.0.1, $\det(M)$ 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 7.11. 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 7.10 use Corollary 7.1.5 and 7.11 to write

$$\det(M) = \det\left(M^T\right) = \det\left(\left(\begin{array}{cc} A^T & \mathbf{0} \\ * & a \end{array}\right)\right) = a\det\left(A^T\right) = a\det(A). \blacksquare$$

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 7.1.12 *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 7.1.13 *Let A be an $n \times n$ matrix where $n \geq 2$. Then*

$$\det(A) = \sum_{j=1}^{n} a_{ij} \operatorname{cof}(A)_{ij} = \sum_{i=1}^{n} a_{ij} \operatorname{cof}(A)_{ij}. \tag{7.12}$$

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 7.1.6,

$$\det(A) = \sum_{j=1}^{n} \det(B_j)$$

Denote by A^{ij} the $(n-1)\times(n-1)$ matrix obtained by deleting the i^{th} row and the j^{th} column of A. Thus $\operatorname{cof}(A)_{ij} \equiv (-1)^{i+j}\det\left(A^{ij}\right)$. At this point, recall that from Proposition 7.1.3, when two rows or two columns in a matrix M, are switched, this results in multiplying the determinant of the old matrix by -1 to get the determinant of the new matrix. Therefore, by Lemma 7.1.11,

$$\begin{aligned}\det(B_j) &= (-1)^{n-j}(-1)^{n-i}\det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij}\end{pmatrix}\right) \\ &= (-1)^{i+j}\det\left(\begin{pmatrix} A^{ij} & * \\ \mathbf{0} & a_{ij}\end{pmatrix}\right) = a_{ij}\operatorname{cof}(A)_{ij}.\end{aligned}$$

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. ■

7.1.8 Formula For The Inverse

Note that this gives an easy way to write a formula for the inverse of an $n\times n$ matrix.

Theorem 7.1.14 *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 7.1.13 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 7.1.6. 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 7.1.13, 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 7.1.10,

$$1 = \det(I) = \det\left(AA^{-1}\right) = \det(A)\det\left(A^{-1}\right)$$

so $\det(A) \neq 0$. ■

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 7.1.15 *Let A be an $n \times n$ matrix and suppose there exists an $n \times n$ matrix B such that $BA = I$. Then A^{-1} exists and $A^{-1} = B$. Also, if there exists C an $n \times n$ matrix such that $AC = I$, then A^{-1} exists and $A^{-1} = C$.*

Proof: Since $BA = I$, Theorem 7.1.10 implies

$$\det B \det A = 1$$

and so $\det A \neq 0$. Therefore from Theorem 7.1.14, 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 7.1.14 says that to find the inverse, take the transpose of the cofactor matrix and divide by the determinant. The transpose of the cofactor matrix is called the adjugate or sometimes the classical adjoint of the matrix A. It is an abomination to call it the adjoint although you do sometimes see it referred to in this way. In words, A^{-1} is equal to one over the determinant of A times the adjugate matrix of A.

7.1.9 Cramer's Rule

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.

7.1.10 Upper Triangular Matrices

Definition 7.1.16 *A matrix M, is upper triangular if $M_{ij} = 0$ whenever $i > j$. Thus such a matrix equals zero below the main diagonal, the entries of the form M_{ii} as shown.*

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

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 7.1.13.

Corollary 7.1.17 *Let M be an upper (lower) triangular matrix. Then $\det(M)$ is obtained by taking the product of the entries on the main diagonal.*

7.2 The Cayley Hamilton Theorem*

Definition 7.2.1 *Let A be an $n \times n$ matrix. The characteristic polynomial is defined as*

$$q_A(t) \equiv \det(tI - A)$$

and the solutions to $q_A(t) = 0$ 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 $q_A(t) = 0$. It is one of the most important theorems in linear algebra[1]. The proof in this section is not the most general proof, but works well when the field of scalars is $\mathbb{R}$ or $\mathbb{C}$. The following lemma will help with its proof.

[1] A special case was first proved by Hamilton in 1853. The general case was announced by Cayley some time later and a proof was given by Frobenius in 1878.

Lemma 7.2.2 *Suppose for all $|\lambda|$ large enough,*

$$A_0 + A_1\lambda + \cdots + A_m\lambda^m = 0,$$

where the A_i are $n \times n$ matrices. Then each $A_i = 0$.

Proof: Multiply by λ^{-m} to obtain

$$A_0\lambda^{-m} + A_1\lambda^{-m+1} + \cdots + A_{m-1}\lambda^{-1} + A_m = 0.$$

Now let $|\lambda| \to \infty$ to obtain $A_m = 0$. With this, multiply by λ to obtain

$$A_0\lambda^{-m+1} + A_1\lambda^{-m+2} + \cdots + A_{m-1} = 0.$$

Now let $|\lambda| \to \infty$ to obtain $A_{m-1} = 0$. Continue multiplying by λ and letting $\lambda \to \infty$ to obtain that all the $A_i = 0$. ■

With the lemma, here is a simple corollary.

Corollary 7.2.3 *Let A_i and B_i be $n \times n$ matrices and suppose*

$$A_0 + A_1\lambda + \cdots + A_m\lambda^m = B_0 + B_1\lambda + \cdots + B_m\lambda^m$$

for all $|\lambda|$ large enough. Then $A_i = B_i$ for all i. If $A_i = B_i$ for each A_i, B_i then one can substitute an $n \times n$ matrix M for λ and the identity will continue to hold.

Proof: Subtract and use the result of the lemma. The last claim is obvious by matching terms. ■

With this preparation, here is a relatively easy proof of the Cayley Hamilton theorem.

Theorem 7.2.4 *Let A be an $n \times n$ matrix and let $q(\lambda) \equiv \det(\lambda I - A)$ be the characteristic polynomial. Then $q(A) = 0$.*

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 7.1.14

$$C(\lambda) = q(\lambda)(\lambda I - A)^{-1}.$$

Say

$$q(\lambda) = a_0 + a_1\lambda + \cdots + \lambda^n$$

Note that each entry in $C(\lambda)$ is a polynomial in λ having degree no more than $n-1$. For example, you might have something like

$$C(\lambda) = \begin{pmatrix} \lambda^2 - 6\lambda + 9 & 3 - \lambda & 0 \\ 2\lambda - 6 & \lambda^2 - 3\lambda & 0 \\ \lambda - 1 & \lambda - 1 & \lambda^2 - 3\lambda + 2 \end{pmatrix}$$

$$= \begin{pmatrix} 9 & 3 & 0 \\ -6 & 0 & 0 \\ -1 & -1 & 2 \end{pmatrix} + \lambda \begin{pmatrix} -6 & -1 & 0 \\ 2 & -3 & 0 \\ 1 & 1 & -3 \end{pmatrix} + \lambda^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Therefore, collecting the terms in the general case,

$$C(\lambda) = C_0 + C_1\lambda + \cdots + C_{n-1}\lambda^{n-1}$$

for C_j some $n \times n$ matrix. Then

$$C(\lambda)(\lambda I - A) = \left(C_0 + C_1\lambda + \cdots + C_{n-1}\lambda^{n-1}\right)(\lambda I - A) = q(\lambda) I$$

Then multiplying out the middle term, it follows that for all $|\lambda|$ sufficiently large,

$$\begin{aligned} a_0 I + a_1 I\lambda + \cdots + I\lambda^n &= C_0\lambda + C_1\lambda^2 + \cdots + C_{n-1}\lambda^n \\ &- \left[C_0 A + C_1 A\lambda + \cdots + C_{n-1} A\lambda^{n-1}\right] \\ = -C_0 A + (C_0 - C_1 A)\lambda &+ (C_1 - C_2 A)\lambda^2 + \cdots + (C_{n-2} - C_{n-1}A)\lambda^{n-1} + C_{n-1}\lambda^n \end{aligned}$$

Then, using Corollary 7.2.3, one can replace λ on both sides with A. Then the right side is seen to equal 0. Hence the left side, $q(A) I$ is also equal to 0. ■

It is good to keep in mind the following example when considering the above proof of the Cayley Hamilton theorem. It was shown to me by Marc van Leeuwen. If $p(\lambda) = q(\lambda)$ for all λ or for all λ large enough where $p(\lambda), q(\lambda)$ are polynomials having matrix coefficients, then it is not necessarily the case that $p(A) = q(A)$ for A a matrix of an appropriate size. Let

$$E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, E_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Then a short computation shows that for all complex λ,

$$(\lambda I + E_1)(\lambda I + E_2) = \left(\lambda^2 + \lambda\right) I = (\lambda I + E_2)(\lambda I + E_1)$$

However,

$$(NI + E_1)(NI + E_2) \neq (NI + E_2)(NI + E_1)$$

The reason this can take place is that N fails to commute with E_i. Of course a scalar commutes with any matrix so there was no difficulty in obtaining that the matrix equation held for arbitrary λ, but this factored equation does not continue to hold if λ is replaced by a matrix. In the above proof of the Cayley Hamilton theorem, this issue was avoided by considering only polynomials which are of the form $C_0 + C_1\lambda + \cdots$ in which the polynomial identity held because the corresponding matrix coefficients were equal. However, you can also argue that in the above proof, the C_i each commute with A. Nevertheless, an earlier proof of the Cayley Hamilton theorem using this approach was misleading because this issue was not made clear.

Chapter 8

Rank Of A Matrix

8.1 Elementary Matrices

The elementary matrices result from doing a row operation to the identity matrix.

Definition 8.1.1 *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 8.1.2 *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.

When you multiply by the identity, nothing happens, but when you multiply by an elementary matrix you end up doing a row operation. The next definition is what is meant by an elementary matrix.

[1] More generally, a permutation matrix is a matrix which comes by permuting the rows of the identity matrix, not just switching two rows.

Definition 8.1.3 *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*[2].

The importance of elementary matrices is that when you multiply on the left by one, it does the row operation which was used to produce the elementary matrix.

Now consider what these elementary matrices look like. First consider the one which involves switching row i and row j where $i < j$. This matrix is of the form

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

Note how the i^{th} and j^{th} rows are switched in the identity matrix and there are thus all ones on the main diagonal except for those two positions indicated. The two exceptional rows are shown. The i^{th} row was the j^{th} and the j^{th} row was the i^{th} in the identity matrix. Now consider what this does to a column vector.

$$\begin{pmatrix} \ddots & & & & \\ & 0 & & 1 & \\ & & \ddots & & \\ & 1 & & 0 & \\ & & & & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ x_i \\ \vdots \\ x_j \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ x_j \\ \vdots \\ x_i \\ \vdots \end{pmatrix}$$

Now denote by P^{ij} the elementary matrix which comes from the identity from switching rows i and j. From what was just explained and Proposition 5.1.15,

$$P^{ij} \begin{pmatrix} \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & & \vdots \\ a_{j1} & a_{j2} & \cdots & a_{jp} \\ \vdots & \vdots & & \vdots \end{pmatrix} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ a_{j1} & a_{j2} & \cdots & a_{jp} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & & \vdots \end{pmatrix}$$

This has established the following lemma.

Lemma 8.1.4 *Let P^{ij} denote the elementary matrix which involves switching the i^{th} and the j^{th} rows. Then*

$$P^{ij}A = B$$

where B is obtained from A by switching the i^{th} and the j^{th} rows.

[2] More generally, a permutation matrix is a matrix which comes by permuting the rows of the identity matrix, which means possibly more than two rows are switched.

Example 8.1.5 *Consider the following.*

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ g & d \\ e & f \end{pmatrix} = \begin{pmatrix} g & d \\ a & b \\ e & f \end{pmatrix}$$

Next consider the row operation which involves multiplying the i^{th} row by a nonzero constant, c. The elementary matrix which results from applying this operation to the i^{th} row of the identity matrix is of the form

$$\begin{pmatrix} \ddots & & & & 0 \\ & 1 & & & \\ & & c & & \\ & & & 1 & \\ 0 & & & & \ddots \end{pmatrix}$$

Now consider what this does to a column vector.

$$\begin{pmatrix} \ddots & & & & 0 \\ & 1 & & & \\ & & c & & \\ & & & 1 & \\ 0 & & & & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ v_{i-1} \\ v_i \\ v_{i+1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ v_{i-1} \\ cv_i \\ v_{i+1} \\ \vdots \end{pmatrix}$$

Denote by $E(c,i)$ this elementary matrix which multiplies the i^{th} row of the identity by the nonzero constant, c. Then from what was just discussed and Proposition 5.1.15,

$$E(c,i) \begin{pmatrix} \vdots & \vdots & & \vdots \\ a_{(i-1)1} & a_{(i-1)2} & \cdots & a_{(i-1)p} \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ a_{(i+1)1} & a_{(i+1)2} & \cdots & a_{(i+1)p} \\ \vdots & \vdots & & \vdots \end{pmatrix} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ a_{(i-1)1} & a_{(i-1)2} & \cdots & a_{(i-1)p} \\ ca_{i1} & ca_{i2} & \cdots & ca_{ip} \\ a_{(i+1)1} & a_{(i+1)2} & \cdots & a_{(i+1)p} \\ \vdots & \vdots & & \vdots \end{pmatrix}$$

This proves the following lemma.

Lemma 8.1.6 *Let $E(c,i)$ denote the elementary matrix corresponding to the row operation in which the i^{th} row is multiplied by the nonzero constant, c. Thus $E(c,i)$ involves multiplying the i^{th} row of the identity matrix by c. Then*

$$E(c,i)A = B$$

where B is obtained from A by multiplying the i^{th} row of A by c.

Example 8.1.7 *Consider this.*

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} = \begin{pmatrix} a & b \\ 5c & 5d \\ e & f \end{pmatrix}$$

Finally consider the third of these row operations. Denote by $E(c\times i+j)$ the elementary matrix which replaces the j^{th} row with the j^{th} row added to c times the i^{th} row. In case $i<j$ this will be of the form

$$\begin{pmatrix} \ddots & & & & 0 \\ & 1 & & & \\ & & \ddots & & \\ & c & & 1 & \\ 0 & & & & \ddots \end{pmatrix}$$

Now consider what this does to a column vector.

$$\begin{pmatrix} \ddots & & & & 0 \\ & 1 & & & \\ & & \ddots & & \\ & c & & 1 & \\ 0 & & & & \ddots \end{pmatrix}\begin{pmatrix} \vdots \\ v_i \\ \vdots \\ v_j \\ \vdots \end{pmatrix}=\begin{pmatrix} \vdots \\ v_i \\ \vdots \\ cv_i+v_j \\ \vdots \end{pmatrix}$$

Now from this and Proposition 5.1.15,

$$E(c\times i+j)\begin{pmatrix} \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & & \vdots \\ a_{j1} & a_{j2} & \cdots & a_{jp} \\ \vdots & \vdots & & \vdots \end{pmatrix}$$

$$=\begin{pmatrix} \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ip} \\ \vdots & \vdots & & \vdots \\ ca_{i1}+a_{j1} & ca_{i2}+a_{j2} & \cdots & ca_{ip}+a_{jp} \\ \vdots & \vdots & & \vdots \end{pmatrix}$$

The case where $i>j$ is handled similarly. This proves the following lemma.

Lemma 8.1.8 *Let $E(c\times i+j)$ denote the elementary matrix obtained from I by replacing the j^{th} row with c times the i^{th} row added to it. Then*

$$E(c\times i+j)A=B$$

where B is obtained from A by replacing the j^{th} row of A with itself added to c times the i^{th} row of A.

Example 8.1.9 *Consider the third row operation.*

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \\ 2a+e & 2b+f \end{pmatrix}$$

The next theorem is the main result.

Theorem 8.1.10 *To perform any of the three row operations on a matrix A, it suffices to do the row operation on the identity matrix obtaining an elementary matrix E and then take the product, EA. Furthermore, if E is an elementary matrix, then there is another elementary matrix $\hat{E}$ such that $E\hat{E} = \hat{E}E = I$.*

Proof: The first part of this theorem has been proved in Lemmas 8.1.4 - 8.1.8. It only remains to verify the claim about the matrix $\hat{E}$. Consider first the elementary matrices corresponding to row operation of type three.

$$E(-c \times i + j) E(c \times i + j) = I.$$

This follows because the first matrix takes c times row i in the identity and adds it to row j. When multiplied on the left by $E(-c \times i + j)$ it follows from the first part of this theorem that you take the i^{th} row of $E(c \times i + j)$ which coincides with the i^{th} row of I since that row was not changed, multiply it by $-c$ and add to the j^{th} row of $E(c \times i + j)$ which was the j^{th} row of I added to c times the i^{th} row of I. Thus $E(-c \times i + j)$ multiplied on the left, undoes the row operation which resulted in $E(c \times i + j)$. The same argument applied to the product $E(c \times i + j) E(-c \times i + j)$ replacing c with $-c$ in the argument yields that this product is also equal to I. Therefore, there is an elementary matrix of the same sort which when multiplied by E on either side gives the identity.

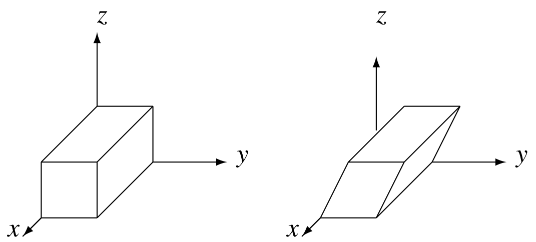
Similar reasoning shows that for $E(c,i)$ the elementary matrix which comes from multiplying the i^{th} row by the nonzero constant c, you can take $\hat{E} = E((1/c), i)$.

Finally, consider P^{ij} which involves switching the i^{th} and the j^{th} rows $P^{ij}P^{ij} = I$ because by the first part of this theorem, multiplying on the left by P^{ij} switches the i^{th} and j^{th} rows of P^{ij} which was obtained from switching the i^{th} and j^{th} rows of the identity. First you switch them to get P^{ij} and then you multiply on the left by P^{ij} which switches these rows again and restores the identity matrix. ■

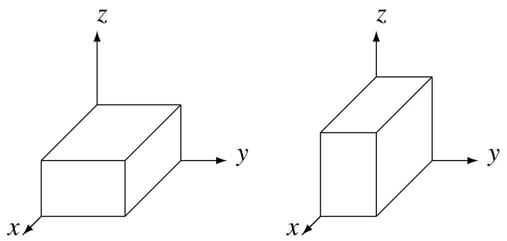
To summarize the last part, if $E(c \times i + j)^{-1} = E(-c \times i + j)$, $\left(P^{ij}\right)^{-1} = P^{ij}$, and

$$E(c,i)^{-1} = E\left(\frac{1}{c}, i\right).$$

The geometric significance of these elementary operations is interesting. The following picture shows the effect of doing $E\left(\frac{1}{2} \times 3 + 1\right)$ on a box. You will see that it shears the box in one direction. Of course there would be corresponding shears in the other directions also. Note that this does not change the volume.



The other elementary matrices have similar simple geometric interpretations. For example, $E(c,i)$ merely multiplies the i^{th} variable by c. It stretches or contracts the box in that direction. If c is negative, it also causes the box to be reflected in this direction. The following picture illustrates the effect of P^{13} on a box in three dimensions. It changes the x and the z values.



8.2 THE Row Reduced Echelon Form Of A Matrix

Recall that putting a matrix in row reduced echelon form involves doing row operations as described on Page 62. In this section we review the description of the row reduced echelon form and prove the row reduced echelon form for a given matrix is unique. That is, every matrix can be row reduced to a unique row reduced echelon form. Of course this is not true of the echelon form. The significance of this is that it becomes possible to use the definite article in referring to **the** row reduced echelon form and hence important conclusions about the original matrix may be logically deduced from an examination of its unique row reduced echelon form. First we need the following definition of some terminology.

Definition 8.2.1 *Let $\mathbf{v}_1,\cdots,\mathbf{v}_k,\mathbf{u}$ be vectors. Then $\mathbf{u}$ is said to be a **linear combination** of the vectors $\{\mathbf{v}_1,\cdots,\mathbf{v}_k\}$ if there exist scalars, $c_1,\cdots,c_k$ such that*

$$\mathbf{u}=\sum_{i=1}^{k}c_i\mathbf{v}_i.$$

*The collection of all linear combinations of the vectors, $\{\mathbf{v}_1,\cdots,\mathbf{v}_k\}$ is known as the **span** of these vectors and is written as $\operatorname{span}(\mathbf{v}_1,\cdots,\mathbf{v}_k)$.*

Another way to say the same thing as expressed in the earlier definition of row reduced echelon form found on Page 61 is the following which is a more useful description when proving the major assertions about the row reduced echelon form.

Definition 8.2.2 *Let* $\mathbf{e}_i$ *denote the column vector which has all zero entries except for the* i^{th} *slot which is one. An* $m \times n$ *matrix is said to be in* ***row reduced echelon form*** *if, in viewing successive columns from left to right, the first nonzero column encountered is* $\mathbf{e}_1$ *and if you have encountered* $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$*, the next column is either* $\mathbf{e}_{k+1}$ *or is a linear combination of the vectors,* $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$*.*

Theorem 8.2.3 *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 equal to zero. Thus the first nonzero column is now $\mathbf{e}_1$. Denote the resulting matrix by A_1. Consider the sub-matrix of A_1 to the right of this column and below the first row. Do exactly the same thing for this sub-matrix that was done for A. This time the $\mathbf{e}_1$ will refer to $\mathbb{F}^{m-1}$. Use the first 1 obtained by the above process which is in the top row of this sub-matrix 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.

Recall the following definition which was discussed earlier.

Definition 8.2.4 *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 but first, here is a fundamental lemma.

In rough terms, the following lemma states that **linear relationships** between columns in a matrix are preserved by row operations. This simple lemma is the main result in understanding all the major questions related to the row reduced echelon form as well as many other topics.

Lemma 8.2.5 *Let A and B be two $m \times n$ matrices and suppose B results from a row operation applied to A. Then the k^{th} column of B is a linear combination of the $i_1, \cdots, i_r$ columns of B if and only if the k^{th} column of A is a linear combination of the $i_1, \cdots, i_r$ columns of A. Furthermore, the scalars in the linear combination are the same. (The linear relationship between the k^{th} column of A and the $i_1, \cdots, i_r$ columns of A is the same as the linear relationship between the k^{th} column of B and the $i_1, \cdots, i_r$ columns of B.)*

Proof: Let A equal the following matrix in which the $\mathbf{a}_k$ are the columns

$$\begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{pmatrix}$$

and let B equal the following matrix in which the columns are given by the $\mathbf{b}_k$

$$\begin{pmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \end{pmatrix}$$

Then by Theorem 8.1.10 on Page 149 $\mathbf{b}_k = E\mathbf{a}_k$ 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,

$$\mathbf{b}_k = E\mathbf{a}_k = \sum_{r \in S} c_r E\mathbf{a}_r = \sum_{r \in S} c_r \mathbf{b}_r. \blacksquare$$

Definition 8.2.6 *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. Note

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n$$

so to say two column vectors are equal is to say they are the same linear combination of the special vectors $\mathbf{e}_j$.

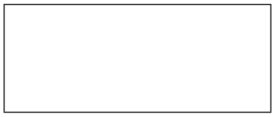
Thus the row reduced echelon form is completely determined by the positions of columns which are not linear combinations of preceding columns (These become the $\mathbf{e}_i$ vectors in the row reduced echelon form.) and the scalars which are used in the linear combinations of these special pivot columns to obtain the other columns. All of these considerations pertain only to linear relations between the columns of the matrix, which by Lemma 8.2.5 are all preserved. Therefore, there is only one row reduced echelon form for any given matrix. The proof of the following corollary is just a more careful exposition of this simple idea.

Corollary 8.2.7 *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. In reading from left to right in B, suppose $\mathbf{e}_1, \cdots, \mathbf{e}_r$ occur first in positions $i_1, \cdots, i_r$ respectively. The description of the row reduced echelon form means that each of these columns is not a linear combination of the preceding columns. Therefore, by Lemma 8.2.5, the same is true of the columns in positions $i_1, i_2, \cdots, i_r$ for C. It follows from the description of the row reduced echelon form that in C, $\mathbf{e}_1, \cdots, \mathbf{e}_r$ occur first in positions $i_1, i_2, \cdots, i_r$. Therefore, both B and C have the sequence $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_r$ occurring first in the positions, $i_1, i_2, \cdots, i_r$. By Lemma 8.2.5, the columns between the i_k and i_{k+1} position in the two matrices are linear combinations involving the same scalars of the columns in the $i_1, \cdots, i_k$ position. Also the columns after the i_r position are linear combinations of the columns in the $i_1, \cdots, i_r$ positions involving the same scalars in both matrices. This is equivalent to the assertion that each of these columns is identical and this proves the corollary. ■

The above corollary shows that you can determine whether two matrices are row equivalent by simply checking their row reduced echelon forms. The matrices are row equivalent if and only if they have the same row reduced echelon form.

Now with the above corollary, here is a very fundamental observation. It concerns a matrix which looks like this: (More columns than rows.)



Corollary 8.2.8 *Suppose A is an $m\times n$ matrix and that $m<n$. That is, the number of rows is less than the number of columns. Then one of the columns of A is a linear combination of the preceding columns of A. Also, there exists a nonzero solution* $\mathbf{x}$ *to the equation* $A\mathbf{x}=\mathbf{0}$.

Proof: Since $m<n$, not all the columns of A can be pivot columns. In reading from left to right, pick the first one which is not a pivot column. Then from the description of the row reduced echelon form, this column is a linear combination of the preceding columns. Denote the j^{th} column of A by $\mathbf{a}_j$. Thus for some $k>1$,

$$\mathbf{a}_k=\sum_{j=1}^{k-1}x_j\mathbf{a}_j,\text{ so }\sum_{j=1}^{k-1}x_j\mathbf{a}_j+(-1)\,\mathbf{a}_k=\mathbf{0}$$

Let $\mathbf{x}=(x_1,\cdots,x_{k-1},-1,0,\cdots,0)^T$. Then $A\mathbf{x}=\mathbf{0}$. ■

Example 8.2.9 *Find the row reduced echelon form of the matrix*

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

The first nonzero column is the second in the matrix. We switch the third and first rows to obtain

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

Now we multiply the top row by -2 and add to the second.

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

Next, add the second row to the bottom and then divide the bottom row by -6

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

Next use the bottom row to obtain zeros in the last column above the 1 and divide the second row by -2

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

Finally, add -1 times the middle row to the top.

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

This is in row reduced echelon form.

Example 8.2.10 *Find the row reduced echelon form for the matrix*

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

►►

You should verify that the row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & -\frac{8}{5} & 0 \\ 0 & 1 & \frac{4}{5} & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Having developed the row reduced echelon form, it is now easy to verify that the right inverse found earlier using the Gauss Jordan procedure is the inverse.

Theorem 8.2.11 *Suppose A, B are $n \times n$ matrices and $AB = I$. Then it follows that $BA = I$ also, and so $B = A^{-1}$. For $n \times n$ matrices, the left inverse, right inverse and inverse are all the same thing. Furthermore, if A^{-1} exists, then it can be found using the technique of row operations described earlier.*

Proof. If $AB = I$ for A, B $n \times n$ matrices, is $BA = I$? If $AB = I$, there exists at most one solution $\mathbf{x}$ to the equation

$$B\mathbf{x} = \mathbf{y}$$

for any choice of $\mathbf{y}$. In fact,

$$\mathbf{x} = A(B\mathbf{x}) = A\mathbf{y}.$$

This means the row reduced echelon form of B must be I. Thus every column is a pivot column. Otherwise, there exists a free variable and the solution, if it exists, would not be unique, contrary to what was just shown must happen if $AB = I$. It follows that a right inverse B^{-1} for B exists. The Gauss Jordan procedure for finding the inverse yields

$$\begin{pmatrix} B & I \end{pmatrix} \mapsto \begin{pmatrix} I & B^{-1} \end{pmatrix}.$$

Now multiply both sides of the equation $AB = I$ on the right by B^{-1}. Then

$$A = A\left(BB^{-1}\right) = (AB)B^{-1} = B^{-1}.$$

Thus A is the right inverse of B, and so $BA = I$. This shows that if $AB = I$, then $BA = I$ also. Exchanging roles of A and B, we see that if $BA = I$, then $AB = I$.

Now suppose A^{-1} exists. Then there exists a unique solution $\mathbf{x}$ to the system $A\mathbf{x} = \mathbf{b}$ given by $\mathbf{x} = A^{-1}\mathbf{b}$. It follows that each column of A is a pivot column. Hence one can row reduce and obtain

$$\begin{pmatrix} A & I \end{pmatrix} \rightarrow \begin{pmatrix} I & A^{-1} \end{pmatrix} \blacksquare$$

Note that this also shows that any invertible matrix can be written as a product of elementary matrices.

Proposition 8.2.12 *If A is an invertible matrix, then it is a product of elementary matrices.*

Proof: From the above, there exists a sequence of row operations which will reduce A to the identity matrix. Also, as explained above, each of these row operations may be obtained by multiplying on the left by an elementary matrix so $E_1E_2\cdots E_mA = I$ Now multiply on the left by the inverse of the elementary matrices in that product which are themselves elementary matrices by Theorem 8.1.10. Thus $E_2\cdots E_mA = E_1^{-1}, E_3\cdots E_mA = E_2^{-1}E_1^{-1}, \ldots, A = E_m^{-1}\cdots E_2^{-1}E_1^{-1}$, a product of elementary matrices. ■

Recall Theorem 8.1.10 which gives the description of the inverse of an elementary matrix.

Example 8.2.13 *Here is a matrix* $\begin{pmatrix} 1 & 1 & 3 & 1 \\ 1 & 2 & 5 & 3 \\ 1 & -1 & -1 & -3 \end{pmatrix}$. *Find a sequence of elementary matrices which, when multiplied on the left, will reduce the above matrix to row reduced echelon form.*

The following sequence of row operations leads to the row reduced echelon form

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

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

Thus the sequence of elementary matrices is

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

From the way we are multiplying on the left the matrices must be listed in this order.

Example 8.2.14 *Here is an invertible matrix.* $A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 3 & 0 \\ 0 & -1 & 2 \end{pmatrix}$. *Express A as a product of elementary matrices.*

Here are a sequence of row operations which row reduce A to the identity.

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

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

Thus the sequence of elementary matrices which produces the row reduced echelon form, in this case, the identity matrix is the following product multiplied on the left times A :

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

Thus, to obtain A, we replace each of these with its inverse and multiply the result in the opposite order to get A as indicated in Proposition 8.2.12. Thus we want

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

You can check that the product of these does equal A.

8.3 The Rank Of A Matrix

8.3.1 The Definition Of Rank

To begin, here is a definition to introduce some terminology.

Definition 8.3.1 *Let A be an $m \times n$ matrix. The* ***column space*** *of A is the span of the columns. The* ***row space*** *is the span of the rows.*

There are three definitions of the **rank** of a matrix which are useful. These are given in the following definition. It turns out that the concept of **determinant rank** is often important but is virtually impossible to find directly. The other two concepts of rank are very easily determined and it is a happy fact that all three yield the same number. This is shown later.

Definition 8.3.2 *A* ***sub-matrix*** *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 $r \times r$ sub-matrix of A has a non zero determinant. The* ***row space*** *of a matrix is the span of the rows and the* ***column space*** *of a matrix is the span of the columns. The* ***row rank*** *of a matrix is the number of nonzero rows in the row reduced echelon form and the* ***column rank*** *is the number columns in the row reduced echelon form which are one of the $\mathbf{e}_k$ vectors. Thus the column rank equals the number of pivot columns. It follows the row rank equals the column rank. This is also called the rank of the matrix. The rank of a matrix A is denoted by* $\text{rank}(A)$.

Example 8.3.3 *Consider the matrix*

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

What is its rank?

You could look at all the 2×2 submatrices

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

Each has determinant equal to 0. Therefore, the rank is less than 2. Now look at the 1×1 submatrices. There exists one of these which has nonzero determinant. For example (1) has determinant equal to 1 and so the rank of this matrix equals 1.

Of course this example was pretty easy but what if you had a 4×7 matrix? You would have to consider all the 4×4 submatrices and then all the 3×3 submatrices and then all the 2×2 matrices and finally all the 1×1 matrices in order to compute the rank. Clearly this is not practical. The following theorem will remove the difficulties just indicated.

The following theorem is proved later.

Theorem 8.3.4 *Let A be an $m \times n$ matrix. Then the row rank, column rank and determinant rank are all the same.*

Example 8.3.5 *Find the rank of the matrix*

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 0 \\ -4 & 3 & 2 & 1 & 2 \\ 3 & 2 & 1 & 6 & 5 \\ 4 & -3 & -2 & 1 & 7 \end{pmatrix}.$$

From the above definition, all you have to do is find the row reduced echelon form and then count up the number of nonzero rows. But the row reduced echelon form of this matrix is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & -\frac{17}{4} \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & -\frac{45}{4} \\ 0 & 0 & 0 & 1 & \frac{9}{2} \end{pmatrix}$$

and so the rank of this matrix is 4.

Find the rank of the matrix

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 0 \\ -4 & 3 & 2 & 1 & 2 \\ 3 & 2 & 1 & 6 & 5 \\ 0 & 7 & 4 & 10 & 7 \end{pmatrix}$$

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 0 & \frac{3}{2} & \frac{5}{2} \\ 0 & 1 & 0 & -4 & -17 \\ 0 & 0 & 1 & \frac{19}{2} & \frac{63}{2} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and so this time the rank is 3.

8.3.2 Finding The Row And Column Space Of A Matrix

The row reduced echelon form also can be used to obtain an efficient description of the row and column space of a matrix. Of course you can get the column space by simply saying that it equals the span of all the columns but often you can get the column space as the span of fewer columns than this. This is what we mean by an "efficient description". This is illustrated in the next example.

Example 8.3.6 *Find the rank of the following matrix and describe the column and row spaces efficiently.*

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 1 & 3 & 6 & 0 & 2 \\ 3 & 7 & 8 & 6 & 6 \end{pmatrix} \tag{8.1}$$

The row reduced echelon form is

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

Therefore, the rank of this matrix equals 2. All columns of this row reduced echelon form are in

$$\text{span}\left(\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\right).$$

For example,

$$\begin{pmatrix} -9 \\ 5 \\ 0 \end{pmatrix} = -9\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 5\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

By Lemma 8.2.5, all columns of the original matrix, are similarly contained in the span of the first two columns of that matrix. For example, consider the third column of the original matrix.

$$\begin{pmatrix} 1 \\ 6 \\ 8 \end{pmatrix} = -9\begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix} + 5\begin{pmatrix} 2 \\ 3 \\ 7 \end{pmatrix}.$$

How did I know to use -9 and 5 for the coefficients? This is what Lemma 8.2.5 says! It says linear relationships are all preserved. Therefore, the column space of the original matrix equals the span of the first two columns. This is the desired efficient description of the column space.

What about an efficient description of the row space? When row operations are used, the resulting vectors remain in the row space. Thus the rows in the row reduced echelon form are in the row space of the original matrix. Furthermore, by reversing the row operations, each row of the original matrix can be obtained as a linear combination of the rows in the row reduced echelon form. It follows that the span of the nonzero rows in the row reduced echelon matrix equals the span of the original rows. In the above example, the row space equals the span of the two vectors, $\begin{pmatrix} 1 & 0 & -9 & 9 & 2 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 & 5 & -3 & 0 \end{pmatrix}$.

Example 8.3.7 *Find the rank of the following matrix and describe the column and row spaces efficiently.*

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 2 \\ 1 & 3 & 6 & 0 & 2 \\ 1 & 2 & 1 & 3 & 2 \\ 1 & 3 & 2 & 4 & 0 \end{pmatrix} \tag{8.2}$$

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \frac{13}{2} \\ 0 & 1 & 0 & 2 & -\frac{5}{2} \\ 0 & 0 & 1 & -1 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

and so the rank is 3, the row space is the span of the vectors,

$$\begin{pmatrix} 0 & 0 & 1 & -1 & \frac{1}{2} \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 & 2 & -\frac{5}{2} \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 & \frac{13}{2} \end{pmatrix},$$

and the column space is the span of the first three columns in the **original matrix**,

$$\text{span}\left(\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 6 \\ 1 \\ 2 \end{pmatrix}\right).$$

Example 8.3.8 *Find the rank of the following matrix and describe the column and row spaces efficiently.*

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

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 1 & 0 & \frac{21}{17} \\ 0 & 1 & 1 & 0 & -\frac{2}{17} \\ 0 & 0 & 0 & 1 & \frac{14}{17} \end{pmatrix}.$$

It follows the rank is three and the column space is the span of the first, second and fourth columns of the **original matrix.**

$$\text{span}\left(\begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 3 \end{pmatrix}\right)$$

while the row space is the span of the vectors

$$\begin{pmatrix} 0 & 0 & 0 & 1 & \frac{14}{17} \end{pmatrix}, \begin{pmatrix} 0 & 1 & 1 & 0 & -\frac{2}{17} \end{pmatrix}, \begin{pmatrix} 1 & 0 & 1 & 0 & \frac{21}{17} \end{pmatrix}.$$

Procedure 8.3.9 *To find the rank of a matrix, obtain the row reduced echelon form for the matrix. Then count the number of nonzero rows or equivalently the number of pivot columns. This is the rank. The row space is the span of the nonzero rows in the row reduced echelon form and the column space is the span of the pivot columns of the **original matrix**.*

8.4 A Short Application To Chemistry

The following example is to chemistry. Sometimes there are numerous chemical reactions and some are in a sense redundant. This example is discussed in the book by Greenberg [7]. Suppose you have the folowing chemical reactions.

$$\begin{gathered} CO+\frac{1}{2}O_2 \to CO_2 \\ H_2+\frac{1}{2}O_2 \to H_2O \\ CH_4+\frac{3}{2}O_2 \to CO+2H_2O \\ CH_4+2O_2 \to CO_2+2H_2O \end{gathered}$$

There are four chemical reactions here but they are not independent reactions. There is some redundancy. What are the independent reactions? Is there a way to consider a shorter list of reactions? To analyze this situation, you can write as a matrix

$$\begin{pmatrix} CO & O_2 & CO_2 & H_2 & H_2O & CH_4 \\ 1 & 1/2 & -1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1 & -1 & 0 \\ -1 & 3/2 & 0 & 0 & -2 & 1 \\ 0 & 2 & -1 & 0 & -2 & 1 \end{pmatrix}$$

The top row of numbers comes from $CO+\frac{1}{2}O_2-CO_2=0$ which represents the first of the chemical reactions. The entries of the matrix

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

are called Stoichiometric coefficients. Rather than listing all of the reactions as above, it would be more efficient to only list those which are in a sense independent by throwing out that which is redundant. This is easy to do. Just take the row reduced echelon from of the above matrix.

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

The top three rows represent "independent" reactions which come from the original four reactions. One can obtain each of the original four rows of the stoichiometric matrix given above by taking a suitable linear combination of rows of this row reduced matrix.

Thus one could consider the simplified reactions

$$\begin{gathered} CO+3H_2-1H_2O-1CH_4=0 \\ O_2+2H_2-2H_2O=0 \\ CO_2+4H_2-2H_2O-1CH_4=0 \end{gathered}$$

In terms of the original notation, these are the reactions

$$\begin{gathered} CO+3H_2 \rightarrow H_2O+CH_4 \\ O_2+2H_2 \rightarrow 2H_2O \\ CO_2+4H_2 \rightarrow 2H_2O+CH_4 \end{gathered}$$

Instead of the four you started with, you could consider the simpler list given above. The idea is that, in terms of what happens chemically, you obtain the same information with the shorter list of reactions and have gotten rid of the redundancy which was present in the original list. You can probably imagine that if you had a very large list of reactions made up from some sort of experimental evidence, such a simplification could be a considerable improvement.

This is motivation for the general notion of a basis for a vector space which is discussed in the next section. The idea of a basis is similar to what was just done, reducing a list of reactions to a shorter list. With vectors, you have the span of some vectors and you want to get the shortest possible list of vectors which will leave the span unchanged.

8.5 Linear Independence And Bases

8.5.1 Linear Independence And Dependence

First we consider the concept of linear independence. We define what it means for vectors in $\mathbb{F}^n$ to be linearly independent and then give equivalent descriptions. In the following definition, the symbol,

$$\left(\begin{array}{cccc} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{array}\right)$$

denotes the matrix which has the vector $\mathbf{v}_1$ as the first column, $\mathbf{v}_2$ as the second column and so forth until $\mathbf{v}_k$ is the k^{th} column.

Definition 8.5.1 *Let $\{\mathbf{v}_1,\cdots,\mathbf{v}_k\}$ be vectors in $\mathbb{F}^n$. Then this collection of vectors is said to be* ***linearly independent*** *if each of the columns of the $n\times k$ matrix*

$$\left(\begin{array}{cccc} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{array}\right)$$

is a pivot column. Thus the row reduced echelon form for this matrix is

$$\left(\begin{array}{cccc} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_k \end{array}\right)$$

and you cannot delete any of these vectors without diminishing the span of the resulting list.

The question whether any vector in the first k columns in a matrix is a pivot column is independent of the presence of later columns. Thus each of $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is a pivot column in

$$\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{pmatrix}$$

if and only if these vectors are each pivot columns in

$$\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k & \mathbf{w}_1 & \cdots & \mathbf{w}_r \end{pmatrix}$$

Here is what the linear independence means in terms of linear relationships.

Corollary 8.5.2 *The collection of vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is linearly independent if and only if none of these vectors is a linear combination of the others.*

Proof: If $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is linearly independent, then every column in

$$\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{pmatrix}$$

is a pivot column which requires that the row reduced echelon form is

$$\begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_k \end{pmatrix}.$$

Now none of the $\mathbf{e}_i$ vectors is a linear combination of the others. By Lemma 8.2.5 on Page 152 none of the $\mathbf{v}_i$ is a linear combination of the others. Recall this lemma says linear relationships between the columns are preserved under row operations.

Next suppose none of the vectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is a linear combination of the others. Then none of the columns in

$$\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{pmatrix}$$

is a linear combination of the others. By Lemma 8.2.5 the same is true of the row reduced echelon form for this matrix. From the description of the row reduced echelon form, it follows that the i^{th} column of the row reduced echelon form must be $\mathbf{e}_i$ since otherwise, it would be a linear combination of the first $i-1$ vectors $\mathbf{e}_1, \cdots, \mathbf{e}_{i-1}$ and by Lemma 8.2.5, it follows $\mathbf{v}_i$ would be the same linear combination of $\mathbf{v}_1, \cdots, \mathbf{v}_{i-1}$ contrary to the assumption that none of the columns in $\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{pmatrix}$ is a linear combination of the others. Therefore, each of the k columns in $\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{pmatrix}$ is a pivot column and so $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is linearly independent. ■

Corollary 8.5.3 *The collection of vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is linearly independent if and only if whenever*

$$\sum_{i=1}^{n} c_i \mathbf{v}_i = \mathbf{0}$$

it follows each $c_i = 0$.

Proof: Suppose first $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is linearly independent. Then by Corollary 8.5.2, none of the vectors is a linear combination of the others. Now suppose

$$\sum_{i=1}^{n} c_i \mathbf{v}_i = \mathbf{0}$$

and not all the $c_i = 0$. Then pick c_i which is not zero, divide by it and solve for $\mathbf{v}_i$ in terms of the other $\mathbf{v}_j$, contradicting the fact that none of the $\mathbf{v}_i$ equals a linear combination of the others.

Now suppose the condition about the sum holds. If $\mathbf{v}_i$ is a linear combination of the other vectors in the list, then you could obtain an equation of the form

$$\mathbf{v}_i = \sum_{j \neq i} c_j \mathbf{v}_j$$

and so

$$\mathbf{0} = \sum_{j \neq i} c_j \mathbf{v}_j + (-1)\mathbf{v}_i,$$

contradicting the condition about the sum. ■

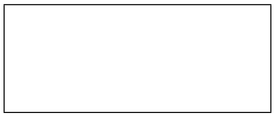
Sometimes we refer to this last condition about sums as follows: The set of vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is linearly independent if and only if there is no nontrivial linear combination which equals zero. (A nontrivial linear combination is one in which not all the scalars equal zero.)

We give the following equivalent definition of linear independence which follows from the above corollaries.

Definition 8.5.4 *A set of vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is linearly independent if and only if none of the vectors is a linear combination of the others or equivalently if there is no nontrivial linear combination of the vectors which equals 0. It is said to be **linearly dependent** if at least one of the vectors **is** a linear combination of the others or equivalently there exists a nontrivial linear combination which equals zero.*

Note the meaning of the words. To say a set of vectors is linearly dependent means at least one is a linear combination of the others. In other words, it is in a sense "dependent" on these other vectors. At this time, the vectors are in $\mathbb{F}^n$ but the above definition makes sense without knowing any description of the vectors. This will be considered later in the book.

The following corollary follows right away from the row reduced echelon form. It concerns a matrix which looks like this: (More columns than rows.)



Corollary 8.5.5 *Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ be a set of vectors in $\mathbb{F}^n$. Then if $k > n$, it must be the case that $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is not linearly independent. In other words, if $k > n$, then $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is dependent.*

Proof: If $k > n$, then the columns of $\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{pmatrix}$ cannot each be a pivot column because there are at most n pivot columns due to the fact the matrix has only n rows. In reading from left to right, pick the first column which is not a pivot column. Then from the description of row reduced echelon form, this column is a linear combination of the preceding columns and so the given vectors are dependent by Corollary 8.5.2. ■

Example 8.5.6 *Are the vectors* $\left\{\begin{pmatrix}1\\2\\3\\0\end{pmatrix}\begin{pmatrix}2\\1\\0\\1\end{pmatrix}\begin{pmatrix}0\\1\\1\\2\end{pmatrix}\begin{pmatrix}3\\2\\2\\-1\end{pmatrix}\right\}$ *linearly independent? If they are linearly dependent, exhibit one of the vectors as a linear combination of the others.*

Form the matrix mentioned above.

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

Then the row reduced echelon form of this matrix is

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

Thus not all the columns are pivot columns and so the vectors are not linear independent. Note the fourth column is of the form

$$1\begin{pmatrix}1\\0\\0\\0\end{pmatrix}+1\begin{pmatrix}0\\1\\0\\0\end{pmatrix}+(-1)\begin{pmatrix}0\\0\\1\\0\end{pmatrix}$$

From Lemma 8.2.5, the same linear relationship exists between the columns of the original matrix. Thus

$$1\begin{pmatrix}1\\2\\3\\0\end{pmatrix}+1\begin{pmatrix}2\\1\\0\\1\end{pmatrix}+(-1)\begin{pmatrix}0\\1\\1\\2\end{pmatrix}=\begin{pmatrix}3\\2\\2\\-1\end{pmatrix}.$$

Note the usefulness of the row reduced echelon form in discovering hidden linear relationships in collections of vectors.

Example 8.5.7 *Determine whether the vectors* $\left\{\begin{pmatrix}1\\2\\3\\0\end{pmatrix}\begin{pmatrix}2\\1\\0\\1\end{pmatrix}\begin{pmatrix}0\\1\\1\\2\end{pmatrix}\begin{pmatrix}3\\2\\2\\0\end{pmatrix}\right\}$ *are linearly independent. If they are linearly dependent, exhibit one of the vectors as a linear combination of the others.*

The matrix used to find this is

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

The row reduced echelon form is

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

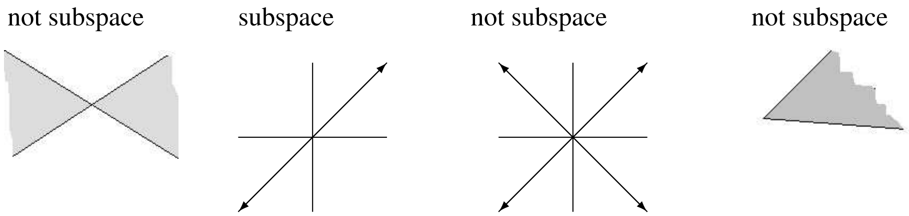
and so every column is a pivot column. Therefore, these vectors are linearly independent and there is no way to obtain one of the vectors as a linear combination of the others.

8.5.2 Subspaces

A subspace is a set of vectors with the property that linear combinations of these vectors remain in the set. Geometrically, subspaces are like lines and planes which contain the origin. More precisely, the following definition is the right way to think of this.

Definition 8.5.8 *Let V be a nonempty collection of vectors in $\mathbb{F}^n$. Then V is called a subspace if whenever α, β are scalars and $\mathbf{u}, \mathbf{v}$ are vectors in V, the linear combination $\alpha \mathbf{u} + \beta \mathbf{v}$ is also in V.*

There is no substitute for the above definition or equivalent algebraic definition! However, it is sometimes helpful to look at pictures at least initially. The following are four subsets of $\mathbb{R}^2$. The first is the shaded area between two lines which intersect at the origin, the second is a line through the origin, the third is the union of two lines through the origin, and the last is the region between two rays from the origin. Note that in the last, multiplication of a vector in the set by a nonnegative scalar results in a vector in the set as does the sum of two vectors in the set. However, multiplication by a negative scalar does not take a vector in the set to another in the set.



Observe how the above definition indicates that the claims posted on the picture are valid.

Subspaces are exactly those subsets of $\mathbb{F}^n$ which are themselves vector spaces. Recall that a vector space is something which satisfies the vector space axioms on Page 19.

Proposition 8.5.9 *Let V be a nonempty collection of vectors in $\mathbb{F}^n$. Then V is a subspace if and only if V is itself a vector space having the same operations as those defined on $\mathbb{F}^n$.*

Proof: Suppose first that V is a subspace. It is obvious all the algebraic laws hold on V because it is a subset of $\mathbb{F}^n$ and they hold on $\mathbb{F}^n$. Thus $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$ along with the other axioms. Does V contain $\mathbf{0}$? Yes because it contains $0\mathbf{u}=\mathbf{0}$. Are the operations defined on V? That is, when you add vectors of V do you get a vector in V? When you multiply a vector in V by a scalar, do you get a vector in V? Yes. This is contained in the definition. Does every vector in V have an additive inverse? Yes because $-\mathbf{v}=(-1)\mathbf{v}$ which is given to be in V provided $\mathbf{v}\in V$. ($(-1)\mathbf{v}+\mathbf{v}=((-1)+1)\mathbf{v}=0\mathbf{v}=\mathbf{0}$.)

Next suppose V is a vector space. Then by definition, it is closed with respect to linear combinations. Hence it is a subspace. ■

It turns out that every subspace equals the span of some vectors. This is the content of the next theorem.

Theorem 8.5.10 *V is a subspace of $\mathbb{F}^n$ if and only if there exist vectors of V*

$$\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$$

such that $V=\operatorname{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$.

Proof: Pick a vector of $V,\mathbf{u}_1$. If $V=\operatorname{span}\{\mathbf{u}_1\}$, then stop. You have found your list of vectors. If $V\neq\operatorname{span}(\mathbf{u}_1)$, then there exists $\mathbf{u}_2$ a vector of V which is not a vector in $\operatorname{span}(\mathbf{u}_1)$. Consider $\operatorname{span}(\mathbf{u}_1,\mathbf{u}_2)$. If $V=\operatorname{span}(\mathbf{u}_1,\mathbf{u}_2)$, stop. Otherwise, pick $\mathbf{u}_3\notin\operatorname{span}(\mathbf{u}_1,\mathbf{u}_2)$. Continue this way. Note that since V is a subspace, these spans are each contained in V. The process must stop with $\mathbf{u}_k$ for some $k\leq n$ since otherwise, the matrix

$$\left(\begin{array}{ccc}\mathbf{u}_1 & \cdots & \mathbf{u}_k\end{array}\right)$$

having these vectors as columns would have n rows and $k>n$ columns. Consequently, it can have no more than n pivot columns and so the first column which is not a pivot column would be a linear combination of the preceding columns contrary to the construction.

Conversely, suppose $V=\operatorname{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$ and let $\sum_{i=1}^k c_i\mathbf{u}_i$ and $\sum_{i=1}^k d_i\mathbf{u}_i$ be two vectors in V. Now let α and β be two scalars. Then

$$\alpha\sum_{i=1}^k c_i\mathbf{u}_i+\beta\sum_{i=1}^k d_i\mathbf{u}_i=\sum_{i=1}^k(\alpha c_i+\beta d_i)\mathbf{u}_i$$

which is one of the things in $\operatorname{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$ showing that $\operatorname{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$ has the properties of a subspace. ■

The following corollary also follows easily.

Corollary 8.5.11 *For V a subspace of $\mathbb{F}^n$, there exist vectors of $V,\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ such that $V=\operatorname{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$ and $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ is linearly independent.*

Proof: Let $V=\operatorname{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$. Then let the vectors $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ be the columns of the following matrix.

$$\left(\begin{array}{ccc}\mathbf{u}_1 & \cdots & \mathbf{u}_k\end{array}\right)$$

Retain only the pivot columns. That is, determine the pivot columns from the row reduced echelon form and these are a basis for $\operatorname{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$. ■

The message is that subspaces of $\mathbb{F}^n$ consist of spans of finite, linearly independent collections of vectors of $\mathbb{F}^n$.

The following fundamental lemma is very useful. It is called the exchange theorem.

Lemma 8.5.12 *Suppose $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is linearly independent and each $\mathbf{x}_k$ is contained in*

$$\operatorname{span}(\mathbf{y}_1, \cdots, \mathbf{y}_s).$$

Then $s \geq r$. In words, spanning sets have at least as many vectors as linearly independent sets.

Proof: Since $\{\mathbf{y}_1, \cdots, \mathbf{y}_s\}$ is a spanning set, there exist scalars a_{ij} such that

$$\mathbf{x}_j = \sum_{i=1}^{s} a_{ij}\mathbf{y}_i$$

Suppose $s < r$. Then the matrix A whose ij^{th} entry is a_{ij} has fewer rows, s than columns, r. By Corollary 8.2.8 there exists $\mathbf{d}$ such that $\mathbf{d} \neq \mathbf{0}$ but $A\mathbf{d} = \mathbf{0}$. In other words,

$$\sum_{j=1}^{r} a_{ij}d_j = 0, \; i = 1, 2, \cdots, s$$

Therefore,

$$\begin{aligned}\sum_{j=1}^{r} d_j\mathbf{x}_j &= \sum_{j=1}^{r} d_j \sum_{i=1}^{s} a_{ij}\mathbf{y}_i \\ &= \sum_{i=1}^{s}\left(\sum_{j=1}^{r} a_{ij}d_j\right)\mathbf{y}_i = \sum_{i=1}^{s} 0\mathbf{y}_i = 0\end{aligned}$$

which contradicts $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is linearly independent, because not all the $d_j = 0$. Thus $s \geq r$. ■

Note how this lemma was totally dependent on algebraic considerations and was independent of context. This will be considered more later in the chapter on abstract vector spaces. I didn't need to know what the $\mathbf{x}_k$, $\mathbf{y}_k$ were, only that the $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ were independent and contained in the span of the $\mathbf{y}_k$.

8.5.3 Basis Of A Subspace

It was just shown in Corollary 8.5.11 that every subspace of $\mathbb{F}^n$ is equal to the span of a linearly independent collection of vectors of $\mathbb{F}^n$. Such a collection of vectors is called a basis.

Definition 8.5.13 *Let V be a subspace of $\mathbb{F}^n$. Then $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ is a* ***basis*** *for V if the following two conditions hold.*

1. $\operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k) = V$.
2. $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ *is linearly independent.*

The plural of basis is ***bases****.*

The main theorem about bases is the following.

Theorem 8.5.14 *For V be a subspace of $\mathbb{F}^n$ and $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$, $\{\mathbf{v}_1,\cdots,\mathbf{v}_m\}$ two bases for V, it follows $k=m$.*

Proof: This follows right away from Lemma 8.5.12. $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ is a spanning set while $\{\mathbf{v}_1,\cdots,\mathbf{v}_m\}$ is linearly independent so $k\geq m$. Also $\{\mathbf{v}_1,\cdots,\mathbf{v}_m\}$ is a spanning set while $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ is linearly independent so $m\geq k$.

Now here is another proof. Suppose $k<m$. Then since $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ is a basis for V, each $\mathbf{v}_i$ is a linear combination of the vectors of $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$. Consider the matrix

$$\left(\begin{array}{cccccc}\mathbf{u}_1 & \cdots & \mathbf{u}_k & \mathbf{v}_1 & \cdots & \mathbf{v}_m\end{array}\right)$$

in which each of the $\mathbf{u}_i$ is a pivot column because the $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ are linearly independent. Therefore, the row reduced echelon form of this matrix is

$$\left(\begin{array}{cccccc}\mathbf{e}_1 & \cdots & \mathbf{e}_k & \mathbf{w}_1 & \cdots & \mathbf{w}_m\end{array}\right) \tag{8.3}$$

where each $\mathbf{w}_j$ has zeroes below the k^{th} row. This is because of Lemma 8.2.5 which implies each $\mathbf{w}_i$ is a linear combination of the $\mathbf{e}_1,\cdots,\mathbf{e}_k$. Discarding the bottom $n-k$ rows of zeroes in the above, yields the matrix

$$\left(\begin{array}{cccccc}\mathbf{e}_1' & \cdots & \mathbf{e}_k' & \mathbf{w}_1' & \cdots & \mathbf{w}_m'\end{array}\right)$$

in which all vectors are in $\mathbb{F}^k$. Since $m>k$, it follows from Corollary 8.5.5 that the vectors, $\{\mathbf{w}_1',\cdots,\mathbf{w}_m'\}$ are dependent. Therefore, some $\mathbf{w}_j'$ is a linear combination of the other $\mathbf{w}_i'$. Therefore, $\mathbf{w}_j$ is a linear combination of the other $\mathbf{w}_i$ in 8.3. By Lemma 8.2.5 again, the same linear relationship exists between the $\{\mathbf{v}_1,\cdots,\mathbf{v}_m\}$ showing that $\{\mathbf{v}_1,\cdots,\mathbf{v}_m\}$ is not linearly independent and contradicting the assumption that $\{\mathbf{v}_1,\cdots,\mathbf{v}_m\}$ is a basis. It follows $m\leq k$. Similarly, $k\leq m$. ■

This is a very important theorem so here is yet another proof of it.

Theorem 8.5.15 *Let V be a subspace and suppose $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ and $\{\mathbf{v}_1,\cdots,\mathbf{v}_m\}$ are two bases for V. Then $k=m$.*

Proof: Suppose $k>m$. Then since the vectors, $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ span V, there exist scalars, c_{ij} such that

$$\sum_{i=1}^{m} c_{ij}\mathbf{v}_i=\mathbf{u}_j.$$

Therefore,

$$\sum_{j=1}^{k} d_j\mathbf{u}_j=\mathbf{0} \text{ if and only if } \sum_{j=1}^{k}\sum_{i=1}^{m} c_{ij}d_j\mathbf{v}_i=\mathbf{0}$$

if and only if

$$\sum_{i=1}^{m}\left(\sum_{j=1}^{k} c_{ij}d_j\right)\mathbf{v}_i=\mathbf{0}$$

Now since$\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is independent, this happens if and only if

$$\sum_{j=1}^{k} c_{ij} d_j = 0,\ i = 1,2,\cdots,m.$$

However, this is a system of m equations in k variables, $d_1,\cdots,d_k$ and $m < k$. Therefore, there exists a solution to this system of equations in which not all the d_j are equal to zero. Recall why this is so. The augmented matrix for the system is of the form $\begin{pmatrix} C & \mathbf{0} \end{pmatrix}$ where C is a matrix which has more columns than rows. Therefore, there are free variables and hence nonzero solutions to the system of equations. However, this contradicts the linear independence of $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ because, as explained above, $\sum_{j=1}^{k} d_j \mathbf{u}_j = \mathbf{0}$. Similarly it cannot happen that $m > k$. ■

The following definition can now be stated.

Definition 8.5.16 *Let V be a subspace of $\mathbb{F}^n$. Then the **dimension** of V is defined to be the number of vectors in a basis.*

Corollary 8.5.17 *The dimension of $\mathbb{F}^n$ is n. The dimension of the space of $m \times n$ matrices is mn.*

Proof: You only need to exhibit a basis for $\mathbb{F}^n$ which has n vectors. Such a basis is $\{\mathbf{e}_1,\cdots,\mathbf{e}_n\}$. As to the vector space of $m \times n$ matrices, a basis consists of the matrices E_{ij} which has a 1 in the ij^{th} position and 0 elsewhere. ■

Corollary 8.5.18 *Suppose $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is linearly independent and each $\mathbf{v}_i$ is a vector in $\mathbb{F}^n$. Then $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is a basis for $\mathbb{F}^n$. Suppose $\{\mathbf{v}_1,\cdots,\mathbf{v}_m\}$ spans $\mathbb{F}^n$. Then $m \geq n$. If $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ spans $\mathbb{F}^n$, then $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is linearly independent.*

Proof: Let $\mathbf{u}$ be a vector of $\mathbb{F}^n$ and consider the matrix

$$\begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n & \mathbf{u} \end{pmatrix}.$$

Since each $\mathbf{v}_i$ is a pivot column, the row reduced echelon form is

$$\begin{pmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_n & \mathbf{w} \end{pmatrix}$$

and so, since $\mathbf{w}$ is in $\operatorname{span}(\mathbf{e}_1,\cdots,\mathbf{e}_n)$, it follows from Lemma 8.2.5 that $\mathbf{u}$ is one of the vectors in $\operatorname{span}(\mathbf{v}_1,\cdots,\mathbf{v}_n)$. Therefore, $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is a basis as claimed.

To establish the second claim, suppose that $m < n$. Then letting $\mathbf{v}_{i_1},\cdots,\mathbf{v}_{i_k}$ be the pivot columns of the matrix

$$\begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_m \end{pmatrix}$$

it follows $k \leq m < n$ and these k pivot columns would be a basis for $\mathbb{F}^n$ having fewer than n vectors, contrary to Theorem 8.5.14 which states every two bases have the same number of vectors in them.

Finally consider the third claim. If $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is not linearly independent, then replace this list with $\{\mathbf{v}_{i_1},\cdots,\mathbf{v}_{i_k}\}$ where these are the pivot columns of the matrix

$$\begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}$$

Then $\{\mathbf{v}_{i_1},\cdots,\mathbf{v}_{i_k}\}$ spans $\mathbb{F}^n$ and is linearly independent so it is a basis having less than n vectors contrary to Theorem 8.5.14 which states every two bases have the same number of vectors in them. ■

Example 8.5.19 *Find the rank of the following matrix. If the rank is r, identify r columns* ***in the original matrix*** *which have the property that every other column may be written as a linear combination of these. Also find a basis for the row and column spaces of the matrices.*

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

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 0 & \frac{27}{70} \\ 0 & 1 & 0 & \frac{1}{10} \\ 0 & 0 & 1 & \frac{33}{70} \end{pmatrix}$$

and so the rank of the matrix is 3. A basis for the column space is the first three columns of the **original matrix.** I know they span because the first three columns of the row reduced echelon form above span the column space of that matrix. They are linearly independent because the first three columns of the row reduced echelon form are linearly independent. By Lemma 8.2.5 all linear relationships are preserved and so these first three vectors form a basis for the column space. The four rows of the row reduced echelon form form a basis for the row space of the original matrix.

Example 8.5.20 *Find the rank of the following matrix. If the rank is r, identify r columns* ***in the original matrix*** *which have the property that every other column may be written as a linear combination of these. Also find a basis for the row and column spaces of the matrices.*

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

The row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & 1 & 0 & -\frac{1}{7} \\ 0 & 1 & 1 & 0 & \frac{4}{7} \\ 0 & 0 & 0 & 1 & -\frac{11}{42} \end{pmatrix}.$$

A basis for the column space of this row reduced echelon form is the first second and fourth columns. Therefore, a basis for the column space in the **original matrix** is the first second and fourth columns. The rank of the matrix is 3. A basis for the row space of the original matrix is the columns of the row reduced echelon form.

8.5.4 Extending An Independent Set To Form A Basis

Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ is a linearly independent set of vectors in $\mathbb{F}^n$. It turns out there is a larger set of vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_m, \mathbf{v}_{m+1}, \cdots, \mathbf{v}_n\}$ which is a basis for $\mathbb{F}^n$. It is easy to do this using the row reduced echelon form. Consider the following matrix having rank n in which the columns are shown.

$$\begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_m & \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \end{pmatrix}.$$

Since the $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ are linearly independent, the row reduced echelon form of this matrix is of the form

$$\begin{pmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_m & \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{pmatrix}$$

Now the pivot columns can be identified and this leads to a basis for the column space of the original matrix which is of the form

$$\left\{\mathbf{v}_1, \cdots, \mathbf{v}_m, \mathbf{e}_{i_1}, \cdots, \mathbf{e}_{i_{n-m}}\right\}.$$

This proves the following theorem.

Theorem 8.5.21 *Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_m\}$ be a linearly independent set of vectors in $\mathbb{F}^n$. Then there is a larger set of vectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_m, \mathbf{v}_{m+1}, \cdots, \mathbf{v}_n\}$ which is a basis for $\mathbb{F}^n$.*

Example 8.5.22 *The vectors,* $\left\{\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}\right\}$ *are linearly independent. Enlarge this set of vectors to form a basis for $\mathbb{R}^4$.*

Using the above technique, consider the following matrix.

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

whose row reduced echelon form is

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

The pivot columns are numbers 1,2,3, and 6. Therefore, a basis is

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

8.5.5 Finding The Null Space Or Kernel Of A Matrix

Let A be an $m \times n$ matrix.

Definition 8.5.23 $\ker(A)$*, also referred to as the null space of A is defined as follows.*

$$\ker(A) = \{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$$

and to find $\ker(A)$ *one must solve the system of equations $A\mathbf{x} = \mathbf{0}$.*

This is not new! There is just some new terminology being used. To repeat, $\ker(A)$ is the solution to the system $A\mathbf{x} = \mathbf{0}$.

Example 8.5.24 *Let*

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

Find $\ker(A)$.

You need to solve the equation $A\mathbf{x} = \mathbf{0}$. To do this you write the augmented matrix and then obtain the row reduced echelon form and the solution. The augmented matrix is

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

Next place this matrix in row reduced echelon form,

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

Note that x_1 and x_2 are basic variables while x_3 is a free variable. Therefore, the solution to this system of equations, $A\mathbf{x} = \mathbf{0}$ is given by

$$\begin{pmatrix} 3t \\ t \\ t \end{pmatrix} : t \in \mathbb{R}.$$

Example 8.5.25 *Let*

$$A = \begin{pmatrix} 1 & 2 & 1 & 0 & 1 \\ 2 & -1 & 1 & 3 & 0 \\ 3 & 1 & 2 & 3 & 1 \\ 4 & -2 & 2 & 6 & 0 \end{pmatrix}$$

Find the null space of A.

You need to solve the equation, $A\mathbf{x} = \mathbf{0}$. The augmented matrix is

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

Its row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & \frac{3}{5} & \frac{6}{5} & \frac{1}{5} & | & 0 \\ 0 & 1 & \frac{1}{5} & -\frac{3}{5} & \frac{2}{5} & | & 0 \\ 0 & 0 & 0 & 0 & 0 & | & 0 \\ 0 & 0 & 0 & 0 & 0 & | & 0 \end{pmatrix}$$

It follows x_1 and x_2 are basic variables and x_3, x_4, x_5 are free variables. Therefore, $\ker(A)$ is given by

$$\left(\begin{array}{c} \left(-\frac{3}{5}\right)s_1 + \left(\frac{-6}{5}\right)s_2 + \left(\frac{1}{5}\right)s_3 \\ \left(-\frac{1}{5}\right)s_1 + \left(\frac{3}{5}\right)s_2 + \left(-\frac{2}{5}\right)s_3 \\ s_1 \\ s_2 \\ s_3 \end{array} \right) : s_1, s_2, s_3 \in \mathbb{R}.$$

We write this in the form

$$s_1 \left(\begin{array}{c} -\frac{3}{5} \\ -\frac{1}{5} \\ 1 \\ 0 \\ 0 \end{array} \right) + s_2 \left(\begin{array}{c} \frac{-6}{5} \\ \frac{3}{5} \\ 0 \\ 1 \\ 0 \end{array} \right) + s_3 \left(\begin{array}{c} \frac{1}{5} \\ -\frac{2}{5} \\ 0 \\ 0 \\ 1 \end{array} \right) : s_1, s_2, s_3 \in \mathbb{R}.$$

In other words, the null space of this matrix equals the span of the three vectors above. Thus

$$\ker(A) = \operatorname{span}\left(\left(\begin{array}{c} -\frac{3}{5} \\ -\frac{1}{5} \\ 1 \\ 0 \\ 0 \end{array} \right), \left(\begin{array}{c} \frac{-6}{5} \\ \frac{3}{5} \\ 0 \\ 1 \\ 0 \end{array} \right), \left(\begin{array}{c} \frac{1}{5} \\ -\frac{2}{5} \\ 0 \\ 0 \\ 1 \end{array} \right) \right).$$

This is the same as

$$\ker(A) = \operatorname{span}\left(\left(\begin{array}{c} \frac{3}{5} \\ \frac{1}{5} \\ -1 \\ 0 \\ 0 \end{array} \right), \left(\begin{array}{c} \frac{6}{5} \\ \frac{-3}{5} \\ 0 \\ -1 \\ 0 \end{array} \right), \left(\begin{array}{c} \frac{-1}{5} \\ \frac{2}{5} \\ 0 \\ 0 \\ -1 \end{array} \right) \right).$$

Notice also that the three vectors above are linearly independent and so the dimension of $\ker(A)$ is 3. This is generally the way it works. The number of free variables equals the dimension of the null space while the number of basic variables equals the number of pivot columns which equals the rank. We state this in the following theorem.

Definition 8.5.26 *The dimension of the null space of a matrix is called the* ***nullity***[3] *and written as* $\text{null}(A)$.

Theorem 8.5.27 *Let A be an $m \times n$ matrix. Then* $\text{rank}(A) + \text{null}(A) = n$.

8.5.6 Rank And Existence Of Solutions To Linear Systems

Consider the linear system of equations,

$$A\mathbf{x} = \mathbf{b} \tag{8.4}$$

[3] Isn't it amazing how many different words are available for use in linear algebra?

where A is an $m\times n$ matrix, $\mathbf{x}$ is a $n\times 1$ column vector, and $\mathbf{b}$ is an $m\times 1$ column vector. Suppose

$$A=\left(\begin{array}{ccc}\mathbf{a}_1 & \cdots & \mathbf{a}_n\end{array}\right)$$

where the $\mathbf{a}_k$ denote the columns of A. Then $\mathbf{x}=(x_1,\cdots,x_n)^T$ is a solution of the system 8.4, if and only if

$$x_1\mathbf{a}_1+\cdots+x_n\mathbf{a}_n=\mathbf{b}$$

which says that $\mathbf{b}$ is a vector in $\operatorname{span}(\mathbf{a}_1,\cdots,\mathbf{a}_n)$. This shows that there exists a solution to the system, 8.4 if and only if $\mathbf{b}$ is contained in $\operatorname{span}(\mathbf{a}_1,\cdots,\mathbf{a}_n)$. In words, there is a solution to 8.4 if and only if $\mathbf{b}$ is in the column space of A. In terms of rank, the following proposition describes the situation.

Proposition 8.5.28 *Let A be an $m\times n$ matrix and let $\mathbf{b}$ be an $m\times 1$ column vector. Then there exists a solution to 8.4 if and only if*

$$\operatorname{rank}\left(\begin{array}{ccc}A & | & \mathbf{b}\end{array}\right)=\operatorname{rank}(A). \tag{8.5}$$

Proof: Place $\left(\begin{array}{ccc}A & | & \mathbf{b}\end{array}\right)$ and A in row reduced echelon form, respectively B and C. If the above condition on rank is true, then both B and C have the same number of nonzero rows. In particular, you cannot have a row of the form

$$\left(\begin{array}{cccc}0 & \cdots & 0 & \blacksquare\end{array}\right)$$

where $\blacksquare\neq 0$ in B. Therefore, there will exist a solution to the system 8.4.

Conversely, suppose there exists a solution. This means there cannot be such a row in B described above. Therefore, B and C must have the same number of zero rows and so they have the same number of nonzero rows. Therefore, the rank of the two matrices in 8.5 is the same. ■

8.6 Fredholm Alternative

Before reading this, it would be helpful to read about the dot product earlier in Section 3.1. However, what you need to know is this: For $\mathbf{x}\in\mathbb{R}^p$ meaning

$$\mathbf{x}=\left(\begin{array}{cccc}x_1 & x_2 & \cdots & x_p\end{array}\right)$$

and $\mathbf{y}$ another such thing in $\mathbb{R}^p,\mathbf{x}\cdot\mathbf{y}$ is defined as $\sum_j x_jy_j\equiv x_1y_1+x_2y_2+\cdots+x_py_p$. This is called the dot product.

There is a very useful version of Proposition 8.5.28 known as the **Fredholm alternative**. I will only present this for the case of real matrices here. Later a much more elegant and general approach is presented which allows for the general case of complex matrices.

The following definition is used to state the Fredholm alternative.

Definition 8.6.1 *Let $S\subseteq\mathbb{R}^m$. Then $S^{\perp}\equiv\{\mathbf{z}\in\mathbb{R}^m:\mathbf{z}\cdot\mathbf{s}=0$ for every $\mathbf{s}\in S\}$. The funny exponent, $\perp$ is called "perp".*

Now note

$$\ker\left(A^T\right) \equiv \left\{\mathbf{z} : A^T\mathbf{z} = \mathbf{0}\right\} = \left\{\mathbf{z} : \sum_{k=1}^{m} z_k \mathbf{a}_k = 0\right\}$$

Here the $\mathbf{a}_k$ are the rows of A because they are the columns of A^T.

Lemma 8.6.2 *Let A be a real $m \times n$ matrix, let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Then*

$$(A\mathbf{x} \cdot \mathbf{y}) = \left(\mathbf{x} \cdot A^T\mathbf{y}\right)$$

Proof: This follows right away from the definition of the dot product and matrix multiplication.

$$(A\mathbf{x} \cdot \mathbf{y}) = \sum_{k,l} A_{kl} x_l y_k = \sum_{k,l} \left(A^T\right)_{lk} x_l y_k = \left(\mathbf{x} \cdot A^T\mathbf{y}\right). \blacksquare$$

Now it is time to state the Fredholm alternative. The first version of this is the following theorem.

Theorem 8.6.3 *Let A be a real $m \times n$ matrix and let $\mathbf{b} \in \mathbb{R}^m$. There exists a solution $\mathbf{x}$ to the equation $A\mathbf{x} = \mathbf{b}$ if and only if $\mathbf{b} \in \ker\left(A^T\right)^{\perp}$.*

Proof: First suppose $\mathbf{b} \in \ker\left(A^T\right)^{\perp}$. Then this says that if $A^T\mathbf{x} = \mathbf{0}$, it follows that

$$\mathbf{b} \cdot \mathbf{x} = \mathbf{x}^T\mathbf{b} = \mathbf{0}.$$

In other words, taking the transpose, if

$$\mathbf{x}^T A = \mathbf{0} \text{ then } \mathbf{x}^T\mathbf{b} = 0.$$

Thus, if P is a product of elementary matrices such that PA is in row reduced echelon form, then if PA has a row of zeros, in the k^{th} position, obtained from the k^{th} row of P times A, then there is also a zero in the k^{th} position of $P\mathbf{b}$. This is because the k^{th} position in $P\mathbf{b}$ is just the k^{th} row of P times $\mathbf{b}$. Thus the row reduced echelon forms of A and $\left(\begin{array}{c|c} A & \mathbf{b} \end{array}\right)$ have the same number of zero rows. Thus $\operatorname{rank}\left(\begin{array}{c|c} A & \mathbf{b} \end{array}\right) = \operatorname{rank}(A)$. By Proposition 8.5.28, there exists a solution $\mathbf{x}$ to the system $A\mathbf{x} = \mathbf{b}$. It remains to prove the converse.

Let $\mathbf{z} \in \ker\left(A^T\right)$ and suppose $A\mathbf{x} = \mathbf{b}$. I need to verify $\mathbf{b} \cdot \mathbf{z} = 0$. By Lemma 8.6.2,

$$\mathbf{b} \cdot \mathbf{z} = A\mathbf{x} \cdot \mathbf{z} = \mathbf{x} \cdot A^T\mathbf{z} = \mathbf{x} \cdot \mathbf{0} = 0 \blacksquare$$

This implies the following corollary which is also called the Fredholm alternative. The "alternative" becomes more clear in this corollary.

Corollary 8.6.4 *Let A be an $m \times n$ matrix. Then A maps $\mathbb{R}^n$ onto $\mathbb{R}^m$ if and only if the only solution to $A^T\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$.*

Proof: If the only solution to $A^T\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$, then $\ker\left(A^T\right) = \{\mathbf{0}\}$ and so $\ker\left(A^T\right)^{\perp} = \mathbb{R}^m$ because every $\mathbf{b} \in \mathbb{R}^m$ has the property that $\mathbf{b} \cdot \mathbf{0} = 0$. Therefore, $A\mathbf{x} = \mathbf{b}$ has a solution for any $\mathbf{b} \in \mathbb{R}^m$ because the $\mathbf{b}$ for which there is a solution are those in $\ker\left(A^T\right)^{\perp}$ by Theorem 8.6.3. In other words, A maps $\mathbb{R}^n$ onto $\mathbb{R}^m$.

Conversely if A is onto, then if $A^T\mathbf{x} = \mathbf{0}$, there exists $\mathbf{y}$ such that $\mathbf{x} = A\mathbf{y}$ and then $A^T A\mathbf{y} = \mathbf{0}$ and so $|A\mathbf{y}|^2 = A\mathbf{y} \cdot A\mathbf{y} = A^T A\mathbf{y} \cdot \mathbf{y} = \mathbf{0} \cdot \mathbf{y} = 0$ and so $\mathbf{x} = A\mathbf{y} = \mathbf{0}$. $\blacksquare$

Here is an amusing example.

Example 8.6.5 *Let A be an $m \times n$ matrix in which $m > n$. Then A cannot map onto $\mathbb{R}^m$.*

The reason for this is that A^T is an $n \times m$ where $m > n$ and so in the augmented matrix

$$\left(A^T|\mathbf{0}\right)$$

there must be some free variables. Thus there exists a nonzero vector $\mathbf{x}$ such that $A^T\mathbf{x} = \mathbf{0}$.

8.6.1 Row, Column, And Determinant Rank

I will now present a review of earlier topics and prove Theorem 8.3.4.

Definition 8.6.6 *A sub-matrix 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 $r \times r$ sub-matrix 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 8.6.7 *If A, an $m \times n$ matrix 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. Thus some $r \times r$ submatrix has non zero determinant and there is no larger square submatrix which has non zero determinant. Suppose such a submatrix is determined by the r columns whose indices are

$$j_1 < \cdots < j_r$$

and the r rows whose indices are

$$i_1 < \cdots < i_r$$

I want to show that every row is a linear combination of these rows. Consider the l^{th} row and let p be an index between 1 and n. Form the following $(r+1) \times (r+1)$ matrix

$$\begin{pmatrix} a_{i_1 j_1} & \cdots & a_{i_1 j_r} & a_{i_1 p} \\ \vdots & & \vdots & \vdots \\ a_{i_r j_1} & \cdots & a_{i_r j_r} & a_{i_r p} \\ a_{l j_1} & \cdots & a_{l j_r} & a_{lp} \end{pmatrix}$$

Of course you can assume $l \notin \{i_1, \cdots, i_r\}$ because there is nothing to prove if the l^{th} row is one of the chosen ones. The above matrix has determinant 0. This is because if $p \notin \{j_1, \cdots, j_r\}$ then the above would be a submatrix of A which is too large to have non zero determinant. On the other hand, if $p \in \{j_1, \cdots, j_r\}$ then the above matrix has two columns which are equal so its determinant is still 0.

Expand the determinant of the above matrix along the last column. Let C_k denote the cofactor associated with the entry $a_{i_k p}$. This is not dependent on the choice of p. Remember, you delete the column and the row the entry is in and take the determinant of what is left and multiply by -1 raised to an appropriate power. Let C denote the cofactor associated with a_{lp}. This is given to be nonzero, it being the determinant of the matrix

$$\begin{pmatrix} a_{i_1 j_1} & \cdots & a_{i_1 j_r} \\ \vdots & & \vdots \\ a_{i_r j_1} & \cdots & a_{i_r j_r} \end{pmatrix}$$

Thus $0 = a_{lp}C + \sum_{k=1}^{r} C_k a_{i_k p}$ which implies $a_{lp} = \sum_{k=1}^{r} \frac{-C_k}{C} a_{i_k p} \equiv \sum_{k=1}^{r} m_k a_{i_k p}$. Since this is true for every p and since m_k does not depend on p, this has shown the l^{th} row is a linear combination of the $i_1, i_2, \cdots, i_r$ rows. ■

Corollary 8.6.8 *The determinant rank equals the row rank.*

Proof: From Theorem 8.6.7, 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 $r \times r$ sub-matrix 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 8.6.9 *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 8.6.8, 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 8.6.10 *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 8.6.7. 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 $\mathbf{x} \equiv \begin{pmatrix} \alpha_1 & \cdots & -1 & \cdots & \alpha_n \end{pmatrix}^T$. 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 7.1.14 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.) ■

Corollary 8.6.11 *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. ■

Recall Proposition 8.2.12. This is reviewed here in terms of rank.

Corollary 8.6.12 *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, $\det(A) \neq 0$ and it follows A must have rank n and so the row reduced echelon form of A is I. Therefore, by Theorem 8.1.10 there is a sequence of elementary matrices, $E_1, \cdots, E_p$ which accomplish successive row operations 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 8.1.10 each of these in this product is an elementary matrix. ■

8.7 Exercises

1. Let $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ be vectors in $\mathbb{R}^n$. The parallelepiped determined by these vectors

$$P(\mathbf{u}_1, \cdots, \mathbf{u}_n)$$

is defined as

$$P(\mathbf{u}_1, \cdots, \mathbf{u}_n) \equiv \left\{ \sum_{k=1}^{n} t_k \mathbf{u}_k : t_k \in [0,1] \text{ for all } k \right\}.$$

Now let A be an $n \times n$ matrix. Show that

$$\{A\mathbf{x} : \mathbf{x} \in P(\mathbf{u}_1, \cdots, \mathbf{u}_n)\}$$

is also a parallelepiped.

2. In the context of Problem 1, draw $P(\mathbf{e}_1, \mathbf{e}_2)$ where $\mathbf{e}_1, \mathbf{e}_2$ are the standard basis vectors for $\mathbb{R}^2$. Thus $\mathbf{e}_1 = (1,0), \mathbf{e}_2 = (0,1)$. Now suppose
$$E = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$
where E is the elementary matrix which takes the third row and adds to the first. Draw
$$\{E\mathbf{x} : \mathbf{x} \in P(\mathbf{e}_1, \mathbf{e}_2)\}.$$
In other words, draw the result of doing E to the vectors in $P(\mathbf{e}_1, \mathbf{e}_2)$. Next draw the results of doing the other elementary matrices to $P(\mathbf{e}_1, \mathbf{e}_2)$.

3. In the context of Problem 1, either draw or describe the result of doing elementary matrices to $P(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. Describe geometrically the conclusion of Corollary 8.6.12.

4. 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) $\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 5 \\ 0 & 0 & 1 & 2 & 0 & 4 \\ 0 & 0 & 0 & 0 & 1 & 3 \end{pmatrix}$

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

6. 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) $\begin{pmatrix} 0 & 1 & 0 & 2 & 1 & 1 & 2 \\ 0 & 3 & 2 & 6 & 1 & 5 & 1 \\ 0 & 1 & 1 & 2 & 0 & 2 & 1 \\ 0 & 2 & 1 & 4 & 0 & 3 & 1 \end{pmatrix}$

7. Suppose A is an $m \times n$ matrix. Explain why the rank of A is always no larger than $\min(m,n)$.

8. A matrix A is called a projection if $A^2 = A$. Here is a matrix.

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

Show that this is a projection. Show that a vector in the column space of a projection matrix is left unchanged by multiplication by A.

9. 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.

10. 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.

11. 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.

12. Let $M = \left\{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 = u_1 = 0 \right\}$. Is M a subspace? Explain.

13. Let $M = \left\{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_3 \geq u_1 \right\}$. Is M a subspace? Explain.

14. Let $\mathbf{w} \in \mathbb{R}^4$ and let $M = \left\{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : \mathbf{w} \cdot \mathbf{u} = 0 \right\}$. Is M a subspace? Explain.

15. Let $M = \left\{ \mathbf{u} = (u_1, u_2, u_3, u_4) \in \mathbb{R}^4 : u_i \geq 0 \text{ for each } i = 1,2,3,4 \right\}$. Is M a subspace? Explain.

16. 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.

17. 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.

18. 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.

19. Study the definition of span. Explain what is meant by the span of a set of vectors. Include pictures.

20. Suppose $\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ is a set of vectors from $\mathbb{F}^n$. Show that $\text{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k)$ contains $\mathbf{0}$.

21. Study the definition of linear independence. Explain in your own words what is meant by linear independence and linear dependence. Illustrate with pictures.

22. Use Corollary 8.5.18 to prove the following theorem: If A, B are $n \times n$ matrices and if $AB = I$, then $BA = I$ and $B = A^{-1}$. **Hint:** First note that if $AB = I$, then it must be the case that A is onto. Explain why this requires $\text{span}(\text{columns of } A) = \mathbb{F}^n$. Now explain why, using the corollary that this requires A to be one to one. Next explain why $A(BA - I) = 0$ and why the fact that A is one to one implies $BA = I$.

23. Here are three vectors. Determine whether they are linearly independent or linearly dependent.
$$\begin{pmatrix} 1 & 2 & 0 \end{pmatrix}^T, \begin{pmatrix} 2 & 0 & 1 \end{pmatrix}^T, \begin{pmatrix} 3 & 0 & 0 \end{pmatrix}^T$$

24. Here are three vectors. Determine whether they are linearly independent or linearly dependent.
$$\begin{pmatrix} 4 & 2 & 0 \end{pmatrix}^T, \begin{pmatrix} 2 & 2 & 1 \end{pmatrix}^T, \begin{pmatrix} 0 & 2 & 2 \end{pmatrix}^T$$

25. Here are three vectors. Determine whether they are linearly independent or linearly dependent.
$$\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}^T, \begin{pmatrix} 4 & 5 & 1 \end{pmatrix}^T, \begin{pmatrix} 3 & 1 & 0 \end{pmatrix}^T$$

26. Here are four vectors. Determine whether they span $\mathbb{R}^3$. Are these vectors linearly independent?
$$\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}^T, \begin{pmatrix} 4 & 3 & 3 \end{pmatrix}^T, \begin{pmatrix} 3 & 1 & 0 \end{pmatrix}^T, \begin{pmatrix} 2 & 4 & 6 \end{pmatrix}^T$$

27. Here are four vectors. Determine whether they span $\mathbb{R}^3$. Are these vectors linearly independent?
$$\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}^T, \begin{pmatrix} 4 & 3 & 3 \end{pmatrix}^T, \begin{pmatrix} 3 & 2 & 0 \end{pmatrix}^T, \begin{pmatrix} 2 & 4 & 6 \end{pmatrix}^T$$

28. 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 $\mathbb{R}^3$.
$$\begin{pmatrix} 1 & 0 & 3 \end{pmatrix}^T, \begin{pmatrix} 4 & 3 & 3 \end{pmatrix}^T, \begin{pmatrix} 1 & 2 & 0 \end{pmatrix}^T, \begin{pmatrix} 2 & 4 & 0 \end{pmatrix}^T$$

29. 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 $\mathbb{R}^3$.

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

30. 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 $\mathbb{R}^3$.

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

31. 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 $\mathbb{R}^3$.

$$\begin{pmatrix} 1 & 0 & 3 \end{pmatrix}^T, \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T, \begin{pmatrix} 1 & 1 & 3 \end{pmatrix}^T, \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}^T$$

32. 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.

33. 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.

34. 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.

35. 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.

36. If you have 6 vectors in $\mathbb{F}^5$, is it possible they are linearly independent? Explain.

37. 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.

38. 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.

39. 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.

40. 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.

41. If $\mathbf{b} \neq \mathbf{0}$, can the solution set of $A\mathbf{x} = \mathbf{b}$ be a plane through the origin? Explain.

42. 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.

43. 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.

44. 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.

45. Suppose there is a unique solution to a system of linear equations. What must be true of the pivot columns in the augmented matrix.

46. 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 $\text{rank}(A|\mathbf{b}) = 2$.

47. 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.

48. 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.

49. Explain why an $n\times n$ matrix A is both one to one and onto if and only if its rank is n.

50. 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

$$\mathbf{x}-\sum_{i=1}^k\mathbf{z}_i\in\ker(B).$$

51. 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?

52. Suppose $\det(A-\lambda I)=0$. Show using Theorem 9.2.9 there exists $\mathbf{x}\neq\mathbf{0}$ such that

$$(A-\lambda I)\mathbf{x}=\mathbf{0}.$$

53. 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 9.2.9 to argue $\det(A-\lambda I)=0$. What sort of equation is this? How many solutions does it have?

54. 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.

55. 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
$$A^TA\mathbf{x}=A^T\mathbf{b}$$
Next show that if $\mathbf{x},\mathbf{x}_1$ are two solutions, then $A\mathbf{x}=A\mathbf{x}_1$. **Hint:** First show that $\left(A^TA\right)^T=A^TA$. Next show if $\mathbf{x}\in\ker\left(A^TA\right)$, then $A\mathbf{x}=\mathbf{0}$. Finally apply the Fredholm alternative. This will give existence of a solution.

56. Show that in the context of Problem 55 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 $A(\mathbb{R}^n)$. This gives an approach to least squares based on rank considerations. In Problem 55, $\mathbf{x}$ is the least squares solution to $A\mathbf{x}=\mathbf{b}$, which may not even have an actual solution.

57. Let A be an $n\times n$ matrix and consider the matrices $\left\{I,A,A^2,\cdots,A^{n^2}\right\}$. Explain why there exist scalars, c_i not all zero such that
$$\sum_{i=1}^{n^2}c_iA^i=0.$$
Then argue there exists a polynomial, $p(\lambda)$ of the form
$$\lambda^m+d_{m-1}\lambda^{m-1}+\cdots+d_1\lambda+d_0$$
such that $p(A)=0$ and if $q(\lambda)$ is another polynomial such that $q(A)=0$, then $q(\lambda)$ is of the form $p(\lambda)l(\lambda)$ for some polynomial, $l(\lambda)$. This extra special polynomial, $p(\lambda)$ is called the **minimal polynomial**. **Hint:** You might consider an $n\times n$ matrix as a vector in $\mathbb{F}^{n^2}$.

58. Here are some invertible matrices. Write them as a product of elementary matrices.

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

(b) $\begin{pmatrix}1&1&-3\\2&3&-7\\1&1&-4\end{pmatrix}$

(c) $\begin{pmatrix}2&3&9\\1&2&6\\1&1&2\end{pmatrix}$

(d) $\begin{pmatrix}2&4&2\\2&2&2\\1&2&2\end{pmatrix}$

59. Here is a matrix: $\begin{pmatrix}1&1&3\\2&1&5\\2&3&7\end{pmatrix}$. Find a product of elementary matrices which, when this product multiplies the given matrix on the left, the result will be in row reduced echelon form.

60. Explain why such a sequence of elementary matrices which row reduces a given matrix to row reduced echelon form is not unique. That is, tell why you could have two different products of elementary matrices which produce the same result.

Chapter 9

Linear Transformations

9.1 Linear Transformations

An $m \times n$ matrix can be used to transform vectors in $\mathbb{F}^n$ to vectors in $\mathbb{F}^m$ through the use of matrix multiplication.

Example 9.1.1 *Consider the matrix* $\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix}$. *Think of it as a function which takes vectors in* $\mathbb{F}^3$ *and makes them in to vectors in* $\mathbb{F}^2$ *as follows. For* $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ *a vector in* $\mathbb{F}^3$, *multiply on the left by the given matrix to obtain the vector in* $\mathbb{F}^2$. *Here are some numerical examples.*

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

$$\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 10 \\ 5 \\ -3 \end{pmatrix} = \begin{pmatrix} 20 \\ 25 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 7 \\ 3 \end{pmatrix} = \begin{pmatrix} 14 \\ 7 \end{pmatrix},$$

More generally,

$$\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x+2y \\ 2x+y \end{pmatrix}$$

The idea is to define a function which takes vectors in $\mathbb{F}^3$ *and delivers new vectors in* $\mathbb{F}^2$.

This is an example of something called a linear transformation.

Definition 9.1.2 *Let* $T : \mathbb{F}^n \mapsto \mathbb{F}^m$ *be a function. Thus for each* $\mathbf{x} \in \mathbb{F}^n, T\mathbf{x} \in \mathbb{F}^m$. *Then* T *is a* ***linear transformation*** *if whenever* α, β *are scalars and* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *are vectors in* $\mathbb{F}^n$,

$$T(\alpha \mathbf{x}_1 + \beta \mathbf{x}_2) = \alpha_1 T\mathbf{x}_1 + \beta T\mathbf{x}_2.$$

A linear transformation is also called a ***homomorphism.*** *In the case that T is in addition to this one to one and onto, it is sometimes called an* ***isomorphism.***

The last two terms are typically used more in abstract algebra than in linear algebra so in this book, such mappings will be referred to as linear transformations. In sloppy language, it distributes across vector addition and you can factor out the scalars.

In words, linear transformations distribute across + and allow you to factor out scalars. At this point, recall the properties of matrix multiplication. The pertinent property is 5.14 on Page 93. Recall it states that for a and b scalars,

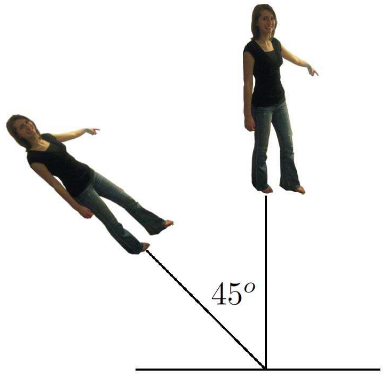
$$A(aB+bC) = aAB+bAC$$

In particular, for A an $m \times n$ matrix and B and $C, n \times 1$ matrices (column vectors) the above formula holds which is nothing more than the statement that matrix multiplication gives an example of a linear transformation.

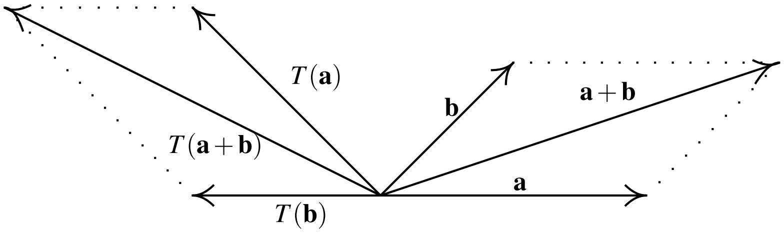
The reason this concept is so important is there are many examples of things which are linear transformations. You might remember from calculus that the operator which consists of taking the derivative is a linear transformation. That is, if f, g are functions (vectors) and α, β are numbers (scalars)

$$\frac{d}{dx}(\alpha f+\beta g) = \alpha \frac{d}{dx} f + \beta \frac{d}{dx} g$$

Another example of a linear transformation is that of rotation through an angle. For example, I may want to rotate every vector through an angle of 45 degrees. Such a rotation would achieve something like the following if applied to each vector corresponding to points on the picture which is standing upright.



More generally, denote a rotation by T. Why is such a transformation linear? Consider the following picture which illustrates a rotation through 135 degrees.



To get $T(\mathbf{a}+\mathbf{b})$, you can add $T\mathbf{a}$ and $T\mathbf{b}$. Here is why. If you add $T\mathbf{a}$ to $T\mathbf{b}$ you get the diagonal of the parallelogram determined by $T\mathbf{a}$ and $T\mathbf{b}$. This diagonal also results from rotating the diagonal of the parallelogram determined by $\mathbf{a}$ and $\mathbf{b}$. This is because the rotation preserves all angles between the vectors as well as their lengths. In particular, it preserves the shape of this parallelogram. Thus both $T\mathbf{a}+T\mathbf{b}$ and $T(\mathbf{a}+\mathbf{b})$ give the same directed line segment. Thus T distributes across $+$ where $+$ refers to vector addition. Similarly, if k is a number $Tk\mathbf{a} = kT\mathbf{a}$ (draw a picture) and so you can factor out scalars also. Thus rotations are an example of a linear transformation.

Definition 9.1.3 *A linear transformation is called* ***one to one*** *(often written as* $1-1$*) if it never takes two different vectors to the same vector. Thus T is one to one if whenever* $\mathbf{x} \neq \mathbf{y}$

$$T\mathbf{x} \neq T\mathbf{y}.$$

Equivalently, if $T(\mathbf{x}) = T(\mathbf{y})$, *then* $\mathbf{x} = \mathbf{y}$.

In the case that a linear transformation comes from matrix multiplication, it is common usage to refer to the matrix as a one to one matrix when the linear transformation it determines is one to one.

Definition 9.1.4 *A linear transformation mapping* $\mathbb{F}^n$ *to* $\mathbb{F}^m$ *is called* ***onto*** *if whenever* $\mathbf{y} \in \mathbb{F}^m$ *there exists* $\mathbf{x} \in \mathbb{F}^n$ *such that* $T(\mathbf{x}) = \mathbf{y}$.

Thus T is onto if everything in $\mathbb{F}^m$ gets hit. In the case that a linear transformation comes from matrix multiplication, it is common to refer to the matrix as onto when the linear transformation it determines is onto. Also it is common usage to write $T\mathbb{F}^n$, $T(\mathbb{F}^n)$,or $\mathrm{Im}(T)$ as the set of vectors of $\mathbb{F}^m$ which are of the form $T\mathbf{x}$ for some $\mathbf{x} \in \mathbb{F}^n$. In the case that T is obtained from multiplication by an $m \times n$ matrix A, it is standard to simply write $A(\mathbb{F}^n)$, $A\mathbb{F}^n$, or $\mathrm{Im}(A)$ to denote those vectors in $\mathbb{F}^m$ which are obtained in the form $A\mathbf{x}$ for some $\mathbf{x} \in \mathbb{F}^n$.

9.2 Constructing The Matrix Of A Linear Transformation

It turns out that if T is any linear transformation which maps $\mathbb{F}^n$ to $\mathbb{F}^m$, there is always an $m \times n$ matrix A with the property that

$$A\mathbf{x} = T\mathbf{x} \tag{9.1}$$

for all $\mathbf{x} \in \mathbb{F}^n$. Here is why. Suppose $T : \mathbb{F}^n \mapsto \mathbb{F}^m$ is a linear transformation and you want to find the matrix defined by this linear transformation as described in 9.1. Then if $\mathbf{x} \in \mathbb{F}^n$ it follows

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i$$

where $\mathbf{e}_i$ is the vector which has zeros in every slot but the i^{th} and a 1 in this slot. Then since T is linear,

$$T\mathbf{x} = \sum_{i=1}^{n} x_i T(\mathbf{e}_i)$$

$$= \begin{pmatrix} | & & | \\ T(\mathbf{e}_1) & \cdots & T(\mathbf{e}_n) \\ | & & | \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \equiv A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

and so you see that the matrix desired is obtained from letting the i^{th} column equal $T(\mathbf{e}_i)$. We state this as the following theorem.

Theorem 9.2.1 *Let T be a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$. Then the matrix A satisfying 9.1 is given by*

$$\begin{pmatrix} | & & | \\ T(\mathbf{e}_1) & \cdots & T(\mathbf{e}_n) \\ | & & | \end{pmatrix}$$

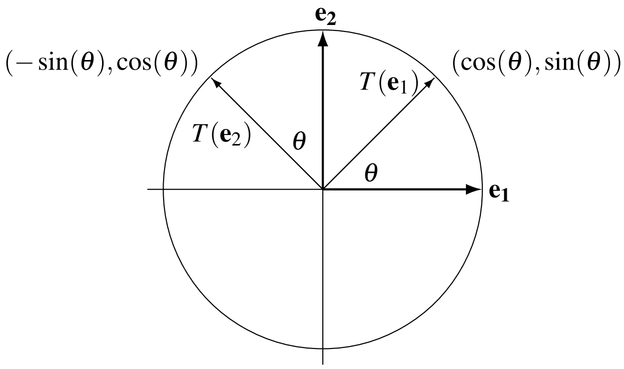
where $T\mathbf{e}_i$ is the i^{th} column of A.

9.2.1 Rotations in $\mathbb{R}^2$

Sometimes you need to find a matrix which represents a given linear transformation which is described in geometrical terms. The idea is to produce a matrix which you can multiply a vector by to get the same thing as some geometrical description. A good example of this is the problem of rotation of vectors discussed above. Consider the problem of rotating through an angle of θ.

Example 9.2.2 *Determine the matrix which represents the linear transformation defined by rotating every vector through an angle of θ.*

Let $\mathbf{e}_1 \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathbf{e}_2 \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. These identify the geometric vectors which point along the positive x axis and positive y axis as shown.



From the above, you only need to find $T\mathbf{e}_1$ and $T\mathbf{e}_2$, the first being the first column of the desired matrix A and the second being the second column. From the definition of the $\cos,\sin$ the coordinates of $T(\mathbf{e}_1)$ are as shown in the picture. The coordinates of $T(\mathbf{e}_2)$ also follow from simple trigonometry. Thus

$$T\mathbf{e}_1 = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, T\mathbf{e}_2 = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}.$$

Therefore, from Theorem 9.2.1,

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

For those who prefer a more algebraic approach, the definition of $(\cos(\theta),\sin(\theta))$ is as the x and y coordinates of the point $(1,0)$. Now the point of the vector from $(0,0)$ to $(0,1)$, $\mathbf{e}_2$ is exactly $\pi/2$ further along along the unit circle. Therefore, when it is rotated through an angle of θ the x and y coordinates are given by

$$(x,y) = (\cos(\theta+\pi/2),\sin(\theta+\pi/2)) = (-\sin\theta,\cos\theta).$$

Example 9.2.3 *Find the matrix of the linear transformation which is obtained by first rotating all vectors through an angle of ϕ and then through an angle θ. Thus you want the linear transformation which rotates all angles through an angle of $\theta+\phi$.*

Let $T_{\theta+\phi}$ denote the linear transformation which rotates every vector through an angle of $\theta+\phi$. Then to get $T_{\theta+\phi}$, you could first do T_ϕ and then do T_θ where T_ϕ is the linear transformation which rotates through an angle of ϕ and T_θ is the linear transformation which rotates through an angle of θ. Denoting the corresponding matrices by $A_{\theta+\phi}$, A_ϕ, and A_θ, you must have for every $\mathbf{x}$

$$A_{\theta+\phi}\mathbf{x} = T_{\theta+\phi}\mathbf{x} = T_\theta T_\phi \mathbf{x} = A_\theta A_\phi \mathbf{x}.$$

Consequently, you must have

$$\begin{aligned} A_{\theta+\phi} &= \begin{pmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{pmatrix} = A_\theta A_\phi \\ &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}. \end{aligned}$$

You know how to multiply matrices. Do so to the pair on the right. This yields

$$\begin{aligned}&\begin{pmatrix}\cos(\theta+\phi) & -\sin(\theta+\phi)\\ \sin(\theta+\phi) & \cos(\theta+\phi)\end{pmatrix}\\ = &\begin{pmatrix}\cos\theta\cos\phi-\sin\theta\sin\phi & -\cos\theta\sin\phi-\sin\theta\cos\phi\\ \sin\theta\cos\phi+\cos\theta\sin\phi & \cos\theta\cos\phi-\sin\theta\sin\phi\end{pmatrix}.\end{aligned}$$

Don't these look familiar? They are the usual trig. identities for the sum of two angles derived here using linear algebra concepts.

You do not have to stop with two dimensions. You can consider rotations and other geometric concepts in any number of dimensions. This is one of the major advantages of linear algebra. You can break down a difficult geometrical procedure into small steps, each corresponding to multiplication by an appropriate matrix. Then by multiplying the matrices, you can obtain a single matrix which can give you numerical information on the results of applying the given sequence of simple procedures. That which you could never visualize can still be understood to the extent of finding exact numerical answers. Another example follows.

Example 9.2.4 *Find the matrix of the linear transformation which is obtained by first rotating all vectors through an angle of $\pi/6$ and then reflecting through the x axis.*

As shown in Example 9.2.3, the matrix of the transformation which involves rotating through an angle of $\pi/6$ is

$$\begin{pmatrix}\cos(\pi/6) & -\sin(\pi/6)\\ \sin(\pi/6) & \cos(\pi/6)\end{pmatrix}=\begin{pmatrix}\frac{1}{2}\sqrt{3} & -\frac{1}{2}\\ \frac{1}{2} & \frac{1}{2}\sqrt{3}\end{pmatrix}$$

The matrix for the transformation which reflects all vectors through the x axis is

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

Therefore, the matrix of the linear transformation which first rotates through $\pi/6$ and then reflects through the x axis is

$$\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}\frac{1}{2}\sqrt{3} & -\frac{1}{2}\\ \frac{1}{2} & \frac{1}{2}\sqrt{3}\end{pmatrix}=\begin{pmatrix}\frac{1}{2}\sqrt{3} & -\frac{1}{2}\\ -\frac{1}{2} & -\frac{1}{2}\sqrt{3}\end{pmatrix}.$$

9.2.2 Rotations About A Particular Vector

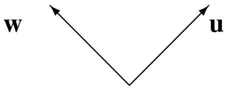
The problem is to find the matrix of the linear transformation which rotates all vectors about a given unit vector $\mathbf{u}$ which is possibly not one of the coordinate vectors $\mathbf{i},\mathbf{j}$, or $\mathbf{k}$. Suppose for $|c|\neq 1$

$$\mathbf{u}=(a,b,c)\,,\ \sqrt{a^2+b^2+c^2}=1.$$

First I will produce a matrix which maps $\mathbf{u}$ to $\mathbf{k}$ such that the right handed rotation about $\mathbf{k}$ corresponds to the right handed rotation about $\mathbf{u}$. Then I will rotate about $\mathbf{k}$ and finally, I will multiply by the inverse of the first matrix to get the desired result.

To begin, find vectors $\mathbf{w},\mathbf{v}$ such that $\mathbf{w}\times\mathbf{v}=\mathbf{u}$. Let

$$\mathbf{w}=\left(-\frac{b}{\sqrt{a^2+b^2}},\frac{a}{\sqrt{a^2+b^2}},0\right).$$



This vector is clearly perpendicular to $\mathbf{u}$. Then $\mathbf{v}=(a,b,c)\times\mathbf{w}\equiv\mathbf{u}\times\mathbf{w}$. Thus from the geometric description of the cross product, $\mathbf{w}\times\mathbf{v}=\mathbf{u}$. Computing the cross product gives

$$\mathbf{v}=(a,b,c)\times\left(-\frac{b}{\sqrt{a^2+b^2}},\frac{a}{\sqrt{a^2+b^2}},0\right)$$

$$=\left(-c\frac{a}{\sqrt{(a^2+b^2)}},-c\frac{b}{\sqrt{(a^2+b^2)}},\frac{a^2}{\sqrt{(a^2+b^2)}}+\frac{b^2}{\sqrt{(a^2+b^2)}}\right)$$

Now I want to have $T\mathbf{w}=\mathbf{i},T\mathbf{v}=\mathbf{j},T\mathbf{u}=\mathbf{k}$. What does this? It is the inverse of the matrix which takes $\mathbf{i}$ to $\mathbf{w}$, $\mathbf{j}$ to $\mathbf{v}$, and $\mathbf{k}$ to $\mathbf{u}$. This matrix is

$$\begin{pmatrix} -\frac{b}{\sqrt{a^2+b^2}} & -\frac{c}{\sqrt{(a^2+b^2)}}a & a \\ \frac{a}{\sqrt{a^2+b^2}} & -\frac{c}{\sqrt{(a^2+b^2)}}b & b \\ 0 & \frac{a^2+b^2}{\sqrt{a^2+b^2}} & c \end{pmatrix}.$$

Its inverse is

$$\begin{pmatrix} -\frac{1}{\sqrt{(a^2+b^2)}}b & \frac{1}{\sqrt{(a^2+b^2)}}a & 0 \\ -\frac{c}{\sqrt{(a^2+b^2)}}a & -\frac{c}{\sqrt{(a^2+b^2)}}b & \sqrt{(a^2+b^2)} \\ a & b & c \end{pmatrix}$$

Therefore, the matrix which does the rotating is

$$\begin{pmatrix} -\frac{b}{\sqrt{a^2+b^2}} & -\frac{c}{\sqrt{(a^2+b^2)}}a & a \\ \frac{a}{\sqrt{a^2+b^2}} & -\frac{c}{\sqrt{(a^2+b^2)}}b & b \\ 0 & \frac{a^2+b^2}{\sqrt{a^2+b^2}} & c \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}\cdot$$

$$\begin{pmatrix} -\frac{1}{\sqrt{(a^2+b^2)}}b & \frac{1}{\sqrt{(a^2+b^2)}}a & 0 \\ -\frac{c}{\sqrt{(a^2+b^2)}}a & -\frac{c}{\sqrt{(a^2+b^2)}}b & \sqrt{(a^2+b^2)} \\ a & b & c \end{pmatrix}$$

This yields a matrix whose columns are

$$\begin{pmatrix} \frac{b^2\cos\theta+c^2a^2\cos\theta+a^4+a^2b^2}{a^2+b^2} \\ \frac{-ba\cos\theta+cb^2\sin\theta+ca^2\sin\theta+c^2ab\cos\theta+ba^3+b^3a}{a^2+b^2} \\ -(\sin\theta)b-(\cos\theta)ca+ca \end{pmatrix},$$

$$\begin{pmatrix} \frac{-ba\cos\theta - ca^2\sin\theta - cb^2\sin\theta + c^2ab\cos\theta + ba^3 + b^3a}{a^2+b^2} \\ \frac{a^2\cos\theta + c^2b^2\cos\theta + a^2b^2 + b^4}{a^2+b^2} \\ (\sin\theta)\,a - (\cos\theta)\,cb + cb \end{pmatrix},$$

$$\begin{pmatrix} (\sin\theta)\,b - (\cos\theta)\,ca + ca \\ -(\sin\theta)\,a - (\cos\theta)\,cb + cb \\ \left(a^2+b^2\right)\cos\theta + c^2 \end{pmatrix}$$

Using the assumption that **u** is a unit vector so that $a^2+b^2+c^2 = 1$, it follows the desired matrix is

$$\begin{pmatrix} \cos\theta - a^2\cos\theta & -ba\cos\theta + ba & (\sin\theta)\,b - (\cos\theta)\,ca \\ +a^2 & -c\sin\theta & +ca \\ -ba\cos\theta + ba & -b^2\cos\theta + b^2 & -(\sin\theta)\,a - (\cos\theta)\,cb \\ +c\sin\theta & +\cos\theta & +cb \\ -(\sin\theta)\,b - (\cos\theta)\,ca & (\sin\theta)\,a - (\cos\theta)\,cb & \left(1-c^2\right)\cos\theta \\ +ca & +cb & +c^2 \end{pmatrix}$$

This was done under the assumption that $|c| \neq 1$. However, if this condition does not hold, you can verify directly that the above still gives the correct answer.

9.2.3 Projections

In Physics it is important to consider the work done by a force field on an object. This involves the concept of projection onto a vector. Suppose you want to find the projection of a vector **v** onto the given vector **u**, denoted by $\text{proj}_{\mathbf{u}}(\mathbf{v})$ This is done using the dot product as follows.

$$\text{proj}_{\mathbf{u}}(\mathbf{v}) = \left(\frac{\mathbf{v}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\right)\mathbf{u}$$

Because of properties of the dot product, the map $\mathbf{v}\mapsto\text{proj}_{\mathbf{u}}(\mathbf{v})$ is linear,

$$\begin{aligned} \text{proj}_{\mathbf{u}}(\alpha\mathbf{v}+\beta\mathbf{w}) &= \left(\frac{\alpha\mathbf{v}+\beta\mathbf{w}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\right)\mathbf{u} = \alpha\left(\frac{\mathbf{v}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\right)\mathbf{u} + \beta\left(\frac{\mathbf{w}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\right)\mathbf{u} \\ &= \alpha\,\text{proj}_{\mathbf{u}}(\mathbf{v}) + \beta\,\text{proj}_{\mathbf{u}}(\mathbf{w}). \end{aligned}$$

Example 9.2.5 *Let the projection map be defined above and let* $\mathbf{u} = (1,2,3)^T$. *Does this linear transformation come from multiplication by a matrix? If so, what is the matrix?*

You can find this matrix in the same way as in the previous example. Let $\mathbf{e}_i$ denote the vector in $\mathbb{R}^n$ which has a 1 in the i^{th} position and a zero everywhere else. Thus a typical vector $\mathbf{x} = (x_1,\cdots,x_n)^T$ can be written in a unique way as

$$\mathbf{x} = \sum_{j=1}^{n} x_j\mathbf{e}_j.$$

From the way you multiply a matrix by a vector, it follows that $\operatorname{proj}_{\mathbf{u}}(\mathbf{e}_i)$ gives the i^{th} column of the desired matrix. Therefore, it is only necessary to find

$$\operatorname{proj}_{\mathbf{u}}(\mathbf{e}_i) \equiv \left(\frac{\mathbf{e}_i \cdot \mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\right)\mathbf{u}$$

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

$$\frac{1}{14}\begin{pmatrix}1 & 2 & 3\\2 & 4 & 6\\3 & 6 & 9\end{pmatrix}.$$

9.2.4 Matrices Which Are One To One Or Onto

Lemma 9.2.6 *Let A be an $m\times n$ matrix. Then $A(\mathbb{F}^n) = \operatorname{span}(\mathbf{a}_1,\cdots,\mathbf{a}_n)$ where $\mathbf{a}_1,\cdots,\mathbf{a}_n$ denote the columns of A. In fact, for $\mathbf{x} = (x_1,\cdots,x_n)^T$,*

$$A\mathbf{x} = \sum_{k=1}^{n} x_k \mathbf{a}_k.$$

Proof: This follows from the definition of matrix multiplication in Definition 5.1.9 on Page 87. ■

The following is a theorem of major significance. First here is an interesting observation.

Observation 9.2.7 *Let A be an $m\times n$ matrix. Then A is one to one if and only if $A\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$.*

Here is why: $A\mathbf{0} = A(\mathbf{0}+\mathbf{0}) = A\mathbf{0}+A\mathbf{0}$ and so $A\mathbf{0} = \mathbf{0}$.

Now suppose A is one to one and $A\mathbf{x} = \mathbf{0}$. Then since $A\mathbf{0} = \mathbf{0}$, it follows $\mathbf{x} = \mathbf{0}$. Thus if A is one to one and $A\mathbf{x} = \mathbf{0}$, then $\mathbf{x} = \mathbf{0}$.

Next suppose the condition that $A\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$ is valid. Then if $A\mathbf{x} = A\mathbf{y}$, then $A(\mathbf{x}-\mathbf{y}) = \mathbf{0}$ and so from the condition, $\mathbf{x}-\mathbf{y} = \mathbf{0}$ so that $\mathbf{x} = \mathbf{y}$. Thus A is one to one.

Theorem 9.2.8 *Suppose A is an $n\times n$ matrix. Then A is one to one if and only if A is onto. Also, if B is an $n\times n$ matrix and $AB = I$, then it follows $BA = I$.*

Proof: First suppose A is one to one. Consider the vectors, $\{A\mathbf{e}_1,\cdots,A\mathbf{e}_n\}$ where $\mathbf{e}_k$ is the column vector which is all zeros except for a 1 in the k^{th} position. This set of vectors is linearly independent because if

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

then since A is linear,

$$A\left(\sum_{k=1}^{n} c_k \mathbf{e}_k\right) = \mathbf{0}$$

and since A is one to one, it follows

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

which implies each $c_k = 0$. Therefore, $\{A\mathbf{e}_1, \cdots, A\mathbf{e}_n\}$ must be a basis for $\mathbb{F}^n$ by Corollary 8.5.18 on Page 170. It follows that for $\mathbf{y} \in \mathbb{F}^n$ there exist constants, c_i such that

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

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

Next suppose A is onto. This implies the span of the columns of A equals $\mathbb{F}^n$ and by Corollary 8.5.18 this implies the columns of A are independent. If $A\mathbf{x} = \mathbf{0}$, then letting $\mathbf{x} = (x_1, \cdots, x_n)^T$, it follows

$$\sum_{i=1}^{n} x_i \mathbf{a}_i = \mathbf{0}$$

and so each $x_i = 0$. If $A\mathbf{x} = A\mathbf{y}$, then $A(\mathbf{x} - \mathbf{y}) = \mathbf{0}$ and so $\mathbf{x} = \mathbf{y}$. This shows A is one to one.

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

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

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

This 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

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & -1 \end{pmatrix} \tag{9.2}$$

Then

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

but

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

There is also an important characterization in terms of determinants. This is proved completely in the section on the mathematical theory of the determinant.

Theorem 9.2.9 *Let A be an $n\times n$ matrix and let T_A denote the linear transformation determined by A. Then the following are equivalent.*

1. *T_A is one to one.*
2. *T_A is onto.*
3. *$\det(A)\neq 0$.*

9.2.5 The General Solution Of A Linear System

Recall the following definition which was discussed above.

Definition 9.2.10 *T is a **linear transformation** if whenever $\mathbf{x},\mathbf{y}$ are vectors and a,b scalars,*

$$T(a\mathbf{x}+b\mathbf{y})=aT\mathbf{x}+bT\mathbf{y}. \tag{9.3}$$

Thus linear transformations distribute across addition and pass scalars to the outside. A linear system is one which is of the form

$$T\mathbf{x}=\mathbf{b}.$$

*If $T\mathbf{x}_p=\mathbf{b}$, then $\mathbf{x}_p$ is called a **particular solution** to the linear system.*

For example, if A is an $m\times n$ matrix and T_A is determined by

$$T_A(\mathbf{x})=A\mathbf{x},$$

then from the properties of matrix multiplication, T_A is a linear transformation. In this setting, we will usually write A for the linear transformation as well as the matrix. There are many other examples of linear transformations other than this. In differential equations, you will encounter linear transformations which act on functions to give new functions. In this case, the functions are considered as vectors. Don't worry too much about this at this time. It will happen later. The fundamental idea is that something is linear if 9.3 holds and if whenever a,b are scalars and $\mathbf{x},\mathbf{y}$ are vectors $a\mathbf{x}+b\mathbf{y}$ is a vector. That is you can add vectors and multiply by scalars.

Definition 9.2.11 *Let T be a linear transformation. Define*

$$\ker(T)\equiv\{\mathbf{x}:T\mathbf{x}=\mathbf{0}\}.$$

*In words, $\ker(T)$ is called the **kernel** of T. As just described, $\ker(T)$ consists of the set of all vectors which T sends to $\mathbf{0}$. This is also called the **null space** of T. It is also called the **solution space** of the equation $T\mathbf{x}=\mathbf{0}$.*

The above definition states that $\ker(T)$ is the set of solutions to the equation,

$$T\mathbf{x}=\mathbf{0}.$$

In the case where T is really a matrix, you have been solving such equations for quite some time. However, sometimes linear transformations act on vectors which are not in $\mathbb{F}^n$. There is more on this in Chapter 16 on Page 16 and this is discussed more carefully then. However, consider the following familiar example.

Example 9.2.12 *Let $\frac{d}{dx}$ denote the linear transformation defined on X, the functions which are defined on $\mathbb{R}$ and have a continuous derivative. Find* $\ker\left(\frac{d}{dx}\right)$.

The example asks for functions, f which the property that $\frac{df}{dx} = 0$. As you know from calculus, these functions are the constant functions. Thus $\ker\left(\frac{d}{dx}\right) =$ constant functions.

When T is a linear transformation, systems of the form $T\mathbf{x} = \mathbf{0}$ are called **homogeneous systems**. Thus the solution to the homogeneous system is known as $\ker(T)$.

Systems of the form $T\mathbf{x} = \mathbf{b}$ where $\mathbf{b} \neq \mathbf{0}$ are called **nonhomogeneous systems**. It turns out there is a very interesting and important relation between the solutions to the homogeneous systems and the solutions to the nonhomogeneous systems.

Theorem 9.2.13 *Suppose $\mathbf{x}_p$ is a solution to the linear system,*

$$T\mathbf{x} = \mathbf{b}$$

Then if $\mathbf{y}$ is any other solution, there exists $\mathbf{x} \in \ker(T)$ such that

$$\mathbf{y} = \mathbf{x}_p + \mathbf{x}.$$

Proof: Consider $\mathbf{y} - \mathbf{x}_p \equiv \mathbf{y} + (-1)\mathbf{x}_p$. Then $T\left(\mathbf{y} - \mathbf{x}_p\right) = T\mathbf{y} - T\mathbf{x}_p = \mathbf{b} - \mathbf{b} = \mathbf{0}$. Let

$$\mathbf{x} \equiv \mathbf{y} - \mathbf{x}_p. \blacksquare$$

Sometimes people remember the above theorem in the following form. The solutions to the nonhomogeneous system, $T\mathbf{x} = \mathbf{b}$ are given by $\mathbf{x}_p + \ker(T)$ where $\mathbf{x}_p$ is a particular solution to $T\mathbf{x} = \mathbf{b}$.

I have been vague about what T is and what $\mathbf{x}$ is on purpose. This theorem is completely algebraic in nature and will work whenever you have linear transformations. In particular, it will be important in differential equations. For now, here is a familiar example.

Example 9.2.14 *Let*

$$A = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 2 & 1 & 1 & 2 \\ 4 & 5 & 7 & 2 \end{pmatrix}$$

Find $\ker(A)$. *Equivalently, find the solution space to the system of equations $A\mathbf{x} = \mathbf{0}$.*

This asks you to find $\{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$. In other words you are asked to solve the system, $A\mathbf{x} = \mathbf{0}$. Let $\mathbf{x} = (x, y, z, w)^T$. Then this amounts to solving

$$\begin{pmatrix} 1 & 2 & 3 & 0 \\ 2 & 1 & 1 & 2 \\ 4 & 5 & 7 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

This is the linear system

$$x + 2y + 3z = 0$$
$$2x + y + z + 2w = 0$$
$$4x + 5y + 7z + 2w = 0$$

and you know how to solve this using row operations, (Gauss Elimination). Set up the augmented matrix

$$\left(\begin{array}{cccccc} 1 & 2 & 3 & 0 & | & 0 \\ 2 & 1 & 1 & 2 & | & 0 \\ 4 & 5 & 7 & 2 & | & 0 \end{array}\right)$$

Then row reduce to obtain the row reduced echelon form,

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

This yields $x = \frac{1}{3}z - \frac{4}{3}w$ and $y = \frac{2}{3}w - \frac{5}{3}z$. Thus $\ker(A)$ consists of vectors of the form,

$$\begin{pmatrix} \frac{1}{3}z - \frac{4}{3}w \\ \frac{2}{3}w - \frac{5}{3}z \\ z \\ w \end{pmatrix} = z\begin{pmatrix} \frac{1}{3} \\ -\frac{5}{3} \\ 1 \\ 0 \end{pmatrix} + w\begin{pmatrix} -\frac{4}{3} \\ \frac{2}{3} \\ 0 \\ 1 \end{pmatrix}.$$

Example 9.2.15 *The* ***general solution*** *of a linear system of equations is just the set of all solutions. Find the general solution to the linear system,*

$$\begin{pmatrix} 1 & 2 & 3 & 0 \\ 2 & 1 & 1 & 2 \\ 4 & 5 & 7 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 9 \\ 7 \\ 25 \end{pmatrix}$$

given that $\begin{pmatrix} 1 & 1 & 2 & 1 \end{pmatrix}^T = \begin{pmatrix} x & y & z & w \end{pmatrix}^T$ *is one solution.*

Note the matrix on the left is the same as the matrix in Example 9.2.14. Therefore, from Theorem 9.2.13, you will obtain all solutions to the above linear system in the form

$$z\begin{pmatrix} \frac{1}{3} \\ -\frac{5}{3} \\ 1 \\ 0 \end{pmatrix} + w\begin{pmatrix} -\frac{4}{3} \\ \frac{2}{3} \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 2 \\ 1 \end{pmatrix}.$$

9.3 Exercises

1. Study the definition of a linear transformation. State it from memory.
2. Show the map $T : \mathbb{R}^n \mapsto \mathbb{R}^m$ defined by $T(\mathbf{x}) = A\mathbf{x}$ where A is an $m \times n$ matrix and $\mathbf{x}$ is an $m \times 1$ column vector is a linear transformation.

3. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/3$.

4. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/4$.

5. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $-\pi/3$.

6. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $2\pi/3$.

7. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/12$. **Hint:** Note that $\pi/12 = \pi/3 - \pi/4$.

8. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $2\pi/3$ and then reflects across the x axis.

9. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/3$ and then reflects across the x axis.

10. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/4$ and then reflects across the x axis.

11. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/6$ and then reflects across the x axis followed by a reflection across the y axis.

12. Find the matrix for the linear transformation which reflects every vector in $\mathbb{R}^2$ across the x axis and then rotates every vector through an angle of $\pi/4$.

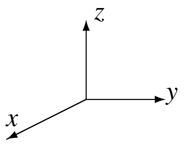
13. Find the matrix for the linear transformation which reflects every vector in $\mathbb{R}^2$ across the y axis and then rotates every vector through an angle of $\pi/4$.

14. Find the matrix for the linear transformation which reflects every vector in $\mathbb{R}^2$ across the x axis and then rotates every vector through an angle of $\pi/6$.

15. Find the matrix for the linear transformation which reflects every vector in $\mathbb{R}^2$ across the y axis and then rotates every vector through an angle of $\pi/6$.

16. Find the matrix for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $5\pi/12$. **Hint:** Note that $5\pi/12 = 2\pi/3 - \pi/4$.

17. Find the matrix of the linear transformation which rotates every vector in $\mathbb{R}^3$ counter clockwise about the z axis when viewed from the positive z axis through an angle of 30° and then reflects through the xy plane.



18. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1,-2,3)^T$.

19. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1,5,3)^T$.

20. Find the matrix for $\text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\mathbf{u} = (1,0,3)^T$.

21. 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.

22. Show that

$$\langle \mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v}), \mathbf{u} \rangle \equiv (\mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v}), \mathbf{u}) \equiv (\mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v})) \cdot \mathbf{u} = 0$$

and conclude every vector in $\mathbb{R}^n$ can be written as the sum of two vectors, one which is perpendicular and one which is parallel to the given vector.

23. Here are some descriptions of functions mapping $\mathbb{R}^n$ to $\mathbb{R}^n$.

    (a) T multiplies the j^{th} component of $\mathbf{x}$ by a nonzero number b.

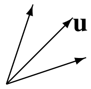
    (b) T replaces the i^{th} component of $\mathbf{x}$ with b times the j^{th} component added to the i^{th} component.

    (c) T switches two components.

    Show these functions are linear and describe their matrices.

24. In Problem 23, sketch the effects of the linear transformations on the unit square in $\mathbb{R}^2$. Give a geometric description of an arbitrary invertible matrix in terms of products of matrices of these special matrices in Problem 23.

25. Let $\mathbf{u} = (a,b)$ be a unit vector in $\mathbb{R}^2$. Find the matrix which reflects all vectors across this vector.



Hint: You might want to notice that $(a,b) = (\cos\theta, \sin\theta)$ for some θ. First rotate through $-\theta$. Next reflect through the x axis which is easy. Finally rotate through θ.

26. Let $\mathbf{u}$ be a unit vector. Show the linear transformation of the matrix $I - 2\mathbf{u}\mathbf{u}^T$ preserves all distances and satisfies

$$\left(I - 2\mathbf{u}\mathbf{u}^T\right)^T \left(I - 2\mathbf{u}\mathbf{u}^T\right) = I.$$

This matrix is called a Householder reflection. More generally, any matrix Q which satisfies $Q^T Q = QQ^T$ is called an orthogonal matrix. Show the linear transformation determined by an orthogonal matrix always preserves the length of a vector in $\mathbb{R}^n$. **Hint:** First either recall, depending on whether you have done Problem 51 on Page 108, or show that for any matrix A,

$$\langle A\mathbf{x}, \mathbf{y} \rangle = \left\langle \mathbf{x}, A^T \mathbf{y} \right\rangle$$

27. Suppose $|\mathbf{x}| = |\mathbf{y}|$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. The problem is to find an orthogonal transformation Q, (see Problem 26) which has the property that $Q\mathbf{x} = \mathbf{y}$ and $Q\mathbf{y} = \mathbf{x}$. Show

$$Q \equiv I - 2\frac{\mathbf{x} - \mathbf{y}}{|\mathbf{x} - \mathbf{y}|^2} (\mathbf{x} - \mathbf{y})^T$$

does what is desired.

28. Let $\mathbf{a}$ be a fixed vector. The function $T_{\mathbf{a}}$ defined by $T_{\mathbf{a}}\mathbf{v} = \mathbf{a} + \mathbf{v}$ has the effect of translating all vectors by adding $\mathbf{a}$. Show this is not a linear transformation. Explain why it is not possible to realize $T_{\mathbf{a}}$ in $\mathbb{R}^3$ by multiplying by a 3×3 matrix.

29. In spite of Problem 28 we can represent both translations and linear transformations by matrix multiplication at the expense of using higher dimensions. This is done by the homogeneous coordinates. I will illustrate in $\mathbb{R}^3$ where most interest in this is found. For each vector $\mathbf{v} = (v_1, v_2, v_3)^T$, consider the vector in $\mathbb{R}^4$ $(v_1, v_2, v_3, 1)^T$. What happens when you do

$$\begin{pmatrix} 1 & 0 & 0 & a_1 \\ 0 & 1 & 0 & a_2 \\ 0 & 0 & 1 & a_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ 1 \end{pmatrix}?$$

Describe how to consider both linear transformations and translations all at once by forming appropriate 4×4 matrices.

30. You want to add $\begin{pmatrix} 1, & 2, & 3 \end{pmatrix}$ to every point in $\mathbb{R}^3$ and then rotate about the z axis counter clockwise through an angle of 30°. Find what happens to the point $\begin{pmatrix} 1, & 1, & 1 \end{pmatrix}$.

31. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

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

32. Using Problem 31 find the general solution to the following linear system.

$$\begin{pmatrix} 1 & -1 & 2 \\ 1 & -2 & 1 \\ 3 & -4 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}.$$

33. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

$$\begin{pmatrix} 0 & -1 & 2 \\ 1 & -2 & 1 \\ 1 & -4 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

34. Using Problem 33 find the general solution to the following linear system.

$$\begin{pmatrix} 0 & -1 & 2 \\ 1 & -2 & 1 \\ 1 & -4 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

35. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

$$\begin{pmatrix} 1 & -1 & 2 \\ 1 & -2 & 0 \\ 3 & -4 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

36. Using Problem 35 find the general solution to the following linear system.

$$\begin{pmatrix} 1 & -1 & 2 \\ 1 & -2 & 0 \\ 3 & -4 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}.$$

37. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

$$\begin{pmatrix} 0 & -1 & 2 \\ 1 & 0 & 1 \\ 1 & -2 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

38. Using Problem 37 find the general solution to the following linear system.

$$\begin{pmatrix} 0 & -1 & 2 \\ 1 & 0 & 1 \\ 1 & -2 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

39. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

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

40. Using Problem 39 find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & -1 & 1 & 0 \\ 3 & -1 & 3 & 2 \\ 3 & 3 & 0 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \\ 3 \end{pmatrix}.$$

41. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

$$\begin{pmatrix} 1 & 1 & 0 & 1 \\ 2 & 1 & 1 & 2 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

42. Using Problem 41 find the general solution to the following linear system.

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

43. 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.

44. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

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

45. Using Problem 44 find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 \\ 3 & 1 & 1 & 2 \\ 3 & 3 & 0 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \\ 3 \end{pmatrix}.$$

46. Write the solution set of the following system as the span of vectors and find a basis for the solution space of the following system.

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

47. Using Problem 46 find the general solution to the following linear system.

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

48. 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 $A\mathbf{x} = \mathbf{0}$. It is the solution space to the system $A\mathbf{x} = \mathbf{0}$.

49. Using Problem 48, find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 11 \\ 7 \\ 18 \\ 7 \end{pmatrix}$$

50. Using Problem 48, find the general solution to the following linear system.

$$\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 2 & 1 & 1 & 2 \\ 1 & 4 & 4 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 6 \\ 7 \\ 13 \\ 7 \end{pmatrix}$$

51. 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.

52. Show that if A is an $m \times n$ matrix, then $\ker(A)$ is a subspace.

53. Verify the linear transformation determined by the matrix of 9.2 maps $\mathbb{R}^3$ onto $\mathbb{R}^2$ but the linear transformation determined by this matrix is not one to one.

54. You are given a linear transformation $T : \mathbf{R}^n \to \mathbf{R}^m$ and you know that

$$T\mathbf{a}_i = \mathbf{b}_i$$

where $\begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}^{-1}$ exists. Show that the matrix A of T with respect to the usual basis vectors ($T\mathbf{x} = A\mathbf{x}$) must be of the form

$$\begin{pmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_n \end{pmatrix} \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}^{-1}$$

55. You have a linear transformation T and

$$T \begin{pmatrix} 1 \\ 2 \\ -6 \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \\ 3 \end{pmatrix}, T \begin{pmatrix} -1 \\ -1 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 5 \end{pmatrix}, T \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \\ -2 \end{pmatrix}$$

Find the matrix of T. That is find A such that $T\mathbf{x} = A\mathbf{x}$.

56. You have a linear transformation T and

$$T \begin{pmatrix} 1 \\ 1 \\ -8 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}, T \begin{pmatrix} -1 \\ 0 \\ 6 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ 1 \end{pmatrix}, T \begin{pmatrix} 0 \\ -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 6 \\ 1 \\ -1 \end{pmatrix}$$

Find the matrix of T. That is find A such that $T\mathbf{x} = A\mathbf{x}$.

57. You have a linear transformation T and

$$T\begin{pmatrix} 1 \\ 3 \\ -7 \end{pmatrix} = \begin{pmatrix} -3 \\ 1 \\ 3 \end{pmatrix}, T\begin{pmatrix} -1 \\ -2 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ -3 \end{pmatrix}, T\begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \\ -3 \end{pmatrix}$$

Find the matrix of T. That is find A such that $T\mathbf{x} = A\mathbf{x}$.

58. You have a linear transformation T and

$$T\begin{pmatrix} 1 \\ 1 \\ -7 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}, T\begin{pmatrix} -1 \\ 0 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, T\begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ -1 \end{pmatrix}$$

Find the matrix of T. That is find A such that $T\mathbf{x} = A\mathbf{x}$.

59. You have a linear transformation T and

$$T\begin{pmatrix} 1 \\ 2 \\ -18 \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \\ 5 \end{pmatrix}, T\begin{pmatrix} -1 \\ -1 \\ 15 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 5 \end{pmatrix}, T\begin{pmatrix} 0 \\ -1 \\ 4 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ -2 \end{pmatrix}$$

Find the matrix of T. That is find A such that $T\mathbf{x} = A\mathbf{x}$.

60. Suppose $\begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}^{-1}$ exists where each $\mathbf{a}_j \in \mathbb{R}^n$ and let vectors

$$\{\mathbf{b}_1, \cdots, \mathbf{b}_n\}$$

in $\mathbb{R}^m$ be given. Show that there **always** exists a linear transformation T such that $T\mathbf{a}_i = \mathbf{b}_i$.

61. Let $V \neq \{\mathbf{0}\}$ be a subspace of $\mathbb{F}^n$. Show directly that there exists a subset of V $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ such that $\operatorname{span}(\mathbf{v}_1, \cdots, \mathbf{v}_r) = V$. Out of all such spanning sets, let the dimension of V be the smallest number of vectors in any spanning set. The spanning set having this smallest number of vectors will be called a minimal spanning set. Thus it is automatically the case that any two of these minimal spanning sets contain the same number of vectors. Now show that $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ is a minimal spanning set if and only if $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ is a basis. This gives a little different way to define a basis which has the advantage of making it obvious that any two bases have the same number of vectors. Hence the dimension is well defined.

62. In general, if V is a subspace of $\mathbb{F}^n$ and W is a subspace of $\mathbb{F}^m$, a function $T : V \to W$ is called a linear transformation if whenever $\mathbf{x}, \mathbf{y}$ are vectors in V and a, b are scalars,

$$T(a\mathbf{x} + b\mathbf{y}) = aT\mathbf{x} + bT\mathbf{y}$$

this is more general than having T be defined only on all of $\mathbb{F}^n$. Why must V be a subspace in order for this to make sense?

Chapter 10

A Few Factorizations

10.1 Definition Of An *LU* factorization

An *LU* factorization of a matrix involves writing the given matrix as the product of a lower triangular matrix which has the main diagonal consisting entirely of ones *L*, and an upper triangular matrix *U* in the indicated order. This is the version discussed here but it is sometimes the case that the *L* has numbers other than 1 down the main diagonal. It is still a useful concept. The *L* goes with "lower" and the *U* with "upper". It turns out many matrices can be written in this way and when this is possible, people get excited about slick ways of solving the system of equations, $A\mathbf{x} = \mathbf{y}$. It is for this reason that you want to study the *LU* factorization. It allows you to work only with triangular matrices. It turns out that it takes about half as many operations to obtain an *LU* factorization as it does to find the row reduced echelon form.

First it should be noted not all matrices have an *LU* factorization and so we will emphasize the techniques for achieving it rather than formal proofs.

Example 10.1.1 *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* factorization. This is what we mean above by saying the method lacks generality.

10.2 Finding An *LU* Factorization By Inspection

Which matrices have an *LU* factorization? It turns out it is those whose row reduced echelon form can be achieved without switching rows and which only involve row operations of type 3 in which row j is replaced with a multiple of row i added to row j for $i < j$.

Example 10.2.1 *Find an LU factorization of* $A = \begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 3 & 2 & 1 \\ 2 & 3 & 4 & 0 \end{pmatrix}$.

One way to find the *LU* factorization is to simply look for it directly. You need

$$\begin{pmatrix} 1 & 2 & 0 & 2 \\ 1 & 3 & 2 & 1 \\ 2 & 3 & 4 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ x & 1 & 0 \\ y & z & 1 \end{pmatrix} \begin{pmatrix} a & d & h & j \\ 0 & b & e & i \\ 0 & 0 & c & f \end{pmatrix}.$$

Then multiplying these you get

$$\begin{pmatrix} a & d & h & j \\ xa & xd+b & xh+e & xj+i \\ ya & yd+zb & yh+ze+c & yj+iz+f \end{pmatrix}$$

and so you can now tell what the various quantities equal. From the first column, you need $a = 1, x = 1, y = 2$. Now go to the second column. You need $d = 2, xd + b = 3$ so $b = 1, yd + zb = 3$ so $z = -1$. From the third column, $h = 0, e = 2, c = 6$. Now from the fourth column, $j = 2, i = -1, f = -5$. Therefore, an *LU* factorization is

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

You can check whether you got it right by simply multiplying these two.

10.3 Using Multipliers To Find An *LU* Factorization

There is also a convenient procedure for finding an *LU* factorization. It turns out that it is only necessary to keep track of the **multipliers** which are used to row reduce to upper triangular form. This procedure is described in the following examples.

Example 10.3.1 *Find an LU factorization for* $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & -4 \\ 1 & 5 & 2 \end{pmatrix}$

Write the matrix next to the identity matrix as shown.

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

The process involves doing row operations to the matrix on the right while simultaneously updating successive columns of the matrix on the left. First take -2 times the first row and add to the second in the matrix on the right.

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

Note the way we updated the matrix on the left. We put a 2 in the second entry of the first column because we used -2 times the first row added to the second row. Now replace the third row in the matrix on the right by -1 times the first row added to the third. Notice that the product of the two matrices is unchanged and equals the original matrix. This is because a row operation was done on the original matrix to get the matrix on the right and then on the left, it was multiplied by an elementary matrix which "undid" the row operation which was done.

The next step is

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

Again, the product is unchanged because we just did and then undid a row operation. Finally, we will add the second row to the bottom row and make the following changes

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -10 \\ 0 & 0 & -11 \end{pmatrix}.$$

At this point, we stop because the matrix on the right is upper triangular. An *LU* factorization is the above.

The justification for this gimmick will be given later in a more general context.

Example 10.3.2 *Find an LU factorization for* $A = \begin{pmatrix} 1 & 2 & 1 & 2 & 1 \\ 2 & 0 & 2 & 1 & 1 \\ 2 & 3 & 1 & 3 & 2 \\ 1 & 0 & 1 & 1 & 2 \end{pmatrix}.$

►►

We will use the same procedure as above. However, this time we will do everything for one column at a time. First multiply the first row by (-1) and then add to the last row. Next take (-2) times the first and add to the second and then (-2) times the first and add to the third.

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

This finishes the first column of L and the first column of U. As in the above, what happened was this. Lots of row operations were done and then these were undone by multiplying by the matrix on the left. Thus the above product equals the original matrix. 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 factorization 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 factorization is of interest in solving systems of equations.

10.4 Solving Systems Using An *LU* Factorization

One reason people care about the LU factorization is it allows the quick solution of systems of equations. Here is an example.

Example 10.4.1 *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 factorization and it turns out that the LU factorization can give the solution quickly. Here is how. The following is an LU factorization 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 $\mathbf{y} = \begin{pmatrix} 1 & -2 & 2 \end{pmatrix}^T$. 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

$$\mathbf{x} = \begin{pmatrix} \frac{7}{5}t - \frac{3}{5} & \frac{9}{5} - \frac{11}{5}t & t & -1 \end{pmatrix}^T, t \in \mathbb{R}.$$

10.5 Justification For The Multiplier Method

Why does the multiplier method work for finding the LU factorization? 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.

Lemma 10.5.1 *Let L be a lower (upper) triangular matrix $m \times m$ which has ones down the main diagonal. Then L^{-1} also is a lower (upper) triangular matrix which has ones down the main diagonal. In the case that L is of the form*

$$L = \begin{pmatrix} 1 & & & \\ a_1 & 1 & & \\ \vdots & & \ddots & \\ a_n & & & 1 \end{pmatrix} \tag{10.1}$$

where all entries are zero except for the left column and main diagonal, it is also the case that L^{-1} is obtained from L by simply multiplying each entry below the main diagonal in L with -1. The same is true if the single nonzero column is in another position.

Proof: Consider the usual setup for finding the inverse $\begin{pmatrix} L & I \end{pmatrix}$. Then each row operation done to L to reduce to row reduced echelon form results in changing only the entries in I below the main diagonal. In the special case of L given in 10.1 or the single nonzero column is in another position, multiplication by -1 as described in the lemma clearly results in L^{-1} . ■

For a simple illustration of the last claim,

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

Now let A be an $m \times n$ matrix, say

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

and assume A can be row reduced to an upper triangular form using only row operation 3. Thus, in particular, $a_{11} \neq 0$. Multiply on the left by $E_1 =$

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{m1}}{a_{11}} & 0 & \cdots & 1 \end{pmatrix}$$

This is the product of elementary matrices which make modifications in the first column only. It is equivalent to taking $-a_{21}/a_{11}$ times the first row and adding to the second. Then taking $-a_{31}/a_{11}$ times the first row and adding to the third and so forth. The quotients in the first column of the above matrix are the multipliers. Thus the result is of the form

$$E_1 A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a'_{1n} \\ 0 & a'_{22} & \cdots & a'_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & a'_{m2} & \cdots & a'_{mn} \end{pmatrix}$$

By assumption, $a'_{22} \neq 0$ and so it is possible to use this entry to zero out all the entries below it in the matrix on the right by multiplication by a matrix of the form $E_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & E \end{pmatrix}$ where E is an $(m-1) \times (m-1)$ matrix of the form

$$E = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\frac{a'_{32}}{a'_{22}} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{a'_{m2}}{a'_{22}} & 0 & \cdots & 1 \end{pmatrix}$$

Again, the entries in the first column below the 1 are the multipliers. Continuing this way, zeroing out the entries below the diagonal entries, finally leads to

$$E_{m-1} E_{n-2} \cdots E_1 A = U$$

where U is upper triangular. Each E_j has all ones down the main diagonal and is lower triangular. Now multiply both sides by the inverses of the E_j in the reverse order. This yields

$$A = E_1^{-1} E_2^{-1} \cdots E_{m-1}^{-1} U$$

By Lemma 10.5.1, this implies that the product of those E_j^{-1} is a lower triangular matrix having all ones down the main diagonal.

The above discussion and lemma gives the justification for the multiplier method. The expressions

$$-a_{21}/a_{11}, -a_{31}/a_{11}, \cdots - a_{m1}/a_{11}$$

denoted respectively by $M_{21}, \cdots, M_{m1}$ to save notation which were obtained in building E_1 are the multipliers**.** Then according to the lemma, to find E_1^{-1} you simply write

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ -M_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -M_{m1} & 0 & \cdots & 1 \end{pmatrix}$$

Similar considerations apply to the other E_j^{-1}. Thus L is a product of the form

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ -M_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -M_{m1} & 0 & \cdots & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & -M_{m(m-1)} & 1 \end{pmatrix}$$

each factor having at most one nonzero column, the position of which moves from left to right in scanning the above product of matrices from left to right. It follows from Theorem 8.1.10 about the effect of multiplying on the left by an elementary matrix that the above product is of the form

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -M_{21} & 1 & \cdots & 0 & 0 \\ \vdots & -M_{32} & \ddots & \vdots & \vdots \\ -M_{(M-1)1} & \vdots & \cdots & 1 & 0 \\ -M_{M1} & -M_{M2} & \cdots & -M_{MM-1} & 1 \end{pmatrix}$$

In words, beginning at the left column and moving toward the right, you simply insert, into the corresponding position in the identity matrix, -1 times the multiplier which was used to zero out an entry in that position below the main diagonal in A, while retaining the main diagonal which consists entirely of ones. This is L.

10.6 The *PLU* Factorization

As indicated above, some matrices don't have an *LU* factorization. Here is an example.

$$M = \begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 2 & 3 & 0 \\ 4 & 3 & 1 & 1 \end{pmatrix} \tag{10.2}$$

In this case, there is another factorization which is useful called a *PLU* factorization. Here P is a permutation matrix.

Example 10.6.1 *Find a PLU factorization for the above matrix in 10.2.*

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 matrix on the right in such a way as to obtain an upper triangular matrix. Therefore, consider the original matrix with the bottom two rows switched.

$$M' = \begin{pmatrix} 1 & 2 & 3 & 2 \\ 4 & 3 & 1 & 1 \\ 1 & 2 & 3 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 2 & 3 & 0 \\ 4 & 3 & 1 & 1 \end{pmatrix} = PM$$

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 matrix on the right is upper triangular and so the LU factorization of the matrix M' has been obtained above.

Thus $M' = PM = LU$ where L and U are given above. Notice that $P^2 = I$ and 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 factorization of a given matrix even though there isn't always an LU factorization.

Example 10.6.2 *Use the PLU factorization 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,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

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

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 \end{pmatrix}^T =$$

$$= \begin{pmatrix} \frac{7}{5}t + \frac{1}{5} & \frac{9}{10} - \frac{11}{5}t & t & -\frac{1}{2} \end{pmatrix}^T : t \in \mathbb{R}.$$

10.7 The *QR* Factorization

As pointed out above, the *LU* factorization is not a mathematically respectable thing because it does not always exist. There is another factorization which does always exist. Much more can be said about it than I will say here. At this time, I will only deal with real matrices and so the inner product will be the usual real dot product. Letting A be an $m \times n$ real matrix and letting $(\cdot, \cdot)$ denote the usual real inner product,

$$\begin{aligned} (A\mathbf{x}, \mathbf{y}) &= \sum_i (A\mathbf{x})_i y_i = \sum_i \sum_j A_{ij} x_j y_i = \sum_j \sum_i \left(A^T\right)_{ji} y_i x_j \\ &= \sum_j \left(A^T \mathbf{y}\right)_j x_j = \left(\mathbf{x}, A^T \mathbf{y}\right) \end{aligned}$$

Thus, when you take the matrix across the comma, you replace with a transpose.

Definition 10.7.1 *An $n \times n$ real matrix Q is called an orthogonal matrix if*

$$QQ^T = Q^T Q = I.$$

Thus an orthogonal matrix is one whose inverse is equal to its transpose.

From the above observation,

$$|Q\mathbf{x}|^2 = (Q\mathbf{x}, Q\mathbf{x}) = \left(\mathbf{x}, Q^T Q\mathbf{x}\right) = (\mathbf{x}, I\mathbf{x}) = (\mathbf{x}, \mathbf{x}) = |\mathbf{x}|^2$$

This shows that orthogonal transformations preserve distances. Conversely you can also show that if you have a matrix which does preserve distances, then it must be orthogonal.

Example 10.7.2 *One of the most important examples of an orthogonal matrix is the so called Householder matrix. You have $\mathbf{v}$ a unit vector and you form the matrix*

$$I - 2\mathbf{v}\mathbf{v}^T$$

This is an orthogonal matrix which is also symmetric. To see this, you use the rules of matrix operations.

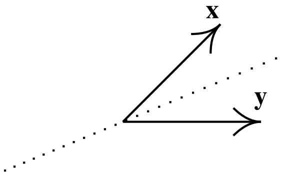
$$\begin{aligned} \left(I - 2\mathbf{v}\mathbf{v}^T\right)^T &= I^T - \left(2\mathbf{v}\mathbf{v}^T\right)^T \\ &= I - 2\mathbf{v}\mathbf{v}^T \end{aligned}$$

so it is symmetric. Now to show it is orthogonal,

$$\begin{aligned}\left(I-2\mathbf{v}\mathbf{v}^T\right)\left(I-2\mathbf{v}\mathbf{v}^T\right) &= I-2\mathbf{v}\mathbf{v}^T-2\mathbf{v}\mathbf{v}^T+4\mathbf{v}\mathbf{v}^T\mathbf{v}\mathbf{v}^T \\ &= I-4\mathbf{v}\mathbf{v}^T+4\mathbf{v}\mathbf{v}^T=I\end{aligned}$$

because $\mathbf{v}^T\mathbf{v}=\mathbf{v}\cdot\mathbf{v}=|\mathbf{v}|^2=1$. *Therefore, this is an example of an orthogonal matrix.*

Next consider the problem illustrated in the following picture.



Find an orthogonal matrix Q which switches the two vectors taking $\mathbf{x}$ to $\mathbf{y}$ and $\mathbf{y}$ to $\mathbf{x}$.

Procedure 10.7.3 *Given two vectors* $\mathbf{x},\mathbf{y}$ *such that* $|\mathbf{x}|=|\mathbf{y}|\neq 0$ *but* $\mathbf{x}\neq\mathbf{y}$ *and you want an orthogonal matrix Q such that* $Q\mathbf{x}=\mathbf{y}$ *and* $Q\mathbf{y}=\mathbf{x}$. *The thing which works is the Householder matrix*

$$Q\equiv I-2\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^2}(\mathbf{x}-\mathbf{y})^T$$

Here is why this works.

$$\begin{aligned}Q(\mathbf{x}-\mathbf{y}) &= (\mathbf{x}-\mathbf{y})-2\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^2}(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) \\ &= (\mathbf{x}-\mathbf{y})-2\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^2}|\mathbf{x}-\mathbf{y}|^2=\mathbf{y}-\mathbf{x}\end{aligned}$$

$$\begin{aligned}Q(\mathbf{x}+\mathbf{y}) &= (\mathbf{x}+\mathbf{y})-2\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^2}(\mathbf{x}-\mathbf{y})^T(\mathbf{x}+\mathbf{y}) \\ &= (\mathbf{x}+\mathbf{y})-2\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^2}((\mathbf{x}-\mathbf{y})\cdot(\mathbf{x}+\mathbf{y})) \\ &= (\mathbf{x}+\mathbf{y})-2\frac{\mathbf{x}-\mathbf{y}}{|\mathbf{x}-\mathbf{y}|^2}\left(|\mathbf{x}|^2-|\mathbf{y}|^2\right)=\mathbf{x}+\mathbf{y}\end{aligned}$$

Hence

$$\begin{aligned}Q\mathbf{x}+Q\mathbf{y} &= \mathbf{x}+\mathbf{y} \\ Q\mathbf{x}-Q\mathbf{y} &= \mathbf{y}-\mathbf{x}\end{aligned}$$

Adding these equations, $2Q\mathbf{x}=2\mathbf{y}$ and subtracting them yields $2Q\mathbf{y}=2\mathbf{x}$.

Definition 10.7.4 *Let A be an $m\times n$ matrix. Then a QR factorization of A consists of two matrices, Q orthogonal and R upper triangular or in other words equal to zero below the main diagonal such that $A=QR$.*

With the solution to this simple problem, here is how to obtain a QR factorization for any matrix A. Let

$$A=(\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_n)$$

where the $\mathbf{a}_i$ are the columns. If $\mathbf{a}_1=\mathbf{0}$, let $Q_1=I$. If $\mathbf{a}_1\neq\mathbf{0}$, let

$$\mathbf{b}\equiv\begin{pmatrix}|\mathbf{a}_1|\\0\\\vdots\\0\end{pmatrix}$$

and form the Householder matrix

$$Q_1\equiv I-2\frac{(\mathbf{a}_1-\mathbf{b})}{|\mathbf{a}_1-\mathbf{b}|^2}(\mathbf{a}_1-\mathbf{b})^T$$

As in the above problem $Q_1\mathbf{a}_1=\mathbf{b}$ and so $Q_1A=\begin{pmatrix}|\mathbf{a}_1| & *\\\mathbf{0} & A_2\end{pmatrix}$ where A_2 is a $m-1\times n-1$ matrix. Now find in the same way as was just done a $n-1\times n-1$ matrix $\widehat{Q}_2$ such that

$$\widehat{Q}_2A_2=\begin{pmatrix}* & *\\\mathbf{0} & A_3\end{pmatrix}$$

Let $Q_2\equiv\begin{pmatrix}1 & 0\\\mathbf{0} & \widehat{Q}_2\end{pmatrix}$.Then

$$Q_2Q_1A=\begin{pmatrix}1 & 0\\\mathbf{0} & \widehat{Q}_2\end{pmatrix}\begin{pmatrix}|\mathbf{a}_1| & *\\\mathbf{0} & A_2\end{pmatrix}=\begin{pmatrix}|\mathbf{a}_1| & * & *\\\vdots & * & *\\0 & \mathbf{0} & A_3\end{pmatrix}$$

Continuing this way until the result is upper triangular, you get a sequence of orthogonal matrices $Q_pQ_{p-1}\cdots Q_1$ such that

$$Q_pQ_{p-1}\cdots Q_1A=R \tag{10.3}$$

where R is upper triangular.

Now if Q_1 and Q_2 are orthogonal, then from properties of matrix multiplication,

$$Q_1Q_2(Q_1Q_2)^T=Q_1Q_2Q_2^TQ_1^T=Q_1IQ_1^T=I$$

and similarly

$$(Q_1Q_2)^TQ_1Q_2=I.$$

Thus the product of orthogonal matrices is orthogonal. Also the transpose of an orthogonal matrix is orthogonal directly from the definition. Therefore, from 10.3

$$A=(Q_pQ_{p-1}\cdots Q_1)^TR\equiv QR,$$

where Q is orthogonal. This suggests the proof of the following theorem.

Theorem 10.7.5 *Let A be any real $m \times n$ matrix. Then there exists an orthogonal matrix Q and an upper triangular matrix R having nonnegative entries down the main diagonal such that*

$$A = QR$$

and this factorization can be accomplished in a systematic manner.

Proof: The theorem is clearly true if A is a $1 \times m$ matrix. Suppose it is true for A an $n \times m$ matrix. Thus, if A is any $n \times m$ matrix, there exists an orthogonal matrix Q such that

$$QA = R$$

where R is upper triangular. Suppose A is an $(n+1) \times m$ matrix. Then, as indicated above, there exists an orthogonal $(n+1) \times (n+1)$ matrix Q_1 such that

$$Q_1 A = \begin{pmatrix} a & \mathbf{b} \\ \mathbf{0} & A_1 \end{pmatrix}$$

where A_1 is $n \times m - 1$ or else, in case $m = 1$, the right side is of the form

$$\begin{pmatrix} a \\ \mathbf{0} \end{pmatrix}$$

and in this case, the conclusion of the theorem follows. If $m - 1 \geq 1$, then by induction, there exists Q_2 an orthogonal $n \times n$ matrix such that $Q_2 A_1 = R_1$. Then

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_2 \end{pmatrix} Q_1 A = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_2 \end{pmatrix} \begin{pmatrix} a & \mathbf{b} \\ \mathbf{0} & A_1 \end{pmatrix} = \begin{pmatrix} a & \mathbf{b} \\ \mathbf{0} & R_1 \end{pmatrix} \equiv R$$

Since $\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_2 \end{pmatrix} Q_1$ is orthogonal, being the product of two orthogonal matrices, the conclusion follows. ■

►►

10.8 MATLAB And Factorizations

MATLAB can find an LU factorizaton of a matrix. Here is an example.

```
>>A=[1,3,2,4;-5,7,2,3;2,3,7,11;1,2,3,4]; [L,U,P]=lu(sym(A))
```

This will give a lower triangular matrix L with ones down the diagonal, an upper triangular matrix U, and a permutation matrix P such that $PA = LU$. Of course if the matrix A has an LU factorization, you will have $P = I$. This was the case here. After you have typed the above, you press enter and you get the following listed in a column. If you just type lu(A), then the answer will come out in decimals.

```
L=                    U=                          P=
[ 1, 0, 0, 0]         [ 1, 3, 2, 4]               1 0 0 0
[ -5, 1, 0, 0]        [ 0, 22, 12, 23]            0 1 0 0
[ 2, -3/22, 1, 0]     [ 0, 0, 51/11, 135/22]      0 0 1 0
[ 1, -1/22, 1/3, 1]   [ 0, 0, 0, -1]              0 0 0 1
```

MATLAB can also find the much more interesting *QR* factorization. Here is how to do it with an example.

```
>>A=[1,2,3;4,2,1;2,6,7;1,-4,2];[Q,R]=qr(A)
```

Then press enter and you get the following.

```
Q=                                   R=
-0.2132 0.1756 -0.2593 0.9255        -4.6904 -3.8376 -4.9036
-0.8528 -0.1892 0.4862 -0.0244       0 6.7285 3.4453
-0.4264 0.6485 -0.5139 -0.3653       0 0 -5.2043
-0.2132 -0.7161 -0.6575 -0.0974      0 0 0
```

If you want to see something horrible, replace qr(A) with qr(sym(A)). This way it gives the exact values. You can check your work by >>Q*Q' and press enter. The Q' means the conjugate transpose. Since everything is real here, this is just the transpose.

10.9 Exercises

1. Find an *LU* factorization of $\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 3 \\ 1 & 2 & 3 \end{pmatrix}$.

2. Find an *LU* factorization of $\begin{pmatrix} 1 & 2 & 3 & 2 \\ 1 & 3 & 2 & 1 \\ 5 & 0 & 1 & 3 \end{pmatrix}$.

3. Find an *LU* factorization of the matrix $\begin{pmatrix} 1 & -2 & -5 & 0 \\ -2 & 5 & 11 & 3 \\ 3 & -6 & -15 & 1 \end{pmatrix}$.

4. Find an *LU* factorization of the matrix $\begin{pmatrix} 1 & -1 & -3 & -1 \\ -1 & 2 & 4 & 3 \\ 2 & -3 & -7 & -3 \end{pmatrix}$.

5. Find an *LU* factorization of the matrix $\begin{pmatrix} 1 & -3 & -4 & -3 \\ -3 & 10 & 10 & 10 \\ 1 & -6 & 2 & -5 \end{pmatrix}$.

6. Find an *LU* factorization of the matrix $\begin{pmatrix} 1 & 3 & 1 & -1 \\ 3 & 10 & 8 & -1 \\ 2 & 5 & -3 & -3 \end{pmatrix}$.

7. Find an *LU* factorization of the matrix $\begin{pmatrix} 3 & -2 & 1 \\ 9 & -8 & 6 \\ -6 & 2 & 2 \\ 3 & 2 & -7 \end{pmatrix}$.

8. Find an *LU* factorization of the matrix $\begin{pmatrix} -3 & -1 & 3 \\ 9 & 9 & -12 \\ 3 & 19 & -16 \\ 12 & 40 & -26 \end{pmatrix}$.

9. Find an *LU* factorization of the matrix $\begin{pmatrix} -1 & -3 & -1 \\ 1 & 3 & 0 \\ 3 & 9 & 0 \\ 4 & 12 & 16 \end{pmatrix}$.

10. Find the *LU* factorization of the coefficient matrix using Dolittle's method and use it to solve the system of equations.

$$\begin{aligned} x+2y &= 5 \\ 2x+3y &= 6 \end{aligned}$$

11. Find the *LU* factorization of the coefficient matrix using Dolittle's method and use it to solve the system of equations.

$$\begin{aligned} x+2y+z &= 1 \\ y+3z &= 2 \\ 2x+3y &= 6 \end{aligned}$$

12. Find the *LU* factorization of the coefficient matrix using Dolittle's method and use it to solve the system of equations.

$$\begin{aligned} x+2y+3z &= 5 \\ 2x+3y+z &= 6 \\ x-y+z &= 2 \end{aligned}$$

13. Find the *LU* factorization of the coefficient matrix using Dolittle's method and use it to solve the system of equations.

$$\begin{aligned} x+2y+3z &= 5 \\ 2x+3y+z &= 6 \\ 3x+5y+4z &= 11 \end{aligned}$$

14. Is there only one *LU* factorization for a given matrix? **Hint:** Consider the equation

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

Look for all possible *LU* factorizations.

15. Find a *PLU* factorization of $\begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 2 & 1 & 1 \end{pmatrix}$.

16. Find a *PLU* factorization of $\begin{pmatrix} 1 & 2 & 1 & 2 & 1 \\ 2 & 4 & 2 & 4 & 1 \\ 1 & 2 & 1 & 3 & 2 \end{pmatrix}$.

17. Find a *PLU* factorization of $\begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 2 & 4 & 1 \\ 3 & 2 & 1 \end{pmatrix}$.

18. Find a *PLU* factorization of $\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 1 \\ 1 & 0 & 2 \\ 2 & 2 & 1 \end{pmatrix}$ and use it to solve the systems

    a. $\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 1 \\ 1 & 0 & 2 \\ 2 & 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 1 \end{pmatrix}$ b. $\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 1 \\ 1 & 0 & 2 \\ 2 & 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}$

19. Find a *PLU* factorization of $\begin{pmatrix} 0 & 2 & 1 & 2 \\ 2 & 1 & -2 & 0 \\ 2 & 3 & -1 & 2 \end{pmatrix}$ and use it to solve the systems

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

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

20. Find a *QR* factorization for the matrix

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

21. Find a *QR* factorization for the matrix

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

22. If you had a QR factorization, $A = QR$, describe how you could use it to solve the equation $A\mathbf{x} = \mathbf{b}$. This is not usually the way people solve this equation. However, the QR factorization is of great importance in certain other problems, especially in finding eigenvalues and eigenvectors.

23. In this problem, is another explanation of the LU factorization. Let

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

Review how to take the inverse of an elementary matrix. Then we can do the following.

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

Next

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

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

Next

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

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

Using this example, describe why, when a matrix can be reduced to echelon form using only row operation 3, then it has an LU factorization.

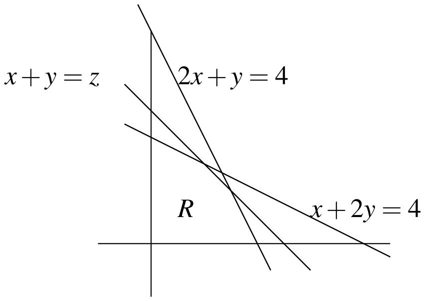
Chapter 11

Linear Programming

11.1 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 patties to use in one week and a limitless supply of special sauce, lettuce, tomatoes, 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 teenagers from the local high school, involves two patties, lots of delicious sauce, condiments galore, and a divider between the two patties. 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 patties is 9000 and so the number of patty used is $x+2y$. This number must satisfy $x+2y \leq 9000$ because there are only 9000 patty 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 patty. 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.

11.2 The Simplex Tableau

Here is some notation.

Definition 11.2.1 *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,

$$\mathbf{x} \geq \mathbf{0}, A\mathbf{x} = \mathbf{b} \tag{11.1}$$

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 referred 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 $\left\{\mathbf{a}_{j_1}, \cdots, \mathbf{a}_{j_m}\right\}$ 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

$$\begin{pmatrix} A & \mathbf{0} & \mathbf{b} \\ -\mathbf{c} & 1 & 0 \end{pmatrix} \tag{11.2}$$

Permute the columns and variables on the left if necessary to write the above in the form

$$\begin{pmatrix} B & F & \mathbf{0} \\ -\mathbf{c}_B & -\mathbf{c}_F & 1 \end{pmatrix} \begin{pmatrix} \mathbf{x}_B \\ \mathbf{x}_F \\ z \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ 0 \end{pmatrix} \tag{11.3}$$

or equivalently in the augmented matrix form keeping track of the variables on the bottom as

$$\begin{pmatrix} B & F & \mathbf{0} & \mathbf{b} \\ -\mathbf{c}_B & -\mathbf{c}_F & 1 & 0 \\ \mathbf{x}_B & \mathbf{x}_F & 0 & 0 \end{pmatrix}. \tag{11.4}$$

Here B pertains to the variables $x_{i_1}, \cdots, x_{j_m}$ and is an $m \times m$ matrix with linearly independent columns, $\left\{\mathbf{a}_{j_1}, \cdots, \mathbf{a}_{j_m}\right\}$, 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

$$\mathbf{x}_B^0 = B^{-1}\mathbf{b} \geq \mathbf{0}. \tag{11.5}$$

This is very important to observe. $B^{-1}\mathbf{b} \geq \mathbf{0}$! This is by the assumption that $\mathbf{x}^0 \geq \mathbf{0}$.

Do row operations on the top part of the matrix

$$\begin{pmatrix} B & F & \mathbf{0} & \mathbf{b} \\ -\mathbf{c}_B & -\mathbf{c}_F & 1 & 0 \end{pmatrix} \tag{11.6}$$

and obtain its row reduced echelon form. Then after these row operations the above becomes

$$\begin{pmatrix} I & B^{-1}F & \mathbf{0} & B^{-1}\mathbf{b} \\ -\mathbf{c}_B & -\mathbf{c}_F & 1 & 0 \end{pmatrix}. \tag{11.7}$$

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

$$\begin{align} & \begin{pmatrix} I & B^{-1}F & \mathbf{0} & B^{-1}\mathbf{b} \\ \mathbf{0} & \mathbf{c}_B B^{-1}F' - \mathbf{c}_F & 1 & \mathbf{c}_B B^{-1}\mathbf{b} \end{pmatrix} \tag{11.8} \\ = & \begin{pmatrix} I & B^{-1}F & \mathbf{0} & B^{-1}\mathbf{b} \\ \mathbf{0} & \mathbf{c}_B B^{-1}F' - \mathbf{c}_F & 1 & \mathbf{c}_B \mathbf{x}_B^0 \end{pmatrix} \notag \\ = & \begin{pmatrix} I & B^{-1}F & \mathbf{0} & B^{-1}\mathbf{b} \\ \mathbf{0} & \mathbf{c}_B B^{-1}F - \mathbf{c}_F & 1 & z^0 \end{pmatrix} \tag{11.9} \end{align}$$

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 11.6 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 11.9 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 11.9 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.

The idea is to obtain an augmented matrix for the constraints such that obvious solutions are also feasible. Then there is an algorithm, to be presented later, which takes you from one obvious feasible solution to another until you obtain the maximum.

Example 11.2.2 *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,

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 10 \\ x_1 + 2x_2 - x_4 &= 2 \\ 2x_1 + x_2 + x_5 &= 6 \end{aligned} .$$

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 11.2, $\begin{pmatrix} A & \mathbf{0} & \mathbf{b} \\ -\mathbf{c} & 1 & 0 \end{pmatrix}$ 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 11.2.3 *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. When you let $x_1, x_2 = 0$, you end up having $x_4 = -6$ which is negative. Consider the second column and select the 2 as a pivot to zero out that which is above and below the 2.

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

This one is good. When you let $x_1 = x_4 = 0$, you find that $x_2 = 3, x_3 = 4, x_5 = 3$. 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 in this larger matrix. 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.

11.3 The Simplex Algorithm

11.3.1 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 $\mathbf{b} \geq \mathbf{0}$.

$$\begin{pmatrix} I & F & \mathbf{0} & \mathbf{b} \\ \mathbf{0} & \mathbf{c} & 1 & z^0 \end{pmatrix} \tag{11.10}$$

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

$$\begin{aligned}\left(\mathbf{x}'_B\right)_k &= b_k - \sum_j F_{kj}\left(\mathbf{x}_F\right)_j \\ &= b_k - F_{ki}x_i \geq 0\end{aligned}$$

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

$$z = (-c_i)x_i + z^0 \tag{11.11}$$

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, you need $\mathbf{x}_F \geq 0$. 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 element, 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

$$b'_k \equiv b_k - \frac{F_{ki}b_j}{F_{ji}} \geq 0 \tag{11.12}$$

for each $k = 1, \cdots, m$ or equivalently,

$$b_k \geq \frac{F_{ki}b_j}{F_{ji}}. \tag{11.13}$$

Now if $F_{ki} \leq 0$ 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. This will work because for $F_{ki} > 0$,

$$\frac{b_p}{F_{pi}} \leq \frac{b_k}{F_{ki}} \Rightarrow b_k \geq \frac{F_{ki} b_p}{F_{pi}}$$

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.

Example 11.3.1 *Maximize* $2x_1 + 3x_2$ *subject to the constraints* $x_1 + x_2 \geq 1, 2x_1 + x_2 \leq 6, x_1 + 2x_2 \leq 6$, $x_1, x_2 \geq 0$.

The constraints are of the form

$$\begin{aligned} x_1 + x_2 - x_3 &= 1 \\ 2x_1 + x_2 + x_4 &= 6 \\ x_1 + 2x_2 + x_5 &= 6 \end{aligned}$$

where the x_3, x_4, x_5 are the slack variables. An augmented matrix for these equations is of the form

$$\begin{pmatrix} 1 & 1 & -1 & 0 & 0 & 1 \\ 2 & 1 & 0 & 1 & 0 & 6 \\ 1 & 2 & 0 & 0 & 1 & 6 \end{pmatrix}$$

Obviously the obvious solution is not feasible. It results in $x_3 < 0$. We need to exchange basic variables. Lets just try something.

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

Now this one is all right because the obvious solution is feasible. Letting $x_2 = x_3 = 0$, it follows that the obvious solution is feasible. Now we add in the objective function as described above.

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

Then do row operations to leave the simple columns the same. Then

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

Now there are negative numbers on the bottom row to the left of the 1. Lets pick the first. (It would be more sensible to pick the second.) The ratios to look at are $5/1, 1/1$ so pick for the pivot the 1 in the second column and first row. This will leave the right column above the lower right corner nonnegative. Thus the next tableau is

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

There is still a negative number there to the left of the 1 in the bottom row. The new ratios are $4/2, 5/1$ so the new pivot is the 2 in the third column. Thus the next tableau is

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

Still, there is a negative number in the bottom row to the left of the 1 so the process does not stop yet. The ratios are $3/(3/2)$ and $3/(1/2)$ and so the new pivot is that 3/2 in the first column. Thus the new tableau is

$$\begin{pmatrix} 0 & 1 & 0 & -\frac{1}{3} & \frac{2}{3} & 0 & 2 \\ \frac{3}{2} & 0 & 0 & 1 & -\frac{1}{2} & 0 & 3 \\ 0 & 0 & 2 & \frac{2}{3} & \frac{2}{3} & 0 & 6 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{4}{3} & 1 & 10 \end{pmatrix}$$

Now stop. The maximum value is 10. This is an easy enough problem to do geometrically and so you can easily verify that this is the right answer. It occurs when $x_4 = x_5 = 0, x_1 = 2, x_2 = 2, x_3 = 3$.

11.3.2 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 $\mathbf{b} \geq \mathbf{0}$.

$$\begin{pmatrix} I & F & \mathbf{0} & \mathbf{b} \\ \mathbf{0} & \mathbf{c} & 1 & z^0 \end{pmatrix} \tag{11.14}$$

Let $x_i^0 = b_i$ for $i = 1, \cdots, m$ and $x_i^0 = 0$ for $i > m$. Then $(\mathbf{x}^0, z^0)$ is a solution to the above system and since $\mathbf{b} \geq \mathbf{0}$, it follows $(\mathbf{x}^0, z^0)$ 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,

$$z = -c_i x_i + z^0$$

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 element and row operations.

Now return to Example 11.2.2. It will be modified to be a maximization problem.

Example 11.3.2 *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

$$\left(\begin{array}{ccccccc} 1 & 0 & -\frac{1}{3} & 0 & \frac{2}{3} & 0 & \frac{14}{3} \\ 0 & 1 & 0 & 0 & 1 & 0 & 8 \\ 0 & 0 & \frac{2}{3} & 1 & -\frac{1}{3} & 0 & \frac{2}{3} \\ 0 & 0 & 1 & 0 & -1 & 1 & -4 \end{array}\right).$$

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,

$$\left(\begin{array}{ccccccc} \frac{3}{2} & 0 & -\frac{1}{2} & 0 & 1 & 0 & 7 \\ -\frac{3}{2} & 1 & \frac{1}{2} & 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} & 1 & 0 & 0 & 3 \\ \frac{3}{2} & 0 & \frac{1}{2} & 0 & 0 & 1 & 3 \end{array}\right).$$

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 variables on the left.

Here is a bucolic example.

Example 11.3.3 *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

$$\begin{array}{rcl} x_1 + 2x_2 + x_3 + 3x_4 & \geq & 5, \\ 5x_1 + 3x_2 + 2x_3 + x_4 & \geq & 8, \\ x_1 + 2x_2 + 2x_3 + x_4 & \geq & 6, \\ 2x_1 + x_2 + x_3 + x_4 & \geq & 7, \\ x_1 + x_2 + x_3 + x_4 & \geq & 4. \end{array}$$

where each $x_i \geq 0$. Add in the slack variables,

$$\begin{array}{rcl} x_1 + 2x_2 + x_3 + 3x_4 - x_5 & = & 5 \\ 5x_1 + 3x_2 + 2x_3 + x_4 - x_6 & = & 8 \\ x_1 + 2x_2 + 2x_3 + x_4 - x_7 & = & 6 \\ 2x_1 + x_2 + x_3 + x_4 - x_8 & = & 7 \\ x_1 + x_2 + x_3 + x_4 - x_9 & = & 4 \end{array}$$

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 element 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

$$\begin{pmatrix} 0 & 1 & 0 & \frac{7}{3} & -1 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{2}{3} \\ 1 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & -\frac{2}{3} & 0 & \frac{8}{3} \\ 0 & 0 & 1 & -2 & 1 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{1}{3} & 0 & 0 & -\frac{1}{3} & -\frac{1}{3} & 1 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{11}{3} & -1 & 1 & \frac{2}{3} & -\frac{7}{3} & 0 & \frac{28}{3} \end{pmatrix}. \tag{11.15}$$

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

$$\begin{pmatrix} 0 & 1 & 0 & \frac{7}{3} & -1 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ 1 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & -\frac{2}{3} & 0 & 0 & \frac{8}{3} \\ 0 & 0 & 1 & -2 & 1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{1}{3} & 0 & 0 & -\frac{1}{3} & -\frac{1}{3} & 1 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{11}{3} & -1 & 1 & \frac{2}{3} & -\frac{7}{3} & 0 & 0 & \frac{28}{3} \\ -2 & -3 & -2 & -3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{11.16}$$

Now you do row operations to keep the simple columns of 11.15 simple in 11.16. 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 11.3.4 *Minimize $z = x_1 - 3x_2 + x_3$ subject to $x_1 + x_2 + x_3 \leq 10, x_1 + x_2 + x_3 \geq 2$, $x_1 + x_2 + 3x_3 \leq 8$ and $x_1 + 2x_2 + x_3 \leq 7$ with all variables nonnegative.*

There exists an answer because the region defined by the constraints 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

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 8 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 2 \\ -2 & -2 & 0 & 0 & 3 & 1 & 0 & 2 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 5 \end{pmatrix} \tag{11.17}$$

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 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 11.17. 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

$$\begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 1 & 0 & 0 & -\frac{1}{2} & 0 & \frac{13}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{7}{2} \\ \frac{1}{2} & 0 & \frac{5}{2} & 0 & 0 & 1 & -\frac{1}{2} & 0 & \frac{9}{2} \\ -\frac{1}{2} & 0 & -\frac{1}{2} & 0 & 1 & 0 & \frac{1}{2} & 0 & \frac{3}{2} \\ -\frac{5}{2} & 0 & -\frac{5}{2} & 0 & 0 & 0 & -\frac{3}{2} & 1 & -\frac{21}{2} \end{pmatrix}.$$

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 11.3.5 *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 $x_1 = 13/2, x_2 = 0, x_3 = 1/2$.

11.4 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 11.4.1 *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,

$$\left(\begin{array}{ccccccc} 2 & 1 & -1 & -1 & 0 & 0 & 3 \\ 1 & 1 & 1 & 0 & -1 & 0 & 2 \\ 1 & 1 & 1 & 0 & 0 & 1 & 7 \end{array}\right) \tag{11.18}$$

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

$$\left(\begin{array}{ccccccccc} 2 & 1 & -1 & -1 & 0 & 0 & 1 & 0 & 3 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 1 & 2 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 7 \end{array}\right) \tag{11.19}$$

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 11.18. Conversely, if you can find a feasible solution for 11.18, then letting x_7 and x_8 both equal zero, you have obtained a feasible solution to 11.19. 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 11.4.2 *There exists a feasible solution to the constraints represented by the augmented matrix of 11.18 and* $\mathbf{x} \geq \mathbf{0}$ *if and only if the minimum of* $x_7 + x_8$ *subject to the constraints of 11.19 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 11.19, 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

$$\left(\begin{array}{cccccccccc} 2 & 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 3 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 1 & 0 & 2 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 7 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 0 \end{array}\right)$$

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,

$$\left(\begin{array}{cccccccccc} 2 & 1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 3 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 1 & 0 & 2 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 7 \\ 3 & 2 & 0 & -1 & -1 & 0 & 0 & 0 & 1 & 5 \end{array}\right)$$

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

$$\begin{pmatrix} 1 & \frac{2}{3} & 0 & -\frac{1}{3} & -\frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{5}{3} \\ 0 & \frac{1}{3} & 1 & \frac{1}{3} & -\frac{2}{3} & 0 & -\frac{1}{3} & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & -1 & 0 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 0 \end{pmatrix} \tag{11.20}$$

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 11.18.

Now consider a maximization problem associated with the above constraints.

Example 11.4.3 *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 11.20 you can immediately assemble an initial simplex tableau. You begin with the first 6 columns and top 3 rows in 11.20. 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 $x_3 = 11/3$.

11.5 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 $z = \mathbf{cx}$ subject to $\mathbf{x} \geq \mathbf{0}$ and $A\mathbf{x} \geq \mathbf{b}$ and

$B.)$ Maximize $w = \mathbf{yb}$ 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 11.5.1 *Let* $\mathbf{x}$ *be a solution of the inequalities of* $A.)$ *and let* $\mathbf{y}$ *be a solution of the inequalities of* $B.)$*. Then*

$$\mathbf{cx} \geq \mathbf{yb}.$$

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 $\mathbf{c} \geq \mathbf{y}A$,

$$\mathbf{cx} \geq \mathbf{y}A\mathbf{x} \geq \mathbf{yb}.$$

It follows from this lemma that if $\mathbf{y}$ satisfies the inequalities 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 $B.)$. ■

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

$$\begin{pmatrix} A & -I & \mathbf{0} & \mathbf{b} \\ -\mathbf{c} & \mathbf{0} & 1 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} A^T & I & \mathbf{0} & \mathbf{c}^T \\ -\mathbf{b}^T & \mathbf{0} & 1 & 0 \end{pmatrix}. \tag{11.21}$$

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

$$\begin{pmatrix} B & F & \mathbf{0} & \mathbf{b} \\ -\mathbf{c}_B & -\mathbf{c}_F & 1 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} B_1 & F_1 & \mathbf{0} & \mathbf{c}^T \\ -\mathbf{b}_{B_1}^T & -\mathbf{b}_{F_1}^T & 1 & 0 \end{pmatrix} \tag{11.22}$$

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,

$$\begin{pmatrix} I & B^{-1}F & \mathbf{0} & B^{-1}\mathbf{b} \\ \mathbf{0} & \mathbf{c}_B B^{-1}F - \mathbf{c}_F & 1 & \mathbf{c}_B B^{-1}\mathbf{b} \end{pmatrix}, \begin{pmatrix} I & B_1^{-1}F_1 & \mathbf{0} & B_1^{-1}\mathbf{c}^T \\ \mathbf{0} & \mathbf{b}_{B_1}^T B_1^{-1}F - \mathbf{b}_{F_1}^T & 1 & \mathbf{b}_{B_1}^T B_1^{-1}\mathbf{c}^T \end{pmatrix} \quad (11.23)$$

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 11.5.2 *Suppose there exists a solution,* $\mathbf{x}$ *to* $A.)$ *where* $\mathbf{x}$ *is a basic feasible solution of the inequalities of* **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 11.23 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 possible you can take $\mathbf{y} = \mathbf{c}_B B^{-1}$? From Lemma 11.5.1 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 11.21 to get 11.22 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

$$\mathbf{y} = \mathbf{c}_B B^{-1}. \blacksquare$$

The proof of the following corollary is similar.

Corollary 11.5.3 *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 11.23, 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 11.5.4 *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

$$\begin{pmatrix} \frac{1}{5} & 1 & \frac{1}{5} & \frac{2}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 & \frac{2}{5} \\ \frac{7}{5} & 0 & \frac{7}{5} & -\frac{1}{5} & \frac{2}{5} & -\frac{3}{5} & 1 & 0 & 0 & 0 & \frac{9}{5} \\ \frac{3}{5} & 0 & \frac{8}{5} & \frac{1}{5} & \frac{3}{5} & -\frac{2}{5} & 0 & 1 & 0 & 0 & \frac{6}{5} \\ \frac{14}{5} & 0 & \frac{4}{5} & \frac{3}{5} & \frac{4}{5} & -\frac{1}{5} & 0 & 0 & 1 & 0 & \frac{13}{5} \\ -\frac{17}{5} & 0 & -\frac{22}{5} & -\frac{19}{5} & -\frac{12}{5} & \frac{8}{5} & 0 & 0 & 0 & 1 & \frac{16}{5} \end{pmatrix}.$$

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 11.5.3,

$$\mathbf{x} = \begin{pmatrix} -\frac{1}{7} & -\frac{2}{7} & \frac{5}{7} & \frac{1}{7} \\ 0 & 0 & 0 & 1 \\ -\frac{1}{7} & \frac{5}{7} & -\frac{2}{7} & -\frac{6}{7} \\ \frac{3}{7} & -\frac{1}{7} & -\frac{1}{7} & -\frac{3}{7} \end{pmatrix} \begin{pmatrix} 5 \\ 6 \\ 7 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{18}{7} \\ 0 \\ \frac{11}{7} \\ \frac{2}{7} \end{pmatrix}$$

which agrees with the original way of doing the problem.

Two good books which give more discussion of linear programming are Strang [17] and Nobel and Daniels [14]. 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.

11.6 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 if 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 $x_1 + 2x_2 \leq 5, x_1 \geq 0, x_2 \geq 0$. You know these two contradict but show they contradict using the simplex algorithm.

4. Find a solution to the following inequalities for $x, y \geq 0$ 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.

Chapter 12

Spectral Theory

12.1 Eigenvalues And Eigenvectors Of A Matrix

Spectral Theory refers to the study of eigenvalues and eigenvectors of a matrix. It is of fundamental importance in many areas. Row operations will no longer be the solution to all issues.

12.1.1 Definition Of Eigenvectors And Eigenvalues

In this section, $\mathbb{F} = \mathbb{C}$. In fact, it is assumed that the field is one for which eigenvalues exist and this will mean $\mathbb{C}$.

To illustrate the idea behind what will be discussed, consider the following example.

Example 12.1.1 *Here is a matrix.*

$$\begin{pmatrix} 0 & 5 & -10 \\ 0 & 22 & 16 \\ 0 & -9 & -2 \end{pmatrix}.$$

Multiply this matrix by the vector $\begin{pmatrix} 5 & -4 & 3 \end{pmatrix}^T$ *and see what happens. Then multiply it by* $\begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T$ *and see what happens. Does this matrix act this way for some other vector?*

First

$$\begin{pmatrix} 0 & 5 & -10 \\ 0 & 22 & 16 \\ 0 & -9 & -2 \end{pmatrix} \begin{pmatrix} -5 \\ -4 \\ 3 \end{pmatrix} = \begin{pmatrix} -50 \\ -40 \\ 30 \end{pmatrix} = 10 \begin{pmatrix} -5 \\ -4 \\ 3 \end{pmatrix}.$$

Next

$$\begin{pmatrix} 0 & 5 & -10 \\ 0 & 22 & 16 \\ 0 & -9 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = 0 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

When you multiply the first vector by the given matrix, it stretched the vector, multiplying it by 10. When you multiplied the matrix by the second vector it sent it to the zero vector. Now consider

$$\begin{pmatrix} 0 & 5 & -10 \\ 0 & 22 & 16 \\ 0 & -9 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -5 \\ 38 \\ -11 \end{pmatrix}.$$

In this case, multiplication by the matrix did not result in merely multiplying the vector by a number.

In the above example, the first two vectors were called eigenvectors and the numbers, 10 and 0 are called eigenvalues. Not every number is an eigenvalue and not every vector is an eigenvector. When you have a **nonzero** vector which, when multiplied by a matrix results in another vector which is parallel to the first or equal to **0,** this vector is called an eigenvector of the matrix. This is the meaning when the vectors are in $\mathbb{R}^n$. Things are less apparent geometrically when the vectors are in $\mathbb{C}^n$. The precise definition in all cases follows.

Definition 12.1.2 *Let M be an $n \times n$ matrix and let $\mathbf{x} \in \mathbb{C}^n$ be a* ***nonzero vector*** *for which*

$$M\mathbf{x} = \lambda \mathbf{x} \tag{12.1}$$

for some scalar λ. Then $\mathbf{x}$ is called an ***eigenvector*** *and λ is called an* ***eigenvalue*** *(****characteristic value****) of the matrix M.*

Note: Eigenvectors are never equal to zero!

The set of all eigenvalues of an $n \times n$ matrix M, is denoted by $\sigma(M)$ and is referred to as the ***spectrum*** *of M.*

The eigenvectors of a matrix M are those vectors, $\mathbf{x}$ for which multiplication by M results in a vector in the same direction or opposite direction to $\mathbf{x}$. Since the zero vector $\mathbf{0}$ has no direction this would make no sense for the zero vector. As noted above, $\mathbf{0}$ is never allowed to be an eigenvector. How can eigenvectors be identified? Suppose $\mathbf{x}$ satisfies 12.1. Then

$$(M - \lambda I)\mathbf{x} = \mathbf{0}$$

for some $\mathbf{x} \neq \mathbf{0}$. (Equivalently, you could write $(\lambda I - M)\mathbf{x} = \mathbf{0}$.) Sometimes we will use

$$(\lambda I - M)\mathbf{x} = \mathbf{0}$$

and sometimes $(M - \lambda I)\mathbf{x} = \mathbf{0}$. It makes absolutely no difference and you should use whichever you like better. Therefore, the matrix $M - \lambda I$ cannot have an inverse because if it did, the equation could be solved,

$$\mathbf{x} = \left((M - \lambda I)^{-1} (M - \lambda I) \right) \mathbf{x} = (M - \lambda I)^{-1} ((M - \lambda I)\mathbf{x}) = (M - \lambda I)^{-1} \mathbf{0} = \mathbf{0},$$

and this would require $\mathbf{x} = \mathbf{0}$, contrary to the requirement that $\mathbf{x} \neq \mathbf{0}$. By Theorem 6.2.1 on Page 117,

$$\det(M - \lambda I) = 0. \tag{12.2}$$

(Equivalently you could write $\det(\lambda I - M) = 0$.) The expression, $\det(\lambda I - M)$ or equivalently, $\det(M - \lambda I)$ is a polynomial called the **characteristic polynomial** and the above equation is called the characteristic equation. For M an $n \times n$ matrix, it follows from the theorem on expanding a matrix by its cofactor that $\det(M - \lambda I)$ is a polynomial of degree n. As such, the equation 12.2 has a solution, $\lambda \in \mathbb{C}$ by the fundamental theorem of algebra. Is it actually an eigenvalue? The answer is yes, and this follows from Observation 9.2.7 on Page 195 along with Theorem 6.2.1 on Page 117. Since $\det(M - \lambda I) = 0$ the matrix $\det(M - \lambda I)$ cannot be one to one and so there exists a nonzero vector $\mathbf{x}$ such that $(M - \lambda I)\mathbf{x} = \mathbf{0}$. This proves the following corollary.

Corollary 12.1.3 *Let M be an $n \times n$ matrix and $\det(M - \lambda I) = 0$. Then there exists a nonzero vector $\mathbf{x} \in \mathbb{C}^n$ such that $(M - \lambda I)\mathbf{x} = \mathbf{0}$.*

12.1.2 Finding Eigenvectors And Eigenvalues

As an example, consider the following.

Example 12.1.4 *Find the eigenvalues and eigenvectors for the matrix*

$$A = \begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix}.$$

You first need to identify the eigenvalues. Recall this requires the solution of the equation $\det(A - \lambda I) = 0$. In this case this equation is

$$\det\left(\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right) = 0$$

When you expand this determinant and simplify, you find the equation you need to solve is

$$(\lambda - 5)\left(\lambda^2 - 20\lambda + 100\right) = 0$$

and so the eigenvalues are $5, 10, 10$. We have listed 10 twice because it is a zero of multiplicity two due to

$$\lambda^2 - 20\lambda + 100 = (\lambda - 10)^2.$$

Having found the eigenvalues, it only remains to find the eigenvectors. First find the eigenvectors for $\lambda = 5$. As explained above, this requires you to solve the equation,

$$\left(\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} - 5\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

That is you need to find the solution to

$$\begin{pmatrix} 0 & -10 & -5 \\ 2 & 9 & 2 \\ -4 & -8 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

By now this is an old problem. You set up the augmented matrix and row reduce to get the solution. Thus the matrix you must row reduce is

$$\left(\begin{array}{ccc|c} 0 & -10 & -5 & 0 \\ 2 & 9 & 2 & 0 \\ -4 & -8 & 1 & 0 \end{array}\right). \tag{12.3}$$

The row reduced echelon form is

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

and so the solution is any vector of the form

$$\begin{pmatrix} \frac{5}{4}t \\ \frac{-1}{2}t \\ t \end{pmatrix} = t \begin{pmatrix} \frac{5}{4} \\ \frac{-1}{2} \\ 1 \end{pmatrix}$$

where $t \in \mathbb{F}$. You would obtain the same collection of vectors if you replaced t with $4t$. Thus a simpler description for the solutions to this system of equations whose augmented matrix is in 12.3 is

$$t \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix} \tag{12.4}$$

where $t \in \mathbb{F}$. Now you need to remember that you can't take $t = 0$ because this would result in the zero vector and

$$\boxed{\underline{\textbf{Eigenvectors}}\ \underline{\textbf{are}}\ \underline{\textbf{never}}\ \underline{\textbf{equal}}\ \underline{\textbf{to}}\ \underline{\textbf{zero}}\textbf{!}}$$

Other than this value, every other choice of z in 12.4 results in an eigenvector. It is a good idea to check your work! To do so, we will take the original matrix and multiply by this vector and see if we get 5 times this vector.

$$\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix} = \begin{pmatrix} 25 \\ -10 \\ 20 \end{pmatrix} = 5 \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix}$$

so it appears this is correct. Always check your work on these problems if you care about getting the answer right.

The parameter, t is sometimes called a **free variable.** The set of vectors in 12.4 is called the **eigenspace** and it is defined by $\ker(A - \lambda I)$. You should observe that in this case the eigenspace has dimension 1 because the eigenspace is the span of a single vector. In general, you obtain the solution from the row echelon form and the number of different free variables gives you the dimension of the eigenspace. Just remember that not every vector in the eigenspace is an eigenvector. The vector **0** is not an eigenvector although it is in the eigenspace because

$$\boxed{\textbf{\underline{Eigenvectors} \underline{are} \underline{never} \underline{equal} \underline{to} \underline{zero}!}}$$

Next consider the eigenvectors for $\lambda = 10$. These vectors are solutions to the equation,

$$\left(\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix} - 10\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

That is you must find the solutions to

$$\begin{pmatrix} -5 & -10 & -5 \\ 2 & 4 & 2 \\ -4 & -8 & -4 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

which reduces to consideration of the augmented matrix

$$\begin{pmatrix} -5 & -10 & -5 & | & 0 \\ 2 & 4 & 2 & | & 0 \\ -4 & -8 & -4 & | & 0 \end{pmatrix}$$

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

$$\begin{pmatrix} -2s-t \\ s \\ t \end{pmatrix} = s\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + t\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

You can't pick t and s both equal to zero because this would result in the zero vector and

$$\boxed{\textbf{\underline{Eigenvectors} \underline{are} \underline{never} \underline{equal} \underline{to} \underline{zero}!}}$$

However, every other choice of t and s does result in an eigenvector for the eigenvalue $\lambda = 10$. As in the case for $\lambda = 5$ you should check your work if you care about getting it right.

$$\begin{pmatrix} 5 & -10 & -5 \\ 2 & 14 & 2 \\ -4 & -8 & 6 \end{pmatrix}\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -10 \\ 0 \\ 10 \end{pmatrix} = 10\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

so it worked. The other vector will also work. Check it.

12.1.3 A Warning

The above example shows how to find eigenvectors and eigenvalues algebraically. You may have noticed it is a bit long. Sometimes students try to first row reduce the matrix before looking for eigenvalues. This is a $\boxed{\underline{\textbf{terrible}}\ \underline{\textbf{idea}}}$ because row operations destroy the eigenvalues. The eigenvalue problem is really not about row operations.

The general eigenvalue problem is the hardest problem in algebra and people still do research on ways to find eigenvalues and their eigenvectors. If you are doing anything which would yield a way to find eigenvalues and eigenvectors for general matrices without too much trouble, the thing you are doing will certainly be wrong. The problems you will see in this book are not too hard because they are cooked up to be easy. General methods to compute eigenvalues and eigenvectors numerically are presented later. These methods work even when the problem is not cooked up to be easy.

Notwithstanding the above discouraging observations, one can sometimes simplify the matrix first before searching for the eigenvalues in other ways.

Lemma 12.1.5 *Let $A = S^{-1}BS$ where A,B are $n \times n$ matrices. Then A,B have the same eigenvalues.*

Proof: Say $A\mathbf{x} = \lambda\mathbf{x}, \mathbf{x} \neq \mathbf{0}$. Then

$$S^{-1}BS\mathbf{x} = \lambda\mathbf{x} \text{ and so } BS\mathbf{x} = \lambda S\mathbf{x}.$$

Since S is one to one, $S\mathbf{x} \neq \mathbf{0}$. Thus if λ is an eigenvalue for A then it is also an eigenvalue for B. The other direction is similar. ■

Note that from the proof of the lemma, the eigenvectors for A,B are **different.**

One can now attempt to simplify a matrix before looking for the eigenvalues by using elementary matrices for S. This is illustrated in the following example.

Example 12.1.6 *Find the eigenvalues for the matrix*

$$\begin{pmatrix} 33 & 105 & 105 \\ 10 & 28 & 30 \\ -20 & -60 & -62 \end{pmatrix}$$

It has big numbers. Use the row operation of adding two times the second row to the bottom and multiply by the inverse of the elementary matrix which does this row operation as illustrated.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 33 & 105 & 105 \\ 10 & 28 & 30 \\ -20 & -60 & -62 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{pmatrix} = \begin{pmatrix} 33 & -105 & 105 \\ 10 & -32 & 30 \\ 0 & 0 & -2 \end{pmatrix}$$

This one has the same eigenvalues as the first matrix but is of course much easier to work with. Next, do the same sort of thing

$$\begin{pmatrix} 1 & -3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 33 & -105 & 105 \\ 10 & -32 & 30 \\ 0 & 0 & -2 \end{pmatrix} \begin{pmatrix} 1 & 3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 0 & 15 \\ 10 & -2 & 30 \\ 0 & 0 & -2 \end{pmatrix} \tag{12.5}$$

At this point, you can get the eigenvalues easily. This last matrix has the same eigenvalues as the first. Thus the eigenvalues are obtained by solving

$$(-2-\lambda)(-2-\lambda)(3-\lambda)=0,$$

and so the eigenvalues of the original matrix are $-2,-2,3$.

At this point, you go back to the **original matrix** A, form $A-\lambda I$, and then the problem from here on does reduce to row operations. In general, if you are so fortunate as to find the eigenvalues as in the above example, then finding the eigenvectors does reduce to row operations and this part of the problem is easy.

However, finding the eigenvalues along with the eigenvectors is anything but easy because for an $n\times n$ matrix A, it involves solving a polynomial equation of degree n. If you only find a good approximation to the eigenvalue, it won't work. It either is or is not an eigenvalue and if it is not, the only solution to the equation, $(A-\lambda I)\mathbf{x}=\mathbf{0}$ will be the zero solution as explained above and

Eigenvectors are never equal to zero!

Another thing worth noting is that when you multiply on the right by an elementary operation, you are merely doing the column operation defined by the elementary matrix. In 12.5 multiplication by the elementary matrix on the right merely involves taking three times the first column and adding to the second. Thus, without referring to the elementary matrices, the transition to the new matrix in 12.5 can be illustrated by

$$\begin{pmatrix} 33 & -105 & 105 \\ 10 & -32 & 30 \\ 0 & 0 & -2 \end{pmatrix} \to \begin{pmatrix} 3 & -9 & 15 \\ 10 & -32 & 30 \\ 0 & 0 & -2 \end{pmatrix} \to \begin{pmatrix} 3 & 0 & 15 \\ 10 & -2 & 30 \\ 0 & 0 & -2 \end{pmatrix}$$

Here is another example.

Example 12.1.7 *Let*

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

First find the eigenvalues. If you like, you could use the above technique to simplify the matrix, obtaining one which has the same eigenvalues, but since the numbers are not large, it is probably better to just expand the determinant without any tricks.

$$\det\left(\begin{pmatrix} 2 & 2 & -2 \\ 1 & 3 & -1 \\ -1 & 1 & 1 \end{pmatrix}-\lambda\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)=0$$

This reduces to $\lambda^3-6\lambda^2+8\lambda=0$ and the solutions are 0, 2, and 4.

0 **Can be an Eigenvalue!**

Now find the eigenvectors. For $\lambda = 0$ the augmented matrix for finding the solutions is

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

and the row reduced echelon form is

$$\left(\begin{array}{cccc} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Therefore, the eigenvectors are of the form $t \left(\begin{array}{ccc} 1 & 0 & 1 \end{array}\right)^T$ where $t \neq 0$.

Next find the eigenvectors for $\lambda = 2$. The augmented matrix for the system of equations needed to find these eigenvectors is

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

and the row reduced echelon form is

$$\left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

and so the eigenvectors are of the form $t \left(\begin{array}{ccc} 0 & 1 & 1 \end{array}\right)^T$ where $t \neq 0$.

Finally find the eigenvectors for $\lambda = 4$. The augmented matrix for the system of equations needed to find these eigenvectors is

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

and the row reduced echelon form is

$$\left(\begin{array}{cccc} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

Therefore, the eigenvectors are of the form $t \left(\begin{array}{ccc} 1 & 1 & 0 \end{array}\right)^T$ where $t \neq 0$.

12.1.4 Triangular Matrices

Although it is usually hard to solve the eigenvalue problem, there is a kind of matrix for which this is not the case. These are the upper or lower triangular matrices. I will illustrate by examples.

Example 12.1.8 *Let* $A = \begin{pmatrix} 1 & 2 & 4 \\ 0 & 4 & 7 \\ 0 & 0 & 6 \end{pmatrix}$. *Find its eigenvalues.*

You need to solve

$$\begin{aligned} 0 &= \det\left(\begin{pmatrix} 1 & 2 & 4 \\ 0 & 4 & 7 \\ 0 & 0 & 6 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right) \\ &= \det \begin{pmatrix} 1-\lambda & 2 & 4 \\ 0 & 4-\lambda & 7 \\ 0 & 0 & 6-\lambda \end{pmatrix} = (1-\lambda)(4-\lambda)(6-\lambda). \end{aligned}$$

Thus the eigenvalues are just the diagonal entries of the original matrix. You can see it would work this way with any such matrix. These matrices are called **upper triangular.** Stated precisely, a matrix A is upper triangular if $A_{ij} = 0$ for all $i > j$. Similarly, it is easy to find the eigenvalues for a lower triangular matrix, on which has all zeros above the main diagonal.

12.1.5 Defective And Nondefective Matrices

Definition 12.1.9 *By the fundamental theorem of algebra, it is possible to write the characteristic equation in the form*

$$(\lambda - \lambda_1)^{r_1} (\lambda - \lambda_2)^{r_2} \cdots (\lambda - \lambda_m)^{r_m} = 0$$

where r_i is some integer no smaller than 1. Thus the eigenvalues are $\lambda_1, \lambda_2, \cdots, \lambda_m$. The ***algebraic multiplicity*** *of λ_j is defined to be r_j.*

Example 12.1.10 *Consider the matrix*

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \tag{12.6}$$

What is the algebraic multiplicity of the eigenvalue $\lambda = 1$?

In this case the characteristic equation is

$$\det(A - \lambda I) = (1 - \lambda)^3 = 0$$

or equivalently,

$$\det(\lambda I - A) = (\lambda - 1)^3 = 0.$$

Therefore, λ is of algebraic multiplicity 3.

Definition 12.1.11 *The* ***geometric multiplicity*** *of an eigenvalue is the dimension of the eigenspace,* $\ker(A - \lambda I)$.

Example 12.1.12 *Find the geometric multiplicity of $\lambda = 1$ for the matrix in 12.6.*

We need to solve

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

The augmented matrix which must be row reduced to get this solution is therefore,

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

This requires $z = y = 0$ and x is arbitrary. Thus the eigenspace is $t \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T, t \in \mathbb{F}$. It follows the geometric multiplicity of $\lambda = 1$ is 1.

Definition 12.1.13 *An $n \times n$ matrix is called* ***defective*** *if the geometric multiplicity is not equal to the algebraic multiplicity for some eigenvalue. Sometimes such an eigenvalue for which the geometric multiplicity is not equal to the algebraic multiplicity is called a defective eigenvalue. If the geometric multiplicity for an eigenvalue equals the algebraic multiplicity, the eigenvalue is sometimes referred to as nondefective.*

Here is another more interesting example of a defective matrix.

Example 12.1.14 *Let*

$$A = \begin{pmatrix} 2 & -2 & -1 \\ -2 & -1 & -2 \\ 14 & 25 & 14 \end{pmatrix}.$$

Find the eigenvectors and eigenvalues.

In this case the eigenvalues are $3, 6, 6$ where we have listed 6 twice because it is a zero of algebraic multiplicity two, the characteristic equation being

$$(\lambda - 3)(\lambda - 6)^2 = 0.$$

It remains to find the eigenvectors for these eigenvalues. First consider the eigenvectors for $\lambda = 3$. You must solve

$$\left(\begin{pmatrix} 2 & -2 & -1 \\ -2 & -1 & -2 \\ 14 & 25 & 14 \end{pmatrix} - 3 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The augmented matrix is

$$\left(\begin{array}{ccc|c} -1 & -2 & -1 & 0 \\ -2 & -4 & -2 & 0 \\ 14 & 25 & 11 & 0 \end{array}\right)$$

and the row reduced echelon form is

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

so the eigenvectors are nonzero vectors of the form

$$\begin{pmatrix} t & -t & t \end{pmatrix}^T = t\begin{pmatrix} 1 & -1 & 1 \end{pmatrix}^T$$

Next consider the eigenvectors for $\lambda = 6$. This requires you to solve

$$\left(\begin{pmatrix} 2 & -2 & -1 \\ -2 & -1 & -2 \\ 14 & 25 & 14 \end{pmatrix} - 6 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

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 $t\begin{pmatrix} -\frac{1}{8} & -\frac{1}{4} & 1 \end{pmatrix}^T$ or simply as

$$t\begin{pmatrix} -1 & -2 & 8 \end{pmatrix}^T$$

where $t \in \mathbb{F}$.

Note that in this example the eigenspace for the eigenvalue $\lambda = 6$ is of dimension 1 because there is only one parameter. However, this eigenvalue is of multiplicity two as a root to the characteristic equation. Thus this eigenvalue is a defective eigenvalue. However, the eigenvalue 3 is nondefective. The matrix is defective because it has a defective eigenvalue.

The word, defective, seems to suggest there is something wrong with the matrix. This is in fact the case. Defective matrices are a lot of trouble in applications and we may wish they never occurred. However, they do occur as the above example shows. When you study linear systems of differential equations, you will have to deal with the case of defective matrices and you will see how awful they are. The reason these matrices are so horrible to work with is that it is impossible to obtain a basis of eigenvectors. When you study differential equations, solutions to first order systems are expressed in terms of eigenvectors of a certain matrix times $e^{\lambda t}$ where λ is an eigenvalue. In order to obtain a general solution of this sort, you must have a basis of eigenvectors. For a defective matrix, such a basis does not exist and so you have to go to something called generalized eigenvectors. Unfortunately, it is **never** explained in beginning differential equations courses why there are enough generalized eigenvectors and eigenvectors to represent the general solution. In fact, this reduces to a difficult question in linear algebra equivalent to the existence of something called the Jordan Canonical form which is much more difficult than everything discussed in the entire differential equations course. If you become interested in this, see Appendix A.

Ultimately, the algebraic issues which will occur in differential equations are a red herring anyway. The real issues relative to existence of solutions to systems of ordinary differential equations are analytical, having much more to do with calculus than with linear algebra although this will likely not be made clear when you take a beginning differential equations class.

In terms of algebra, this lack of a basis of eigenvectors says that it is impossible to obtain a diagonal matrix which is similar to the given matrix.

Although there may be repeated roots to the characteristic equation, 12.2 and it is not known whether the matrix is defective in this case, there is an important theorem which holds when considering eigenvectors which correspond to distinct eigenvalues.

Theorem 12.1.15 *Suppose $M\mathbf{v}_i = \lambda_i \mathbf{v}_i, i = 1, \cdots, r$, $\mathbf{v}_i \neq 0$, and that if $i \neq j$, then $\lambda_i \neq \lambda_j$. Then the set of eigenvectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ is linearly independent.*

Proof. Suppose the claim of the lemma is not true. Then there exists a subset of this set of vectors

$$\{\mathbf{w}_1, \cdots, \mathbf{w}_r\} \subseteq \{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$$

such that

$$\sum_{j=1}^{r} c_j \mathbf{w}_j = \mathbf{0} \tag{12.7}$$

where each $c_j \neq 0$. Say $M\mathbf{w}_j = \mu_j \mathbf{w}_j$ where

$$\{\mu_1, \cdots, \mu_r\} \subseteq \{\lambda_1, \cdots, \lambda_k\},$$

the μ_j being distinct eigenvalues of M. Out of all such subsets, let this one be such that r is as small as possible. Then necessarily, $r > 1$ because otherwise, $c_1 \mathbf{w}_1 = \mathbf{0}$ which would imply $\mathbf{w}_1 = \mathbf{0}$, which is not allowed for eigenvectors.

Now apply M to both sides of 12.7.

$$\sum_{j=1}^{r} c_j \mu_j \mathbf{w}_j = \mathbf{0}. \tag{12.8}$$

Next pick $\mu_k \neq 0$ and multiply both sides of 12.7 by μ_k. Such a μ_k exists because $r > 1$. Thus

$$\sum_{j=1}^{r} c_j \mu_k \mathbf{w}_j = \mathbf{0} \tag{12.9}$$

Subtract the sum in 12.9 from the sum in 12.8 to obtain

$$\sum_{j=1}^{r} c_j \left(\mu_k - \mu_j\right) \mathbf{w}_j = \mathbf{0}$$

Now one of the constants $c_j\left(\mu_k - \mu_j\right)$ equals 0, when $j = k$. Therefore, r was not as small as possible after all. ■

Here is another proof in case you did not follow the above.

Theorem 12.1.16 *Suppose $M\mathbf{v}_i = \lambda_i \mathbf{v}_i, i = 1, \cdots, r$, $\mathbf{v}_i \neq 0$, and that if $i \neq j$, then $\lambda_i \neq \lambda_j$. Then the set of eigenvectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ is linearly independent.*

Proof: Suppose the conclusion is not true. Then in the matrix

$$\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_r \end{pmatrix}$$

not every column is a pivot column. Let the pivot columns be $\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}$, $k < r$. Then there exists $\mathbf{v} \in \{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$, $M\mathbf{v} = \lambda_{\mathbf{v}}\mathbf{v}$, $\mathbf{v} \notin \{\mathbf{w}_1, \cdots, \mathbf{w}_k\}$, such that

$$\mathbf{v} = \sum_{i=1}^{k} c_i \mathbf{w}_i. \tag{12.10}$$

Then doing M to both sides yields

$$\lambda_{\mathbf{v}}\mathbf{v} = \sum_{i=1}^{k} c_i \lambda_{\mathbf{w}_i} \mathbf{w}_i \tag{12.11}$$

But also you could multiply both sides of 12.10 by $\lambda_{\mathbf{v}}$ to get

$$\lambda_{\mathbf{v}}\mathbf{v} = \sum_{i=1}^{k} c_i \lambda_{\mathbf{v}} \mathbf{w}_i.$$

And now subtracting this from 12.11 yields

$$\mathbf{0} = \sum_{i=1}^{k} c_i \left(\lambda_{\mathbf{v}} - \lambda_{\mathbf{w}_i}\right) \mathbf{w}_i$$

and by independence of the $\{\mathbf{w}_1, \cdots, \mathbf{w}_k\}$, this requires $c_i\left(\lambda_{\mathbf{v}} - \lambda_{\mathbf{w}_i}\right) = 0$ for each i. Since the eigenvalues are distinct, $\lambda_{\mathbf{v}} - \lambda_{\mathbf{w}_i} \neq 0$ and so each $c_i = 0$. But from 12.10, this requires $\mathbf{v} = \mathbf{0}$ which is impossible because $\mathbf{v}$ is an eigenvector and

$$\boxed{\textbf{\underline{Eigenvectors} \underline{are} \underline{never} \underline{equal} \underline{to} \underline{zero}!}} \ ■$$

Definition 12.1.17 *An $n \times n$ matrix A is called nondefective if and only if there exists a basis of eigenvectors for $\mathbb{F}^n$.*

In fact the geometric multiplicity is never larger than the algebraic multiplicity. Let A denote an $n \times n$ matrix in what follows and we assume there is an eigenvalue λ and $\mathbb{F}$ will denote the field of scalars.

Theorem 12.1.18 *Let λ be an eigenvalue for an $n \times n$ matrix. Then its geometric multiplicity is never larger than its algebraic multiplicity.*

Proof: Let $\{\mathbf{v}_1,\cdots,\mathbf{v}_r\}$ be a basis for the eigenspace corresponding to some λ. Then by Theorem 8.5.21, there is a longer list $\{\mathbf{v}_1,\cdots,\mathbf{v}_r,\mathbf{u}_1,\cdots,\mathbf{u}_{n-r}\}$ which is a basis for $\mathbb{F}^n$. Thus the matrix $S\equiv\begin{pmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_r & \mathbf{u}_1 & \cdots & \mathbf{u}_{n-r}\end{pmatrix}$ is invertible. Say its inverse is

$$S^{-1}=\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_r & \mathbf{b}_1 & \cdots & \mathbf{b}_{n-r}\end{pmatrix}^T$$

where $\mathbf{a}_i^T\mathbf{v}_j=\delta_{ij}$. Then $S^{-1}AS$ must be

$$\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_r & \mathbf{b}_1 & \cdots & \mathbf{b}_{n-r}\end{pmatrix}^T\begin{pmatrix}\lambda\mathbf{v}_1 & \cdots & \lambda\mathbf{v}_r & A\mathbf{u}_1 & \cdots & A\mathbf{u}_{n-r}\end{pmatrix}$$

and so $S^{-1}AS$ is of the form

$$\begin{pmatrix}D & M\\ 0 & N\end{pmatrix} \tag{*}$$

where D is an $r\times r$ diagonal matrix having λ down the main diagonal, M an $r\times(n-r)$, and N a $(n-r)\times(n-r)$. Now

$$\begin{aligned}\det\left(S^{-1}AS-\mu I\right) &= \det\left(S^{-1}(A-\mu I)S\right)\\ &= \det\left(S^{-1}\right)\det(S)\det(A-\mu I)=\det(A-\mu I)\end{aligned}$$

and so this matrix in $*$ has the same eigenvalues with the same multiplicities as the matrix A. So consider the characteristic polynomial with variable μ of the matrix in $*$. Expanding repeatedly along the first column, one obtains the characteristic polynomial is of the form

$$q(\mu)=(\mu-\lambda)^r\det(\mu I-N)=0$$

and so the multiplicity of λ is at least r. ■

This yields easily the following corollary which ties this theorem to Theorem 12.1.16.

Corollary 12.1.19 *Let A be an $n\times n$ matrix and suppose the characteristic polynomial factors completely and that the eigenvalues are $\lambda_1,\cdots,\lambda_m$. If no eigenvalue is defective, then A is nondefective. If some eigenvalue is defective, then A is defective.*

Proof: Let $\left\{\mathbf{v}_j^i\right\}_{j=1}^{r_i}\equiv\beta_i$ be a basis for the eigenspace of λ_i. First I claim that

$$\{\beta_1,\cdots,\beta_m\}$$

is linearly independent. To see this, suppose $\mathbf{w}_i\in\operatorname{span}(\beta_i)$ and $\sum_{i=1}^m\mathbf{w}_i=\mathbf{0}$. Then by Theorem 12.1.16, each $\mathbf{w}_i=\mathbf{0}$ since otherwise, you would have a nontrivial linear combination of eigenvectors associated with distinct eigenvalues which is $\mathbf{0}$. Now suppose

$$\sum_i\sum_j c_j^i\mathbf{v}_j^i=\mathbf{0}$$

Letting $\mathbf{w}_i=\sum_j c_j^i\mathbf{v}_j^i$, it follows from what was just observed that each $\mathbf{w}_i=\mathbf{0}$. Now the independence of the vectors in β_i implies that for each $i, c_j^i=0$ for each j. Thus these vectors $\{\beta_1, \cdots, \beta_m\}$ are linearly independent as claimed. Letting $|\beta_i|$ denote the dimension of span (β_i), which equals the geometric multiplicity of λ_i and letting m_i denote the algebraic multiplicity of λ_i it follows that

$$\sum_i |\beta_i| \leq n = \sum_i m_i$$

If each $|\beta_i| = m_i$, then A is nondefective because $\{\beta_1, \cdots, \beta_m\}$ will then be a basis of eigenvectors. This is the case of no defective eigenvalues.

In case $|\beta_i| < m_i$ for some i, then A must be defective because if not, you would have a basis of eigenvectors and you could let γ_j be those which pertain to λ_j. Then γ_j would be an independent set of vectors and therefore, the number of vectors in γ_j denoted as $\left|\gamma_j\right|$ is no more than $\left|\beta_j\right|$. But then,

$$\sum_j \left|\gamma_j\right| \leq \sum_j \left|\beta_j\right| < \sum_j m_j = n$$

and so, you would have fewer than n vectors in this basis of eigenvectors which cannot occur. Hence A is defective. Thus if an eigenvalue is defective, the matrix is defective and if no eigenvalue is defective, then the matrix is nondefective. ■

12.1.6 Diagonalization

First of all, here is what it means for two matrices to be similar.

Definition 12.1.20 *Let A, B be two $n \times n$ matrices. Then they are* ***similar*** *if and only if there exists an invertible matrix S such that*

$$A = S^{-1} B S$$

Proposition 12.1.21 *Define for $n \times n$ matrices $A \sim B$ if A is similar to B. Then*

$$A \sim A,$$

$$\text{If } A \sim B \text{ then } B \sim A$$

$$\text{If } A \sim B \text{ and } B \sim C \text{ then } A \sim C$$

Proof: It is clear that $A \sim A$ because you could just take $S = I$. If $A \sim B$, then for some S invertible,

$$A = S^{-1} B S \text{ and so } S A S^{-1} = B$$

But then

$$\left(S^{-1}\right)^{-1} A S^{-1} = B$$

which shows that $B \sim A$.

Now suppose $A \sim B$ and $B \sim C$. Then there exist invertible matrices S, T such that

$$A = S^{-1} B S, \; B = T^{-1} C T.$$

Therefore,

$$A = S^{-1} T^{-1} C T S = (TS)^{-1} C (TS)$$

showing that A is similar to C. ■

For your information, when $\sim$ satisfies the above conditions, it is called a similarity relation. Similarity relations are very significant in mathematics.

When a matrix is similar to a diagonal matrix, the matrix is said to be diagonalizable. I think this is one of the worst monstrosities for a word that I have ever seen. Nevertheless, it is commonly used in linear algebra. It turns out to be the same as nondefective which will follow easily from later material. The following is the precise definition.

Definition 12.1.22 *Let A be an $n \times n$ matrix. Then A is* ***diagonalizable*** *if there exists an invertible matrix S such that*

$$S^{-1}AS = D$$

where D is a diagonal matrix. This means D has a zero as every entry except for the main diagonal. More precisely, $D_{ij} = 0$ unless $i = j$. Such matrices look like the following.

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

where $*$ might not be zero.

The most important theorem about diagonalizability[1] is the following major result. First here is a simple observation.

Observation 12.1.23 *Let $S = \begin{pmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_n \end{pmatrix}$ where S is $n \times n$. Then here is the result of multiplying on the right by a diagonal matrix.*

$$\begin{pmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} = \begin{pmatrix} \lambda_1 \mathbf{s}_1 & \cdots & \lambda_n \mathbf{s}_n \end{pmatrix}$$

This follows from the way we multiply matrices. The i^{th} entry of the j^{th} column of the product on the left is of the form $s_i \lambda_j$. Thus the j^{th} column of the matrix on the left is just $\lambda_j \mathbf{s}_j$.

Theorem 12.1.24 *An $n \times n$ matrix is diagonalizable if and only if $\mathbb{F}^n$ has a basis of eigenvectors of A. Furthermore, you can take the matrix S described above, to be given as*

$$S = \begin{pmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \cdots & \mathbf{s}_n \end{pmatrix}$$

where here the $\mathbf{s}_k$ are the eigenvectors in the basis for $\mathbb{F}^n$. If A is diagonalizable, the eigenvalues of A are the diagonal entries of the diagonal matrix.

Proof: To say that A is diagonalizable, is to say that

$$S^{-1}AS = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

[1] This word has 9 syllables. It is a little like the name of a volcano in Iceland.

the λ_i being elements of $\mathbb{F}$. This is to say that for $S = \begin{pmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_n \end{pmatrix}$, $\mathbf{s}_k$ being the k^{th} column,

$$A \begin{pmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_n \end{pmatrix} = \begin{pmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_n \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

which is equivalent, from the way we multiply matrices, that

$$\begin{pmatrix} A\mathbf{s}_1 & \cdots & A\mathbf{s}_n \end{pmatrix} = \begin{pmatrix} \lambda_1 \mathbf{s}_1 & \cdots & \lambda_n \mathbf{s}_n \end{pmatrix}$$

which is equivalent to saying that the columns of S are eigenvectors and the diagonal matrix has the eigenvectors down the main diagonal. Since S^{-1} is invertible, these eigenvectors are a basis. Similarly, if there is a basis of eigenvectors, one can take them as the columns of S and reverse the above steps, finally concluding that A is diagonalizable. ■

Example 12.1.25 *Let* $A = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 4 & -1 \\ -2 & -4 & 4 \end{pmatrix}$. *Find a matrix, S such that*

$$S^{-1}AS = D,$$

where D is a diagonal matrix.

Solving $\det(\lambda I - A) = 0$ yields the eigenvalues are 2 and 6 with 2 an eigenvalue of multiplicity two. Solving $(2I - A)\mathbf{x} = \mathbf{0}$ to find the eigenvectors, you find that the eigenvectors are

$$a \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + b \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

where a, b are scalars. An eigenvector for $\lambda = 6$ is $\begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}$. Let the matrix S be

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

That is, the columns are the eigenvectors. Then

$$S^{-1} = \begin{pmatrix} -\frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \end{pmatrix}.$$

Then $S^{-1}AS =$

$$\begin{pmatrix} -\frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 1 & 4 & -1 \\ -2 & -4 & 4 \end{pmatrix} \begin{pmatrix} -2 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -2 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 6 \end{pmatrix}.$$

We know the result from the above theorem, but it is nice to see it work in a specific example just the same. You may wonder if there is a need to find S^{-1}. The following is an example of a situation where this is needed. It is one of the major applications of diagonalizability.

Example 12.1.26 *Here is a matrix.* $A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix}$ *Find* A^{50}.

Sometimes this sort of problem can be made easy by using diagonalization. In this case there are eigenvectors,

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

the first two corresponding to $\lambda = 1$ and the last corresponding to $\lambda = 2$. Then let the eigenvectors be the columns of the matrix, S. Thus

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

Then also

$$S^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix}$$

and $S^{-1}AS =$

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

Now it follows

$$A = SDS^{-1} = \begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix}.$$

Note that

$$\left(SDS^{-1}\right)^2 = SDS^{-1}SDS^{-1} = SD^2S^{-1}$$

and

$$\left(SDS^{-1}\right)^3 = SDS^{-1}SDS^{-1}SDS^{-1} = SD^3S^{-1},$$

etc. In general, you can see that

$$\left(SDS^{-1}\right)^n = SD^nS^{-1}$$

In other words, $A^n = SD^nS^{-1}$. Therefore,

$$A^{50} = SD^{50}S^{-1} = \begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}^{50} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix}.$$

It is easy to raise a diagonal matrix to a power.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}^{50} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2^{50} \end{pmatrix}.$$

It follows $A^{50} =$

$$\begin{pmatrix} 0 & -1 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2^{50} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix} = \begin{pmatrix} 2^{50} & -1+2^{50} & 0 \\ 0 & 1 & 0 \\ 1-2^{50} & 1-2^{50} & 1 \end{pmatrix}$$

That isn't too hard. However, this would have been horrendous if you had tried to multiply A^{50} by hand.

This technique of diagonalization is also important in solving the differential equations resulting from vibrations. Sometimes you have systems of differential equation and when you diagonalize an appropriate matrix, you "decouple" the equations. This is very nice. It makes hard problems trivial.

The above example is entirely typical. If $A = SDS^{-1}$ then $A^m = SD^mS^{-1}$ and it is easy to compute D^m. More generally, you can define functions of the matrix using power series in this way.

12.1.7 The Matrix Exponential

When A is diagonalizable, one can easily define what is meant by e^A. Here is how. You know

$$S^{-1}AS = D$$

where D is a diagonal matrix. You also know that if D is of the form

$$\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \tag{12.12}$$

then

$$D^m = \begin{pmatrix} \lambda_1^m & & 0 \\ & \ddots & \\ 0 & & \lambda_n^m \end{pmatrix}$$

and that $A^m = SD^mS^{-1}$as shown above. Recall why this was. $A = SDS^{-1}$ and so

$$A^m = \overbrace{SDS^{-1}SDS^{-1}SDS^{-1} \cdots SDS^{-1}}^{\text{n times}} = SD^mS^{-1}$$

Now formally write the following power series for e^A

$$e^A \equiv \sum_{k=0}^{\infty} \frac{A^k}{k!} = \sum_{k=0}^{\infty} \frac{SD^kS^{-1}}{k!} = S\sum_{k=0}^{\infty} \frac{D^k}{k!} S^{-1}$$

If D is given above in 12.12, the above sum is of the form

$$S\sum_{k=0}^{\infty} \begin{pmatrix} \frac{1}{k!}\lambda_1^k & & 0 \\ & \ddots & \\ 0 & & \frac{1}{k!}\lambda_n^k \end{pmatrix} S^{-1} = S \begin{pmatrix} \sum_{k=0}^{\infty} \frac{1}{k!}\lambda_1^k & & 0 \\ & \ddots & \\ 0 & & \sum_{k=0}^{\infty} \frac{1}{k!}\lambda_n^k \end{pmatrix} S^{-1}$$

$$= S \begin{pmatrix} e^{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_n} \end{pmatrix} S^{-1}$$

and this last thing is the definition of what is meant by e^A.

Example 12.1.27 *Let*

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

Find e^A.

The eigenvalues happen to be $1,2,3$ and eigenvectors associated with these eigenvalues are

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

Then let

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

and so

$$S^{-1} = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

and

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

Then the matrix exponential is

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

$$\begin{pmatrix} e^2 & e^2-e^3 & e^2-e^3 \\ e^2-e & e^2 & e^2-e \\ -e^2+e & -e^2+e^3 & -e^2+e+e^3 \end{pmatrix}$$

Isn't that nice? You could also talk about $\sin(A)$ or $\cos(A)$ etc. You would just have to use a different power series.

This matrix exponential is actually a useful idea when solving autonomous systems of first order linear differential equations. These are equations which are of the form $\mathbf{x}' = A\mathbf{x}$ where $\mathbf{x}$ is a vector in $\mathbb{R}^n$ or $\mathbb{C}^n$ and A is an $n \times n$ matrix. Then it turns out that the solution to the above system of equations is $\mathbf{x}(t) = e^{At}\mathbf{c}$ where $\mathbf{c}$ is a constant vector.

12.1.8 Complex Eigenvalues

Sometimes you have to consider eigenvalues which are complex numbers. This occurs in differential equations for example. You do these problems exactly the same way as you do the ones in which the eigenvalues are real. Here is an example.

Example 12.1.28 *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

$$\det\left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right) = 0.$$

This reduces to $(\lambda - 1)\left(\lambda^2 - 4\lambda + 5\right) = 0$. The solutions are $\lambda = 1, \lambda = 2+i, \lambda = 2-i$.

There is nothing new about finding the eigenvectors for $\lambda = 1$ so consider the eigenvalue $\lambda = 2+i$. You need to solve

$$\left((2+i)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

In other words, you must consider the augmented matrix

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

for the solution. Divide the top row by $(1+i)$ and then take $-i$ times the second row and add to the bottom. This yields

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

Now multiply the second row by $-i$ to obtain

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

Therefore, the eigenvectors are of the form $t\begin{pmatrix} 0 & i & 1 \end{pmatrix}^T$. You should find the eigenvectors for $\lambda = 2-i$. These are $t\begin{pmatrix} 0 & -i & 1 \end{pmatrix}^T$. As usual, if you want to get it right you had better check it.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 0 \\ -i \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ -1-2i \\ 2-i \end{pmatrix} = (2-i)\begin{pmatrix} 0 \\ -i \\ 1 \end{pmatrix}$$

so it worked.

12.2 Some Applications Of Eigenvalues And Eigenvectors

12.2.1 Principal Directions

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 factorization**, 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 elasticity, the stress is often taken to depend on U which is known in this context as the right **Cauchy Green strain tensor**. In this context, the eigenvalues will always be positive. The symmetry of U allows the proof of a theorem which says that if λ_M is the largest eigenvalue, then in every other direction other than the one corresponding to the eigenvector for λ_M the material is stretched less than λ_M and if λ_m is the smallest eigenvalue, then in every other direction other than the one corresponding to an eigenvector of λ_m the material is stretched more than λ_m. 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 we are emphasizing the application to mechanics in which the eigenvectors of the symmetric matrix U determine the **principal directions**, those directions in which the material is stretched the most or the least.

Example 12.2.1 *Find the principal 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 stretched by a factor of $1/2$, 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 principal direction for the eigenvalue, 3 in which the material is stretched to the maximum extent is $\begin{pmatrix} 3 & 1 & 1 \end{pmatrix}^T$. A direction vector (or unit vector) in this direction is

$$\begin{pmatrix} \frac{3}{11}\sqrt{11} & \frac{1}{11}\sqrt{11} & \frac{1}{11}\sqrt{11} \end{pmatrix}^T$$

You should show that the direction in which the material is compressed the most is in the direction

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

Note this is meaningful information which you would have a hard time finding without the theory of eigenvectors and eigenvalues.

12.2.2 Migration Matrices

There are applications which are of great importance which feature only one eigenvalue.

Definition 12.2.2 *Let n locations be denoted by the numbers $1,2,\cdots,n$. Also suppose it is the case that each year a_{ij} denotes the proportion of residents in location j which move to location i. Also suppose no one escapes or emigrates from without these n locations. This last assumption requires $\sum_i a_{ij} = 1$. Such matrices in which the columns are nonnegative numbers which sum to one are called **Markov matrices**. In this context describing migration, they are also called **migration matrices**.*

Example 12.2.3 *Here is an example of one of these matrices.*

$$\begin{pmatrix} .4 & .2 \\ .6 & .8 \end{pmatrix}$$

Thus if it is considered as a migration matrix, $.4$ is the proportion of residents in location 1 which stay in location one in a given time period while $.6$ is the proportion of residents in location 1 which move to location 2 and $.2$ is the proportion of residents in location 2 which move to location 1. Considered as a Markov matrix, these numbers are usually identified with probabilities.

If $\mathbf{v} = (x_1,\cdots,x_n)^T$ where x_i is the population of location i at a given instant, you obtain the population of location i one year later by computing $\sum_j a_{ij}x_j = (A\mathbf{v})_i$. Therefore, the population of location i after k years is $\left(A^k\mathbf{v}\right)_i$. An obvious application of this would be to a situation in which you rent trailers which can go to various parts of a city and you observe through experiments the proportion of trailers which go from point i to point j in a single day. Then you might want to find how many trailers would be in all the locations after 8 days.

Proposition 12.2.4 *Let $A = (a_{ij})$ be a migration matrix. Then 1 is always an eigenvalue for A.*

Proof: Remember that $\det\left(B^T\right) = \det(B)$. Therefore,

$$\det(A-\lambda I) = \det\left((A-\lambda I)^T\right) = \det\left(A^T - \lambda I\right)$$

because $I^T = I$. Thus the characteristic equation for A is the same as the characteristic equation for A^T and so A and A^T have the same eigenvalues. We will show that 1 is an eigenvalue for A^T and then it will follow that 1 is an eigenvalue for A.

Remember that for a migration matrix, $\sum_i a_{ij} = 1$. Therefore, if $A^T = (b_{ij})$ so $b_{ij} = a_{ji}$, it follows that

$$\sum_j b_{ij} = \sum_j a_{ji} = 1.$$

Therefore, from matrix multiplication,

$$A^T \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \sum_j b_{ij} \\ \vdots \\ \sum_j b_{ij} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

which shows that $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ is an eigenvector for A^T corresponding to the eigenvalue, $\lambda = 1$. As explained above, this shows that $\lambda = 1$ is an eigenvalue for A because A and A^T have the same eigenvalues. ■

Example 12.2.5 *Consider the migration matrix* $\begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix}$ *for locations* $1, 2,$ *and* 3. *Suppose initially there are 100 residents in location 1, 200 in location 2 and 400 in location 4. Find the population in the three locations after 10 units of time.*

From the above, it suffices to consider

$$\begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix}^{10} \begin{pmatrix} 100 \\ 200 \\ 400 \end{pmatrix} = \begin{pmatrix} 115.08582922 \\ 120.13067244 \\ 464.78349834 \end{pmatrix}$$

Of course you would need to round these numbers off.

A related problem asks for how many there will be in the various locations after a long time. It turns out that if some power of the migration matrix has all positive entries, then there is a limiting vector $\mathbf{x} = \lim_{k\to\infty} A^k \mathbf{x}_0$ where $\mathbf{x}_0$ is the initial vector describing the number of inhabitants in the various locations initially. This vector will be an eigenvector for the eigenvalue 1 because

$$\mathbf{x} = \lim_{k\to\infty} A^k \mathbf{x}_0 = \lim_{k\to\infty} A^{k+1}\mathbf{x}_0 = A \lim_{k\to\infty} A^k \mathbf{x} = A\mathbf{x},$$

and the sum of its entries will equal the sum of the entries of the initial vector $\mathbf{x}_0$ because this sum is preserved for every multiplication by A since

$$\sum_i \sum_j a_{ij} x_j = \sum_j x_j \left(\sum_i a_{ij} \right) = \sum_j x_j.$$

Here is an example. It is the same example as the one above but here it will involve the long time limit.

Example 12.2.6 *Consider the migration matrix* $\begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix}$ *for locations* $1, 2,$ *and* 3. *Suppose initially there are 100 residents in location 1, 200 in location 2 and 400 in location 4. Find the population in the three locations after a long time.*

You just need to find the eigenvector which goes with the eigenvalue 1 and then normalize it so the sum of its entries equals the sum of the entries of the initial vector. Thus you need to find a solution to

$$\left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} .6 & 0 & .1 \\ .2 & .8 & 0 \\ .2 & .2 & .9 \end{pmatrix} \right) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

The augmented matrix is

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

and its row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & -.25 & 0 \\ 0 & 1 & -.25 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Therefore, the eigenvectors are

$$s\begin{pmatrix} \frac{1}{4} & \frac{1}{4} & 1 \end{pmatrix}^T$$

and all that remains is to choose the value of s such that

$$\frac{1}{4}s+\frac{1}{4}s+s=100+200+400$$

This yields $s=\frac{1400}{3}$ and so the long time limit would equal

$$\frac{1400}{3}\begin{pmatrix} (1/4) \\ (1/4) \\ 1 \end{pmatrix}=\begin{pmatrix} 116.666666666666\,7 \\ 116.666666666666\,7 \\ 466.666666666666\,7 \end{pmatrix}.$$

You would of course need to round these numbers off. You see that you are not far off after just 10 units of time. Therefore, you might consider this as a useful procedure because it is probably easier to solve a simple system of equations than it is to raise a matrix to a large power.

Example 12.2.7 *Suppose a migration matrix is* $\begin{pmatrix} \frac{1}{5} & \frac{1}{2} & \frac{1}{5} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{11}{20} & \frac{1}{4} & \frac{3}{10} \end{pmatrix}$. *Find the comparison between the populations in the three locations after a long time.*

This amounts to nothing more than finding the eigenvector for $\lambda=1$. Solve

$$\left(\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}-\begin{pmatrix} \frac{1}{5} & \frac{1}{2} & \frac{1}{5} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{11}{20} & \frac{1}{4} & \frac{3}{10} \end{pmatrix}\right)\begin{pmatrix} x \\ y \\ z \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The augmented matrix is

$$\left(\begin{array}{ccc|c} \frac{4}{5} & -\frac{1}{2} & -\frac{1}{5} & 0 \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{2} & 0 \\ -\frac{11}{20} & -\frac{1}{4} & \frac{7}{10} & 0 \end{array}\right)$$

The row echelon form is

$$\begin{pmatrix} 1 & 0 & -\frac{16}{19} & 0 \\ 0 & 1 & -\frac{18}{19} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and so an eigenvector is

$$\begin{pmatrix} 16 & 18 & 19 \end{pmatrix}^T$$

Thus there will be $\frac{18}{16}^{th}$ more in location 2 than in location 1. There will be $\frac{19}{18}^{th}$ more in location 3 than in location 2.

You see the eigenvalue problem makes these sorts of determinations fairly simple.

There are many other things which can be said about these sorts of **migration problems**. They include things like the gambler's ruin problem which asks for the probability that a compulsive gambler will eventually lose all his money. However those problems are not so easy although they still involve eigenvalues and eigenvectors.

12.2.3 Discrete Dynamical Systems

The migration matrices discussed above give an example of a discrete dynamical system. They are discrete, not because they are somehow tactful and polite but because they involve discrete values taken at a sequence of points rather than on a whole interval of time. An example of a situation which can be studied in this way is a predator prey model. Consider the following model where x is the number of prey and y the number of predators. These are functions of $k \in \mathbb{N}$ where $1, 2, \cdots$ are the ends of intervals of time which may be of interest in the problem.

$$\begin{pmatrix} x(n+1) \\ y(n+1) \end{pmatrix} = \begin{pmatrix} A & -B \\ C & D \end{pmatrix} \begin{pmatrix} x(n) \\ y(n) \end{pmatrix}$$

This says that x increases if there are more x and decreases as there are more y. As for y, it increases if there are more y and also if there are more x.

Example 12.2.8 *Suppose a dynamical system is of the form*

$$\begin{pmatrix} x(n+1) \\ y(n+1) \end{pmatrix} = \begin{pmatrix} 1.5 & -0.5 \\ 1.0 & 0 \end{pmatrix} \begin{pmatrix} x(n) \\ y(n) \end{pmatrix}$$

Find solutions to the dynamical system for given initial conditions.

In this case, the eigenvalues of the matrix are 1, and .5. The matrix is of the form

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

and so given an initial condition

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

the solution to the dynamical system is

$$\begin{aligned} \begin{pmatrix} x(n) \\ y(n) \end{pmatrix} &= \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & .5 \end{pmatrix}^n\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & (.5)^n \end{pmatrix}\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \\ &= \begin{pmatrix} y_0\left((.5)^n - 1\right) - x_0\left((.5)^n - 2\right) \\ y_0\left(2(.5)^n - 1\right) - x_0\left(2(.5)^n - 2\right) \end{pmatrix} \end{aligned}$$

In the limit as $n \to \infty$, you get

$$\begin{pmatrix} 2x_0 - y_0 \\ 2x_0 - y_0 \end{pmatrix}$$

Thus for large n,

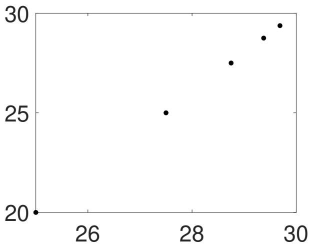
$$\begin{pmatrix} x(n) \\ y(n) \end{pmatrix} \approx \begin{pmatrix} 2x_0 - y_0 \\ 2x_0 - y_0 \end{pmatrix}$$

Letting the initial condition be

$$\begin{pmatrix} 20 \\ 10 \end{pmatrix}$$

one can graph these solutions for various values of n. Here are the solutions for values of n between 1 and 5

$$\begin{pmatrix} 25.0 \\ 20.0 \end{pmatrix}\begin{pmatrix} 27.5 \\ 25.0 \end{pmatrix}\begin{pmatrix} 28.75 \\ 27.5 \end{pmatrix}\begin{pmatrix} 29.375 \\ 28.75 \end{pmatrix}\begin{pmatrix} 29.688 \\ 29.375 \end{pmatrix}$$



Another very different kind of behavior is also observed. It is possible for the ordered pairs to spiral around the origin.

Example 12.2.9 *Suppose a dynamical system is of the form*

$$\begin{pmatrix} x(n+1) \\ y(n+1) \end{pmatrix} = \begin{pmatrix} 0.7 & 0.7 \\ -0.7 & 0.7 \end{pmatrix} \begin{pmatrix} x(n) \\ y(n) \end{pmatrix}$$

Find solutions to the dynamical system for given initial conditions.

In this case, the eigenvalues are complex, $.7+.7i$ and $.7-.7i$. Suppose the initial condition is

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

what is a formula for the solutions to the dynamical system? Some computations show that the eigen pairs are

$$\begin{pmatrix} 1 \\ i \end{pmatrix} \longleftrightarrow .7+.7i, \begin{pmatrix} 1 \\ -i \end{pmatrix} \longleftrightarrow .7-.7i$$

Thus the matrix is of the form

$$\begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \begin{pmatrix} .7+.7i & 0 \\ 0 & .7-.7i \end{pmatrix} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2}i \\ \frac{1}{2} & \frac{1}{2}i \end{pmatrix}$$

and so,

$$\begin{pmatrix} x(n) \\ y(n) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \begin{pmatrix} (.7+.7i)^n & 0 \\ 0 & (.7-.7i)^n \end{pmatrix} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2}i \\ \frac{1}{2} & \frac{1}{2}i \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

The explicit solution is given by

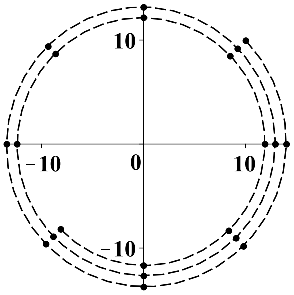
$$\begin{aligned} x &= x_0 \left(\frac{1}{2}((0.7-0.7i))^n + \frac{1}{2}((0.7+0.7i))^n \right) \\ &\quad + y_0 \left(\frac{1}{2} i ((0.7-0.7i))^n - \frac{1}{2} i ((0.7+0.7i))^n \right) \end{aligned}$$

$$\begin{aligned} y &= y_0 \left(\frac{1}{2}((0.7-0.7i))^n + \frac{1}{2}((0.7+0.7i))^n \right) \\ &\quad - x_0 \left(\frac{1}{2} i ((0.7-0.7i))^n - \frac{1}{2} i ((0.7+0.7i))^n \right) \end{aligned}$$

Suppose the initial condition is

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \end{pmatrix}$$

Then one obtains the following sequence of values which are graphed below by letting $n = 1, 2, \cdots, 20$



In this picture, the dots are the values and the dashed line is to help to picture what is happening.

These points are getting gradually closer to the origin, but they are circling the origin in the clockwise direction as they do so. Also, since both eigenvalues are slightly smaller than 1 in absolute value,

$$\lim_{n\to\infty}\begin{pmatrix} x(n) \\ y(n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

This type of behavior along with complex eigenvalues is typical of the deviations from an equilibrium point in the Lotka Volterra system of differential equations which is a famous model for predator prey interactions. These differential equations are given by

$$X' = X(a - bY),\ Y' = -Y(c - dX)$$

where a, b, c, d are positive constants. For example, you might have X be the population of moose and Y the population of wolves on an island.

Note how reasonable these equations are. The top says that the rate at which the moose population increases would be aX if there were no predators Y. However, this is modified by multiplying instead by $(a - bY)$ because if there are predators, these will militate against the population of moose. By definition, the wolves eat the moose and when the moose is eaten, it is not around anymore to make new moose. The more predators there are, the more pronounced is this effect. As to the predator equation, you can see that the equations predict that if there are many prey around, then the rate of growth of the predators would seem to be high. However, this is modified by the term $-cY$ because if there are many predators, there would be competition for the available food supply and this would tend to decrease Y'.

The behavior near an equilibrium point, which is a point where the right side of the differential equations equals zero, is of great interest. In this case, the equilibrium point is

$$Y = \frac{a}{b}, X = \frac{c}{d}$$

Then one defines new variables according to the formula

$$x + \frac{c}{d} = X,\ Y = y + \frac{a}{b}$$

In terms of these new variables, the differential equations become

$$\begin{aligned} x' &= \left(x + \frac{c}{d}\right)\left(a - b\left(y + \frac{a}{b}\right)\right) \\ y' &= -\left(y + \frac{a}{b}\right)\left(c - d\left(x + \frac{c}{d}\right)\right) \end{aligned}$$

Multiplying out the right sides yields

$$\begin{aligned} x' &= -bxy - b\frac{c}{d}y \\ y' &= dxy + \frac{a}{b}dx \end{aligned}$$

The interest is for x,y small and so these equations are essentially equal to

$$x' = -b\frac{c}{d}y,\ y' = \frac{a}{b}dx$$

Replace x' with the difference quotient $\frac{x(t+h)-x(t)}{h}$ where h is a small positive number and y' with a similar difference quotient. For example one could have h correspond to one day or even one hour. Thus, for h small enough, the following would seem to be a good approximation to the differential equations.

$$\begin{aligned} x(t+h) &= x(t) - hb\frac{c}{d}y \\ y(t+h) &= y(t) + h\frac{a}{b}dx \end{aligned}$$

Let $1,2,3,\cdots$ denote the ends of discrete intervals of time having length h chosen above. Then the above equations take the form

$$\left(\begin{array}{c} x(n+1) \\ y(n+1) \end{array}\right) = \left(\begin{array}{cc} 1 & \frac{-hbc}{d} \\ \frac{had}{b} & 1 \end{array}\right)\left(\begin{array}{c} x(n) \\ y(n) \end{array}\right)$$

Note that the eigenvalues of this matrix are always complex.

You are not interested in time intervals of length h for h very small. Instead, you are interested in much longer lengths of time. Thus, replacing the time interval with mh,

$$\left(\begin{array}{c} x(n+m) \\ y(n+m) \end{array}\right) = \left(\begin{array}{cc} 1 & \frac{-hbc}{d} \\ \frac{had}{b} & 1 \end{array}\right)^m\left(\begin{array}{c} x(n) \\ y(n) \end{array}\right)$$

For example, if $m=2$, you would have

$$\left(\begin{array}{c} x(n+2) \\ y(n+2) \end{array}\right) = \left(\begin{array}{cc} 1-ach^2 & -2b\frac{c}{d}h \\ 2\frac{a}{b}dh & 1-ach^2 \end{array}\right)\left(\begin{array}{c} x(n) \\ y(n) \end{array}\right)$$

Note that the eigenvalues of the new matrix will likely still be complex. You can also notice that the upper right corner will be negative by considering higher powers of the matrix. Thus letting $1,2,3,\cdots$ denote the ends of discrete intervals of time, the desired discrete dynamical system is of the form

$$\left(\begin{array}{c} x(n+1) \\ y(n+1) \end{array}\right) = \left(\begin{array}{cc} A & -B \\ C & D \end{array}\right)\left(\begin{array}{c} x(n) \\ y(n) \end{array}\right)$$

where A,B,C,D are positive constants and the matrix will likely have complex eigenvalues because it is a power of a matrix which has complex eigenvalues.

You can see from the above discussion that if the eigenvalues of the matrix used to define the dynamical system are less than 1 in absolute value, then the origin is stable in the sense that as $n \to \infty$, the solution converges to the origin. If either eigenvalue is larger than 1 in absolute value, then the solutions to the dynamical system will usually be unbounded, unless the initial condition is chosen very carefully. The next example exhibits the case where one eigenvalue is larger than 1 and the other is smaller than 1.

Example 12.2.10 *The Fibonacci sequence is the sequence which is defined recursively in the form*

$$x(0) = 1 = x(1),\ x(n+2) = x(n+1) + x(n)$$

This sequence is extremely important in the study of reproducing rabbits. It can be considered as a dynamical system as follows. Let $y(n) = x(n+1)$. Then the above recurrence relation can be written as

$$\begin{pmatrix} x(n+1) \\ y(n+1) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x(n) \\ y(n) \end{pmatrix}, \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The eigenvectors and eigenvalues of the matrix are

$$\begin{pmatrix} -\frac{1}{2}\sqrt{5} - \frac{1}{2} \\ 1 \end{pmatrix} \leftrightarrow \frac{1}{2} - \frac{1}{2}\sqrt{5}, \begin{pmatrix} \frac{1}{2}\sqrt{5} - \frac{1}{2} \\ 1 \end{pmatrix} \leftrightarrow \frac{1}{2}\sqrt{5} + \frac{1}{2}$$

You can see from a short computation that one of these is smaller than 1 in absolute value while the other is larger than 1 in absolute value.

$$\begin{pmatrix} \frac{1}{2}\sqrt{5} - \frac{1}{2} & -\frac{1}{2}\sqrt{5} - \frac{1}{2} \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{2}\sqrt{5} - \frac{1}{2} & -\frac{1}{2}\sqrt{5} - \frac{1}{2} \\ 1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{2}\sqrt{5} + \frac{1}{2} & 0 \\ 0 & \frac{1}{2} - \frac{1}{2}\sqrt{5} \end{pmatrix}$$

Then it follows that for the given initial condition the solution to this dynamical system is of the form

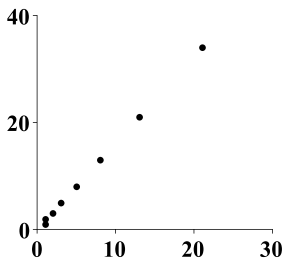
$$\begin{pmatrix} x(n) \\ y(n) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\sqrt{5} - \frac{1}{2} & -\frac{1}{2}\sqrt{5} - \frac{1}{2} \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \left(\frac{1}{2}\sqrt{5} + \frac{1}{2}\right)^n & 0 \\ 0 & \left(\frac{1}{2} - \frac{1}{2}\sqrt{5}\right)^n \end{pmatrix} \cdot$$

$$\begin{pmatrix} \frac{1}{5}\sqrt{5} & \frac{1}{10}\sqrt{5} + \frac{1}{2} \\ -\frac{1}{5}\sqrt{5} & \frac{1}{5}\sqrt{5}\left(\frac{1}{2}\sqrt{5} - \frac{1}{2}\right) \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

It follows that

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

This might not be the first thing you would think of. Here is a picture of the ordered pairs $(x(n), y(n))$ for $n = 0, 1, \cdots, n$. There is so much more that can be said about dynamical systems.



It is a major topic of study in differential equations and what is given above is just an introduction.

12.3 The Estimation of Eigenvalues

There are many other important applications of eigenvalue problems. We have just given a few such applications here. As pointed out, this is a very hard problem but sometimes you don't need to find the eigenvalues exactly. There are ways to estimate the eigenvalues for matrices from just looking at the matrix. The most famous is known as **Gerschgorin's theorem**. This theorem gives a rough idea where the eigenvalues are just from looking at the matrix.

Theorem 12.3.1 *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 $\sigma(A)$, the spectrum of A.

Theorem 12.3.2 *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 $A = (a_{ij})$, let $|x_k| \geq |x_j|$ for all x_j. Thus $|x_k| \neq 0$.

$$\sum_{j \neq k} a_{kj} x_j = (\lambda - a_{kk}) x_k.$$

Then $|x_k| \sum_{j \neq k} |a_{kj}| \geq \sum_{j \neq k} |a_{kj}| |x_j| \geq \left| \sum_{j \neq k} a_{kj} x_j \right| = |\lambda - a_{ii}| |x_k|$. Now dividing by $|x_k|$, it follows λ is contained in the k^{th} Gerschgorin disc. ■

Example 12.3.3 *Suppose the matrix is*

$$A=\left(\begin{array}{ccc}21 & -16 & -6\\ 14 & 60 & 12\\ 7 & 8 & 38\end{array}\right)$$

Estimate the eigenvalues.

The exact eigenvalues are $35,56,$ and 28. The Gerschgorin disks are

$$D_1=\{\lambda\in\mathbb{C}:|\lambda-21|\leq 22\},$$

$$D_2=\{\lambda\in\mathbb{C}:|\lambda-60|\leq 26\},$$

and

$$D_3=\{\lambda\in\mathbb{C}:|\lambda-38|\leq 15\}.$$

Gerschgorin's theorem says these three disks contain the eigenvalues. Now 35 is in $D_3, 56$ is in D_2 and 28 is in D_1.

More can be said when the Gerschgorin disks are disjoint but this is an advanced topic which requires the theory of functions of a complex variable. If you are interested and have a background in complex variable techniques, this is in [13]

12.4 MATLAB and Eigenvalues

To find the eigenvalues enter A and follow with ;. Then type eig(A) and press return. It will give numerical approximation of the eigenvalues. If you want to have it find the exact values, you type eig(sym(A)) and press return. For example, if you type >>A=[1,1,0;-1,0,-1;2,1,3]; and then eig(sym(A)) and return, you will get the eigenvalues 1,1,2 listed in a column. This is correct. The matrix has a repeated eigenvalue of 1. If you want to get the eigenvectors also, you would type >>A=[1,1,0;-1,0,-1;2,1,3]; and then [V,D]=eig(sym(A)) and enter or if you want numerical answers, which will sometimes be all that is available, you would type [V,D]=eig(A). It will find the matrix V such that $AV=VD$ where D is a diagonal. In the case just considered, it will only find two columns for V because this is a defective matrix. In general, however, this would give $V^{-1}AV=D$ and the columns of V are the eigenvectors.

12.5 Exercises

1. State the eigenvalue problem from an algebraic perspective.

2. State the eigenvalue problem from a geometric perspective.

3. Consider the linear transformation which projects all vectors in $\mathbb{R}^2$ onto the span of the vector $(1,2)$. Show that the matrix of this linear transformation is

$$\left(\begin{array}{cc}1/5 & 2/5\\ 2/5 & 4/5\end{array}\right)$$

Now based on geometric considerations only, show that 1 is an eigenvalue and that an eigenvector is $(1,2)^T$. Also explain why 0 will also be an eigenvalue.

4. If A is the matrix of a linear transformation which rotates all vectors in $\mathbb{R}^2$ through 30°, explain why A cannot have any real eigenvalues.

5. If A is an $n \times n$ matrix and c is a nonzero constant, compare the eigenvalues of A and cA.

6. If A is an invertible $n \times n$ matrix, compare the eigenvalues of A and A^{-1}. More generally, for m an arbitrary integer, compare the eigenvalues of A and A^m.

7. Let A, B be invertible $n \times n$ matrices which commute. That is, $AB = BA$. Suppose $\mathbf{x}$ is an eigenvector of B. Show that then $A\mathbf{x}$ must also be an eigenvector for B.

8. Suppose A is an $n \times n$ matrix and it satisfies $A^m = A$ for some m a positive integer larger than 1. Show that if λ is an eigenvalue of A then $|\lambda|$ equals either 0 or 1.

9. Show that if $A\mathbf{x} = \lambda\mathbf{x}$ and $A\mathbf{y} = \lambda\mathbf{y}$, then whenever a, b are scalars,
$$A\left(a\mathbf{x} + b\mathbf{y}\right) = \lambda\left(a\mathbf{x} + b\mathbf{y}\right).$$
Does this imply that $a\mathbf{x} + b\mathbf{y}$ is an eigenvector? Explain.

10. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} -1 & -1 & 7 \\ -1 & 0 & 4 \\ -1 & -1 & 5 \end{pmatrix}.$$
Determine whether the matrix is defective.

11. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} -3 & -7 & 19 \\ -2 & -1 & 8 \\ -2 & -3 & 10 \end{pmatrix}.$$
Determine whether the matrix is defective.

12. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} -7 & -12 & 30 \\ -3 & -7 & 15 \\ -3 & -6 & 14 \end{pmatrix}.$$
Determine whether the matrix is defective.

13. Find the eigenvalues and eigenvectors of the matrix
$$\begin{pmatrix} 7 & -2 & 0 \\ 8 & -1 & 0 \\ -2 & 4 & 6 \end{pmatrix}.$$
Determine whether the matrix is defective.

14. Find the eigenvalues and eigenvectors of the matrix

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

Determine whether the matrix is defective.

15. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 6 & 8 & -23 \\ 4 & 5 & -16 \\ 3 & 4 & -12 \end{pmatrix}$$

Determine whether the matrix is defective.

16. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 5 & 2 & -5 \\ 12 & 3 & -10 \\ 12 & 4 & -11 \end{pmatrix}.$$

Determine whether the matrix is defective.

17. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 20 & 9 & -18 \\ 6 & 5 & -6 \\ 30 & 14 & -27 \end{pmatrix}.$$

Determine whether the matrix is defective.

18. 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.

19. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} 3 & -1 & -2 \\ 11 & 3 & -9 \\ 8 & 0 & -6 \end{pmatrix}.$$

Determine whether the matrix is defective.

20. Find the eigenvalues and eigenvectors of the matrix

$$\begin{pmatrix} -2 & 1 & 2 \\ -11 & -2 & 9 \\ -8 & 0 & 7 \end{pmatrix}.$$

Determine whether the matrix is defective.

21. Find the eigenvalues and eigenvectors of the matrix

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

Determine whether the matrix is defective.

22. Find the complex eigenvalues and eigenvectors of the matrix

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

23. 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.

24. 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.

25. 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.

26. 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.

27. 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.

28. Let A be a real 3×3 matrix which has a complex eigenvalue of the form $a + ib$ where $b \neq 0$. Could A be defective? Explain. Either give a proof or an example.

29. Let T be the linear transformation which reflects vectors about the x axis. Find a matrix for T and then find its eigenvalues and eigenvectors.

30. Let T be the linear transformation which rotates all vectors in $\mathbb{R}^2$ counterclockwise through an angle of $\pi/2$. Find a matrix of T and then find eigenvalues and eigenvectors.

31. Let A be the 2×2 matrix of the linear transformation which rotates all vectors in $\mathbb{R}^2$ through an angle of θ. For which values of θ does A have a real eigenvalue?

32. Let T be the linear transformation which reflects all vectors in $\mathbb{R}^3$ through the xy plane. Find a matrix for T and then obtain its eigenvalues and eigenvectors.

33. Find the principal direction for stretching for the matrix

$$\begin{pmatrix} \frac{13}{9} & \frac{2}{15}\sqrt{5} & \frac{8}{45}\sqrt{5} \\ \frac{2}{15}\sqrt{5} & \frac{6}{5} & \frac{4}{15} \\ \frac{8}{45}\sqrt{5} & \frac{4}{15} & \frac{61}{45} \end{pmatrix}.$$

The eigenvalues are 2 and 1.

34. Find the principal directions for the matrix

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

35. 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.

36. Suppose the migration matrix for three locations is

$$\left(\begin{array}{ccc} .1 & .1 & .3 \\ .3 & .7 & 0 \\ .6 & .2 & .7 \end{array}\right).$$

Find a comparison for the populations in the three locations after a long time.

37. You own a trailer rental company in a large city and you have four locations, one in the South East, one in the North East, one in the North West, and one in the South West. Denote these locations by SE,NE,NW, and SW respectively. Suppose you observe that in a typical day, .8 of the trailers starting in SE stay in SE, .1 of the trailers in NE go to SE, .1 of the trailers in NW end up in SE, .2 of the trailers in SW end up in SE, .1 of the trailers in SE end up in NE,.7 of the trailers in NE end up in NE,.2 of the trailers in NW end up in NE,.1 of the trailers in SW end up in NE, .1 of the trailers in SE end up in NW, .1 of the trailers in NE end up in NW, .6 of the trailers in NW end up in NW, .2 of the trailers in SW end up in NW, 0 of the trailers in SE end up in SW, .1 of the trailers in NE end up in SW, .1 of the trailers in NW end up in SW, .5 of the trailers in SW end up in SW. You begin with 20 trailers in each location. Approximately how many will you have in each location after a long time? Will any location ever run out of trailers?

38. Let A be the $n \times n$, $n > 1$, matrix of the linear transformation which comes from the projection $\mathbf{v} \mapsto \text{proj}_{\mathbf{w}}(\mathbf{v})$. Show that A cannot be invertible. Also show that A has an eigenvalue equal to 1 and that for λ an eigenvalue, $|\lambda| \leq 1$.

39. Let $\mathbf{v}$ be a unit vector in $\mathbb{R}^n$ and let $A = I - 2\mathbf{v}\mathbf{v}^T$. Show that A has an eigenvalue equal to -1.

40. 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**. This is presented in the text. You should write it down in your own words filling in the details without looking at the text.

41. 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 40. 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. Like the last problem, you should try to do this yourself without consulting the text. These problems are a nice review of the meaning of matrix multiplication.

42. Suppose A is an $n \times n$ matrix which is **diagonally dominant**. This means

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|.$$

Show that A^{-1} must exist.

43. Is it possible for a nonzero matrix to have only 0 as an eigenvalue?

44. 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.

45. 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 $a-ib$.

46. 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.

47. 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.

48. A discreet dynamical system is of the form
$$\mathbf{x}(k+1) = A\mathbf{x}(k),\ \mathbf{x}(0) = \mathbf{x}_0$$
where A is an $n \times n$ matrix and $\mathbf{x}(k)$ is a vector in $\mathbb{R}^n$. Show first that
$$\mathbf{x}(k) = A^k\mathbf{x}_0$$
for all $k \geq 1$. If A is nondefective so that it has a basis of eigenvectors, $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ where
$$A\mathbf{v}_j = \lambda_j\mathbf{v}_j$$
you can write the initial condition $\mathbf{x}_0$ in a unique way as a linear combination of these eigenvectors. Thus
$$\mathbf{x}_0 = \sum_{j=1}^n a_j\mathbf{v}_j$$
Now explain why
$$\mathbf{x}(k) = \sum_{j=1}^n a_j A^k\mathbf{v}_j = \sum_{j=1}^n a_j\lambda_j^k\mathbf{v}_j$$
which gives a formula for $\mathbf{x}(k)$, the solution of the dynamical system.

49. Suppose A is an $n \times n$ matrix and let $\mathbf{v}$ be an eigenvector such that $A\mathbf{v} = \lambda\mathbf{v}$. Also suppose the characteristic polynomial of A is
$$\det(\lambda I - A) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0$$

Explain why

$$\left(A^n + a_{n-1}A^{n-1} + \cdots + a_1 A + a_0 I\right)\mathbf{v} = \mathbf{0}$$

If A is nondefective, give a very easy proof of the Cayley Hamilton theorem based on this. Recall this theorem says A satisfies its characteristic equation,

$$A^n + a_{n-1}A^{n-1} + \cdots + a_1 A + a_0 I = 0.$$

50. Suppose an $n \times n$ nondefective matrix A has only 1 and -1 as eigenvalues. Find A^{12}.

51. Suppose the characteristic polynomial of an $n \times n$ matrix A is $1 - \lambda^n$. Find A^{mn} where m is an integer. **Hint:** Note first that A is nondefective. Why?

52. Sometimes sequences come in terms of a recursion formula. An example is the Fibonacci sequence.

$$x_0 = 1 = x_1,\ x_{n+1} = x_n + x_{n-1}$$

Show this can be considered as a discreet dynamical system as follows.

$$\begin{pmatrix} x_{n+1} \\ x_n \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} x_n \\ x_{n-1} \end{pmatrix},\ \begin{pmatrix} x_1 \\ x_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Now use the technique of Problem 48 to find a formula for x_n. This was done in the chapter. Next change the initial conditions to $x_0 = 0, x_1 = 1$ and find the solution.

53. Let A be an $n \times n$ matrix having characteristic polynomial

$$\det(\lambda I - A) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0$$

Show that $a_0 = (-1)^n \det(A)$.

54. Find $\begin{pmatrix} \frac{3}{2} & 1 \\ -\frac{1}{2} & 0 \end{pmatrix}^{35}$. Next find

$$\lim_{n\to\infty} \begin{pmatrix} \frac{3}{2} & 1 \\ -\frac{1}{2} & 0 \end{pmatrix}^n$$

55. Find e^A where A is the matrix $\begin{pmatrix} \frac{3}{2} & 1 \\ -\frac{1}{2} & 0 \end{pmatrix}$ in the above problem.

56. Consider the dynamical system $\begin{pmatrix} x(n+1) \\ y(n+1) \end{pmatrix} = \begin{pmatrix} .8 & .8 \\ -.8 & .8 \end{pmatrix}\begin{pmatrix} x(n) \\ y(n) \end{pmatrix}$. Show eigenvalues and eigenvectors are $0.8 + 0.8i \longleftrightarrow \begin{pmatrix} -i \\ 1 \end{pmatrix}, 0.8 - 0.8i \longleftrightarrow \begin{pmatrix} i \\ 1 \end{pmatrix}$. Find a formula for the solution to the dynamical system for given initial condition $(x_0, y_0)^T$. Show that the magnitude of $(x(n), y(n))^T$ must diverge provided the initial condition is not zero. Next graph the vector field for

$$\begin{pmatrix} .8 & .8 \\ -.8 & .8 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} x \\ y \end{pmatrix}$$

Note that this vector field seems to indicate a conclusion different than what you just obtained. Therefore, in this context of discreet dynamical systems the consideration of such a picture is not all that reliable.

Chapter 13

Matrices And The Inner Product

13.1 Symmetric And Orthogonal Matrices

13.1.1 Orthogonal Matrices

Remember that to find the inverse of a matrix was often a long process. However, it was very easy to take the transpose of a matrix. For some matrices, the transpose equals the inverse and when the matrix has all real entries, and this is true, it is called an orthogonal matrix. Recall the following definition given earlier.

Definition 13.1.1 *A real $n \times n$ matrix U is called an **Orthogonal** matrix if $UU^T = U^T U = I$.*

Example 13.1.2 *Show the matrix*

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

is orthogonal.

$$UU^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Example 13.1.3 *Let $U = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}$. Is U orthogonal?*

The answer is yes. This is because the columns form an orthonormal set of vectors as well as the rows. As discussed above this is equivalent to $U^T U = I$.

$$U^T U = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}^T \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

When you say that U is orthogonal, you are saying that $\sum_j U_{ij}U_{jk}^T = \sum_j U_{ij}U_{kj} = \delta_{ik}$. In words, the dot product of the i^{th} row of U with the k^{th} row gives 1 if $i = k$ and 0 if $i \neq k$. The same is true of the columns because $U^TU = I$ also. Therefore, $\sum_j U_{ij}^T U_{jk} = \sum_j U_{ji}U_{jk} = \delta_{ik}$ which says that the one column dotted with another column gives 1 if the two columns are the same and 0 if the two columns are different.

More succinctly, this states that if $\mathbf{u}_1, \cdots, \mathbf{u}_n$ are the columns of U, an orthogonal matrix, then

$$\mathbf{u}_i \cdot \mathbf{u}_j = \delta_{ij} \equiv \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases} . \tag{13.1}$$

Definition 13.1.4 *A set of vectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ is said to be an* ***orthonormal*** *set if 13.1.*

Theorem 13.1.5 *If $\{\mathbf{u}_1, \cdots, \mathbf{u}_m\}$ is an orthonormal set of vectors then it is linearly independent.*

Proof: Using the properties of the dot product, $\mathbf{0} \cdot \mathbf{u} = (\mathbf{0} + \mathbf{0}) \cdot \mathbf{u} = \mathbf{0} \cdot \mathbf{u} + \mathbf{0} \cdot \mathbf{u}$ and so, subtracting $\mathbf{0} \cdot \mathbf{u}$ from both sides yields $\mathbf{0} \cdot \mathbf{u} = 0$. Now suppose $\sum_j c_j \mathbf{u}_j = \mathbf{0}$. Then from the properties of the dot product,

$$c_k = \sum_j c_j \delta_{jk} = \sum_j c_j (\mathbf{u}_j \cdot \mathbf{u}_k) = \left(\sum_j c_j \mathbf{u}_j\right) \cdot \mathbf{u}_k = \mathbf{0} \cdot \mathbf{u}_k = 0.$$

Since k was arbitrary, this shows that each $c_k = 0$ and this has shown that if $\sum_j c_j \mathbf{u}_j = \mathbf{0}$, then each $c_j = 0$. This is what it means for the set of vectors to be linearly independent. ■

Example 13.1.6 *Let $U = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{\sqrt{6}}{3} \end{pmatrix}$. Is U an orthogonal matrix?*

The answer is yes. This is because the columns (rows) form an orthonormal set of vectors.

The importance of orthogonal matrices is that they change components of vectors relative to different Cartesian coordinate systems. Geometrically, the orthogonal matrices are exactly those which preserve all distances in the sense that if $\mathbf{x} \in \mathbb{R}^n$ and U is orthogonal, then $||U\mathbf{x}|| = ||\mathbf{x}||$ because

$$||U\mathbf{x}||^2 = (U\mathbf{x})^T U\mathbf{x} = \mathbf{x}^T U^T U\mathbf{x} = \mathbf{x}^T I\mathbf{x} = ||\mathbf{x}||^2$$

Observation 13.1.7 *Suppose U is an orthogonal matrix. Then* $\det(U) = \pm 1$.

This is easy to see from the properties of determinants. Thus

$$\det(U)^2 = \det\left(U^T\right)\det(U) = \det\left(U^TU\right) = \det(I) = 1.$$

Orthogonal matrices are divided into two classes, proper and improper. The proper orthogonal matrices are those whose determinant equals 1 and the improper ones are those whose determinants equal -1. The reason for the distinction is that the improper orthogonal matrices are sometimes considered to have no physical significance since they cause a change in orientation which would correspond to material passing through itself in a non physical manner. Thus in considering which coordinate systems must be considered in certain applications, you only need to consider those which are related by a proper orthogonal transformation. Geometrically, the linear transformations determined by the proper orthogonal matrices correspond to the composition of rotations.

13.1.2 Symmetric And Skew Symmetric Matrices

Definition 13.1.8 *A real $n \times n$ matrix A, is* ***symmetric*** *if $A^T = A$. If $A = -A^T$, then A is called* ***skew symmetric.***

Theorem 13.1.9 *The eigenvalues of a real symmetric matrix are real. The eigenvalues of a real skew symmetric matrix are 0 or pure imaginary.*

Proof: The proof of this theorem is in [13]. It is best understood as a special case of more general considerations. However, here is a proof in this special case.

Recall that for a complex number $a+ib$, the complex conjugate, denoted by $\overline{a+ib}$ is given by the formula $\overline{a+ib} = a-ib$. The notation, $\overline{\mathbf{x}}$ will denote the vector which has every entry replaced by its complex conjugate.

Suppose A is a real symmetric matrix and $A\mathbf{x} = \lambda\mathbf{x}$. Then

$$\overline{\lambda}\overline{\mathbf{x}}^T\mathbf{x} = \left(\overline{A\mathbf{x}}\right)^T\mathbf{x} = \overline{\mathbf{x}}^T A^T\mathbf{x} = \overline{\mathbf{x}}^T A\mathbf{x} = \lambda\overline{\mathbf{x}}^T\mathbf{x}.$$

Dividing by $\overline{\mathbf{x}}^T\mathbf{x}$ on both sides yields $\overline{\lambda} = \lambda$ which says λ is real. (Why?)

Next suppose $A = -A^T$ so A is skew symmetric and $A\mathbf{x} = \lambda\mathbf{x}$. Then

$$\overline{\lambda}\overline{\mathbf{x}}^T\mathbf{x} = \left(\overline{A\mathbf{x}}\right)^T\mathbf{x} = \overline{\mathbf{x}}^T A^T\mathbf{x} = -\overline{\mathbf{x}}^T A\mathbf{x} = -\lambda\overline{\mathbf{x}}^T\mathbf{x}$$

and so, dividing by $\overline{\mathbf{x}}^T\mathbf{x}$ as before, $\overline{\lambda} = -\lambda$. Letting $\lambda = a+ib$, this means $a-ib = -a-ib$ and so $a = 0$. Thus λ is pure imaginary. ■

Example 13.1.10 *Let $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. This is a skew symmetric matrix. Find its eigenvalues.*

Its eigenvalues are obtained by solving the equation $\det\begin{pmatrix} -\lambda & -1 \\ 1 & -\lambda \end{pmatrix} = \lambda^2+1 = 0$. You see the eigenvalues are $\pm i$, pure imaginary.

Example 13.1.11 *Let $A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$. This is a symmetric matrix. Find its eigenvalues.*

Its eigenvalues are obtained by solving the equation, $\det\begin{pmatrix} 1-\lambda & 2 \\ 2 & 3-\lambda \end{pmatrix} = -1-4\lambda+\lambda^2 = 0$ and the solution is $\lambda = 2+\sqrt{5}$ and $\lambda = 2-\sqrt{5}$.

Definition 13.1.12 *An $n \times n$ matrix $A = (a_{ij})$ is called a **diagonal matrix** if $a_{ij} = 0$ whenever $i \neq j$. For example, a diagonal matrix is of the form indicated below where $*$ denotes a number.*

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

Theorem 13.1.13 *Let A be a real symmetric matrix. Then there exists an orthogonal matrix U such that $U^T AU$ is a diagonal matrix. Moreover, the diagonal entries are the eigenvalues of A.*

Proof: The proof is given later.

Corollary 13.1.14 *If A is a real $n \times n$ symmetric matrix, then there exists an orthonormal set of eigenvectors, $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$.*

Proof: Since A is symmetric, then by Theorem 13.1.13, there exists an orthogonal matrix U such that $U^T AU = D$, a diagonal matrix whose diagonal entries are the eigenvalues of A. Therefore, since A is symmetric and all the matrices are real,

$$\overline{D} = \overline{D^T} = \overline{U^T A^T U} = U^T A^T U = U^T AU = D$$

showing D is real because each entry of D equals its complex conjugate.[1]

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, $U^T AU = D$ implies

$$AU = \begin{pmatrix} A\mathbf{u}_1 & A\mathbf{u}_2 & \cdots & A\mathbf{u}_n \end{pmatrix} = UD = \begin{pmatrix} \lambda_1\mathbf{u}_1 & \lambda_2\mathbf{u}_2 & \cdots & \lambda_n\mathbf{u}_n \end{pmatrix}$$

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 orthogonal, the ij^{th} entry of $U^T U$ equals δ_{ij} and so $\delta_{ij} = \mathbf{u}_i^T \mathbf{u}_j = \mathbf{u}_i \cdot \mathbf{u}_j$. This proves the corollary because it shows the vectors $\{\mathbf{u}_i\}$ form an orthonormal basis. ■

Example 13.1.15 *Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix*

$$\begin{pmatrix} \frac{19}{9} & -\frac{8}{15}\sqrt{5} & \frac{2}{45}\sqrt{5} \\ -\frac{8}{15}\sqrt{5} & -\frac{1}{5} & -\frac{16}{15} \\ \frac{2}{45}\sqrt{5} & -\frac{16}{15} & \frac{94}{45} \end{pmatrix}$$

given that the eigenvalues are 3, -1, and 2.

[1] Recall that for a complex number, $x+iy$, the complex conjugate, denoted by $\overline{x+iy}$ is defined as $x-iy$.

The augmented matrix which needs to be row reduced to find the eigenvectors for $\lambda = 3$ is

$$\left(\begin{array}{ccc|c} \frac{19}{9}-3 & -\frac{8}{15}\sqrt{5} & \frac{2}{45}\sqrt{5} & 0 \\ -\frac{8}{15}\sqrt{5} & -\frac{1}{5}-3 & -\frac{16}{15} & 0 \\ \frac{2}{45}\sqrt{5} & -\frac{16}{15} & \frac{94}{45}-3 & 0 \end{array}\right)$$

and the row reduced echelon form for this is

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

Therefore, eigenvectors for $\lambda = 3$ are $z\begin{pmatrix} \frac{1}{2}\sqrt{5} & -\frac{3}{4} & 1 \end{pmatrix}^T$ where $z \neq 0$.

The augmented matrix, which must be row reduced to find the eigenvectors for $\lambda = -1$, is

$$\left(\begin{array}{ccc|c} \frac{19}{9}+1 & -\frac{8}{15}\sqrt{5} & \frac{2}{45}\sqrt{5} & 0 \\ -\frac{8}{15}\sqrt{5} & -\frac{1}{5}+1 & -\frac{16}{15} & 0 \\ \frac{2}{45}\sqrt{5} & -\frac{16}{15} & \frac{94}{45}+1 & 0 \end{array}\right)$$

and the row reduced echelon form is

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

Therefore, the eigenvectors for $\lambda = -1$ are $z\begin{pmatrix} \frac{1}{2}\sqrt{5} & 3 & 1 \end{pmatrix}^T$, $z \neq 0$

The augmented matrix which must be row reduced to find the eigenvectors for $\lambda = 2$ is

$$\left(\begin{array}{ccc|c} \frac{19}{9}-2 & -\frac{8}{15}\sqrt{5} & \frac{2}{45}\sqrt{5} & 0 \\ -\frac{8}{15}\sqrt{5} & -\frac{1}{5}-2 & -\frac{16}{15} & 0 \\ \frac{2}{45}\sqrt{5} & -\frac{16}{15} & \frac{94}{45}-2 & 0 \end{array}\right)$$

and its row reduced echelon form is

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

so the eigenvectors for $\lambda = 2$ are $z\begin{pmatrix} -\frac{2}{5}\sqrt{5} & 0 & 1 \end{pmatrix}^T$, $z \neq 0$.

It remains to find an orthonormal basis. You can check that the dot product of any of these vectors with another of them gives zero and so it suffices choose z in each case such that the resulting vector has length 1. First consider the vectors for $\lambda = 3$. It is required to choose z such that $z\begin{pmatrix} \frac{1}{2}\sqrt{5} & -\frac{3}{4} & 1 \end{pmatrix}^T$ is a unit vector. In other words, you need

$$z\begin{pmatrix} \frac{1}{2}\sqrt{5} \\ -\frac{3}{4} \\ 1 \end{pmatrix} \cdot z\begin{pmatrix} \frac{1}{2}\sqrt{5} \\ -\frac{3}{4} \\ 1 \end{pmatrix} = 1.$$

But the above dot product equals $\frac{45}{16}z^2$ and this equals 1 when $z = \frac{4}{15}\sqrt{5}$. Therefore, the eigenvector which is desired is $\begin{pmatrix} \frac{2}{3} & -\frac{1}{5}\sqrt{5} & \frac{4}{15}\sqrt{5} \end{pmatrix}^T$.

Next find the eigenvector for $\lambda = -1$. The same process requires that $1 = \frac{45}{4}z^2$ which happens when $z = \frac{2}{15}\sqrt{5}$. Therefore, an eigenvector for $\lambda = -1$ which has unit length is

$$\frac{2}{15}\sqrt{5}\begin{pmatrix} \frac{1}{2}\sqrt{5} \\ 3 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{5}\sqrt{5} \\ \frac{2}{15}\sqrt{5} \end{pmatrix}.$$

Finally, consider $\lambda = 2$. This time you need $1 = \frac{9}{5}z^2$ which occurs when $z = \frac{1}{3}\sqrt{5}$. Therefore, the eigenvector is

$$\frac{1}{3}\sqrt{5}\begin{pmatrix} -\frac{2}{5}\sqrt{5} \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\frac{2}{3} \\ 0 \\ \frac{1}{3}\sqrt{5} \end{pmatrix}.$$

Now recall that the vectors form an orthonormal set of vectors if the matrix having them as columns is orthogonal. That matrix is

$$\begin{pmatrix} \frac{2}{3} & \frac{1}{3} & -\frac{2}{3} \\ -\frac{1}{5}\sqrt{5} & \frac{2}{5}\sqrt{5} & 0 \\ \frac{4}{15}\sqrt{5} & \frac{2}{15}\sqrt{5} & \frac{1}{3}\sqrt{5} \end{pmatrix}.$$

Is this orthogonal? To find out, multiply by its transpose. Thus

$$\begin{pmatrix} \frac{2}{3} & -\frac{1}{5}\sqrt{5} & \frac{4}{15}\sqrt{5} \\ \frac{1}{3} & \frac{2}{5}\sqrt{5} & \frac{2}{15}\sqrt{5} \\ -\frac{2}{3} & 0 & \frac{1}{3}\sqrt{5} \end{pmatrix}\begin{pmatrix} \frac{2}{3} & \frac{1}{3} & -\frac{2}{3} \\ -\frac{1}{5}\sqrt{5} & \frac{2}{5}\sqrt{5} & 0 \\ \frac{4}{15}\sqrt{5} & \frac{2}{15}\sqrt{5} & \frac{1}{3}\sqrt{5} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Since the identity was obtained this shows the above matrix is orthogonal and that therefore, the columns form an orthonormal set of vectors. The problem asks for you to find an orthonormal basis. However, you will show in Problem 23 that an orthonormal set of n vectors in $\mathbb{R}^n$ is always a basis. Therefore, since there are three of these vectors, they must constitute a basis.

Example 13.1.16 *Find an orthonormal set of three eigenvectors for the matrix*

$$\begin{pmatrix} \frac{13}{9} & \frac{2}{15}\sqrt{5} & \frac{8}{45}\sqrt{5} \\ \frac{2}{15}\sqrt{5} & \frac{6}{5} & \frac{4}{15} \\ \frac{8}{45}\sqrt{5} & \frac{4}{15} & \frac{61}{45} \end{pmatrix}$$

given the eigenvalues are 2, *and* 1.

The eigenvectors which go with $\lambda = 2$ are obtained from row reducing the matrix

$$\begin{pmatrix} \frac{13}{9} - 2 & \frac{2}{15}\sqrt{5} & \frac{8}{45}\sqrt{5} & | & 0 \\ \frac{2}{15}\sqrt{5} & \frac{6}{5} - 2 & \frac{4}{15} & | & 0 \\ \frac{8}{45}\sqrt{5} & \frac{4}{15} & \frac{61}{45} - 2 & | & 0 \end{pmatrix}$$

and its row reduced echelon form is

$$\begin{pmatrix} 1 & 0 & -\frac{1}{2}\sqrt{5} & | & 0 \\ 0 & 1 & -\frac{3}{4} & | & 0 \\ 0 & 0 & 0 & | & 0 \end{pmatrix}$$

which shows the eigenvectors for $\lambda = 2$ are $z \begin{pmatrix} \frac{1}{2}\sqrt{5} & \frac{3}{4} & 1 \end{pmatrix}^T$ and a choice for z which will produce a unit vector is $z = \frac{4}{15}\sqrt{5}$. Therefore, the vector we want is

$$\begin{pmatrix} \frac{2}{3} & \frac{1}{5}\sqrt{5} & \frac{4}{15}\sqrt{5} \end{pmatrix}^T.$$

Next consider the eigenvectors for $\lambda = 1$. The matrix which must be row reduced is

$$\begin{pmatrix} \frac{13}{9} - 1 & \frac{2}{15}\sqrt{5} & \frac{8}{45}\sqrt{5} & | & 0 \\ \frac{2}{15}\sqrt{5} & \frac{6}{5} - 1 & \frac{4}{15} & | & 0 \\ \frac{8}{45}\sqrt{5} & \frac{4}{15} & \frac{61}{45} - 1 & | & 0 \end{pmatrix}$$

and its row reduced echelon form is

$$\left(\begin{array}{cccc|c} 1 & \frac{3}{10}\sqrt{5} & \frac{2}{5}\sqrt{5} & | & 0 \\ 0 & 0 & 0 & | & 0 \\ 0 & 0 & 0 & | & 0 \end{array}\right).$$

Therefore, the eigenvectors are of the form$\left(\begin{array}{ccc} -\frac{3}{10}\sqrt{5}y-\frac{2}{5}\sqrt{5}z & y & z \end{array}\right)^T$, y,z arbitrary. This is a two dimensional eigenspace.

Before going further, we want to point out that no matter how we choose y and z the resulting vector will be orthogonal to the eigenvector for $\lambda = 2$. This is a special case of a general result which states that eigenvectors for distinct eigenvalues of a symmetric matrix are orthogonal. This is explained in Problem 15. For this case you need to show the following dot product equals zero.

$$\left(\begin{array}{c} \frac{2}{3} \\ \frac{1}{5}\sqrt{5} \\ \frac{4}{15}\sqrt{5} \end{array}\right) \cdot \left(\begin{array}{c} -\frac{3}{10}\sqrt{5}y-\frac{2}{5}\sqrt{5}z \\ y \\ z \end{array}\right) \tag{13.2}$$

This is left for you to do.

Continuing with the task of finding an orthonormal basis, Let $y = 0$ first. This results in eigenvectors of the form $\left(\begin{array}{ccc} -\frac{2}{5}\sqrt{5}z & 0 & z \end{array}\right)^T$ and letting $z = \frac{1}{3}\sqrt{5}$ you obtain a unit vector. Thus the second vector will be

$$\left(\begin{array}{c} -\frac{2}{5}\sqrt{5}\left(\frac{1}{3}\sqrt{5}\right) \\ 0 \\ \frac{1}{3}\sqrt{5} \end{array}\right) = \left(\begin{array}{c} -\frac{2}{3} \\ 0 \\ \frac{1}{3}\sqrt{5} \end{array}\right).$$

It remains to find the third vector in the orthonormal basis. This merely involves choosing y and z in 13.2 in such a way that the resulting vector has dot product with the two given vectors equal to zero. Thus you need

$$\left(\begin{array}{c} -\frac{3}{10}\sqrt{5}y-\frac{2}{5}\sqrt{5}z \\ y \\ z \end{array}\right) \cdot \left(\begin{array}{c} -\frac{2}{3} \\ 0 \\ \frac{1}{3}\sqrt{5} \end{array}\right) = \frac{1}{5}\sqrt{5}y+\frac{3}{5}\sqrt{5}z = 0.$$

The dot product with the eigenvector for $\lambda = 2$ is automatically equal to zero and so all that you need is the above equation. This is satisfied when $z = -\frac{1}{3}y$. Therefore, the vector we want is of the form

$$\left(\begin{array}{c} -\frac{3}{10}\sqrt{5}y-\frac{2}{5}\sqrt{5}\left(-\frac{1}{3}y\right) \\ y \\ \left(-\frac{1}{3}y\right) \end{array}\right) = \left(\begin{array}{c} -\frac{1}{6}\sqrt{5}y \\ y \\ -\frac{1}{3}y \end{array}\right)$$

and it only remains to choose y in such a way that this vector has unit length. This occurs when $y = \frac{2}{5}\sqrt{5}$. Therefore, the vector we want is

$$\frac{2}{5}\sqrt{5}\begin{pmatrix} -\frac{1}{6}\sqrt{5} \\ 1 \\ -\frac{1}{3} \end{pmatrix} = \begin{pmatrix} -\frac{1}{3} \\ \frac{2}{5}\sqrt{5} \\ -\frac{2}{15}\sqrt{5} \end{pmatrix}.$$

The three eigenvectors which constitute an orthonormal basis are

$$\begin{pmatrix} -\frac{1}{3} \\ \frac{2}{5}\sqrt{5} \\ -\frac{2}{15}\sqrt{5} \end{pmatrix}, \begin{pmatrix} -\frac{2}{3} \\ 0 \\ \frac{1}{3}\sqrt{5} \end{pmatrix}, \text{ and } \begin{pmatrix} \frac{2}{3} \\ \frac{1}{5}\sqrt{5} \\ \frac{4}{15}\sqrt{5} \end{pmatrix}.$$

To check the work and see if this is really an orthonormal set of vectors, make them the columns of a matrix and see if the resulting matrix is orthogonal. The matrix is

$$\begin{pmatrix} -\frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\ \frac{2}{5}\sqrt{5} & 0 & \frac{1}{5}\sqrt{5} \\ -\frac{2}{15}\sqrt{5} & \frac{1}{3}\sqrt{5} & \frac{4}{15}\sqrt{5} \end{pmatrix}.$$

This matrix times its transpose equals

$$\begin{pmatrix} -\frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \\ \frac{2}{5}\sqrt{5} & 0 & \frac{1}{5}\sqrt{5} \\ -\frac{2}{15}\sqrt{5} & \frac{1}{3}\sqrt{5} & \frac{4}{15}\sqrt{5} \end{pmatrix}\begin{pmatrix} -\frac{1}{3} & \frac{2}{5}\sqrt{5} & -\frac{2}{15}\sqrt{5} \\ -\frac{2}{3} & 0 & \frac{1}{3}\sqrt{5} \\ \frac{2}{3} & \frac{1}{5}\sqrt{5} & \frac{4}{15}\sqrt{5} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and so this is indeed an orthonormal basis.

Because of the repeated eigenvalue, there would have been many other orthonormal bases which could have been obtained. It was pretty arbitrary for to take $y = 0$ in the above argument. We could just as easily have taken $z = 0$ or even $y = z = 1$. Any such change would have resulted in a different orthonormal basis. Geometrically, what is happening is the eigenspace for $\lambda = 1$ was two dimensional. It can be visualized as a plane in three dimensional space which passes through the origin. There are infinitely many different pairs of perpendicular unit vectors in this plane.

13.1.3 Diagonalizing A Symmetric Matrix

Recall the following definition:

Definition 13.1.17 *An $n \times n$ matrix $A = (a_{ij})$ is called a diagonal matrix if $a_{ij} = 0$ whenever $i \neq j$. For example, a diagonal matrix is of the form indicated below where $*$ denotes* *a number.*

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

Definition 13.1.18 *An $n \times n$ matrix A is said to be* ***non defective*** *or* ***diagonalizable*** *if there exists an invertible matrix S such that $S^{-1}AS = D$ where D is a diagonal matrix as described above.*

Some matrices are non defective and some are not. As indicated in Theorem 13.1.13 if A is a real symmetric matrix, there exists an orthogonal matrix U such that $U^T AU = D$ a diagonal matrix. Therefore, every symmetric matrix is non defective because if U is an orthogonal matrix, its inverse is U^T. In the following example, this orthogonal matrix will be found.

Example 13.1.19 *Let* $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{3}{2} \end{pmatrix}$. *Find an orthogonal matrix U such that $U^T AU$ is a diagonal matrix.*

In this case, a tedious computation shows the eigenvalues are 2 and 1. First we will find an eigenvector for the eigenvalue 2. This involves row reducing the following augmented matrix.

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

The row reduced echelon form is

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

and so an eigenvector is $\begin{pmatrix} 0 & 1 & 1 \end{pmatrix}^T$. However, it is desired that the eigenvectors obtained all be unit vectors and so dividing this vector by its length gives

$$\begin{pmatrix} 0 & \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix}^T.$$

Next consider the case of the eigenvalue, 1. The matrix which needs to be row reduced in this case is

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

The row reduced echelon form is

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

Therefore, the eigenvectors are of the form $\begin{pmatrix} s & -t & t \end{pmatrix}^T$. Two of these which are orthonormal are

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} 0 \\ -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}.$$

An orthogonal matrix which works in the process is then obtained by letting these vectors be the columns.

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

It remains to verify this works. U^TAU is of the form

$$\begin{pmatrix} 0 & -\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \\ 1 & 0 & 0 \\ 0 & \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{3}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{3}{2} \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

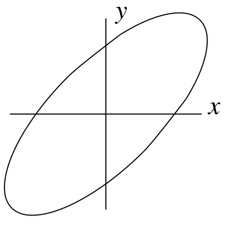
the desired diagonal matrix.

One of the applications for this technique has to do with rotation of axes so that with respect to the new axes, the graph the level curve of a quadratic form is oriented parallel to the coordinate axes. This makes it much easier to understand. This is discussed more in the exercises. However, here is a simple example.

Example 13.1.20 *Consider the following level curve.*

$$5x^2 - 6xy + 5y^2 = 8$$

Its graph is given in the following picture.



You can write this in terms of a symmetric matrix as follows.

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 8$$

Change the variables as follows.

$$\left(\begin{array}{cc}\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2}\\ \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2}\end{array}\right)^{T}\left(\begin{array}{cc}2 & 0\\ 0 & 8\end{array}\right)\left(\begin{array}{cc}\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2}\\ \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2}\end{array}\right)=\left(\begin{array}{cc}5 & -3\\ -3 & 5\end{array}\right)$$

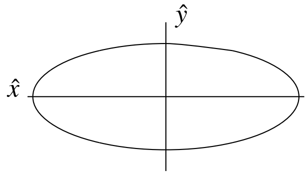
and so

$$\left(\begin{array}{cc}x & y\end{array}\right)\left(\begin{array}{cc}\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2}\\ \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2}\end{array}\right)^{T}\left(\begin{array}{cc}2 & 0\\ 0 & 8\end{array}\right)\left(\begin{array}{cc}\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2}\\ \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2}\end{array}\right)\left(\begin{array}{c}x\\ y\end{array}\right)=8$$

Let

$$\left(\begin{array}{cc}\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2}\\ \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2}\end{array}\right)\left(\begin{array}{c}x\\ y\end{array}\right)=\left(\begin{array}{c}\hat{x}\\ \hat{y}\end{array}\right)$$

Then in terms of these new variables, you get $2\hat{x}^2+8\hat{y}^2=8$. *This is an ellipse which is parallel to the coordinate axes. Its graph is of the form*



Thus this change of variables chooses new axes such that with respect to these new axes, the ellipse is oriented parallel to the coordinate axes. These new axes are called the principal axes.

In general a quadratic form is an expression of the form $\mathbf{x}^T A\mathbf{x}$ where A is a symmetric matrix. When you write something like $\mathbf{x}^T A\mathbf{x}=c$ you are considering a level surface or level curve of some sort. By diagonalizing the matrix as shown above, you can choose new variables such that in the new variables, there are no “mixed” terms like xy or yz. Geometrically this has the effect of choosing new coordinate axes such that with respect to these new axes, the various axes of symmetry of the level surfaces or curves are parallel to the coordinate axes. Therefore, this is a desirable simplification. Other quadratic forms in two variables lead to parabolas or hyperbolas. In three dimensions there are also names associated with these quadratic surfaces usually involving the semi word “oid”. They are typically discussed in calculus courses where they are invariably oriented parallel to the coordinate axes. However, the process of diagonalization just explained will allow one to start with one which is not oriented this way and reduce it to one which is.

13.2 Fundamental Theory And Generalizations

13.2.1 Block Multiplication Of Matrices

Consider the following problem

$$\left(\begin{array}{cc}A & B\\ C & D\end{array}\right)\left(\begin{array}{cc}E & F\\ G & H\end{array}\right)$$

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

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{r1} & \cdots & A_{rm} \end{pmatrix} \tag{13.3}$$

where A_{ij} is a $s_i \times p_j$ matrix where s_i is constant for $j = 1, \cdots, m$ for each $i = 1, \cdots, r$. Such a matrix is called a **block matrix,** also a **partitioned matrix**. How do you get the block A_{ij}? Here is how for A an $m \times n$ matrix:

$$\overbrace{\begin{pmatrix} \mathbf{0} & I_{s_i \times s_i} & \mathbf{0} \end{pmatrix}}^{s_i \times m} A \overbrace{\begin{pmatrix} \mathbf{0} \\ I_{p_j \times p_j} \\ \mathbf{0} \end{pmatrix}}^{n \times p_j}. \tag{13.4}$$

In the block column matrix on the right, you need to have $c_j - 1$ rows of zeros above the small $p_j \times p_j$ identity matrix where the columns of A involved in A_{ij} are $c_j, \cdots, c_j + p_j - 1$ and in the block row matrix on the left, you need to have $r_i - 1$ columns of zeros to the left of the $s_i \times s_i$ identity matrix where the rows of A involved in A_{ij} are $r_i, \cdots, r_i + s_i$. An important observation to make is that the matrix on the right specifies columns to use in the block and the one on the left specifies the rows used. Thus the block A_{ij} in this case is a matrix of size $s_i \times p_j$. There is no overlap between the blocks of A. Thus the identity $n \times n$ identity matrix corresponding to multiplication on the right of A is of the form

$$\begin{pmatrix} I_{p_1 \times p_1} & & 0 \\ & \ddots & \\ 0 & & I_{p_m \times p_m} \end{pmatrix}$$

these little identity matrices don't overlap. A similar conclusion follows from consideration of the matrices $I_{s_i \times s_i}$.

Next consider the question of multiplication of two block matrices. Let B be a block matrix of the form

$$\begin{pmatrix} B_{11} & \cdots & B_{1p} \\ \vdots & \ddots & \vdots \\ B_{r1} & \cdots & B_{rp} \end{pmatrix} \tag{13.5}$$

and A is a block matrix of the form

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{p1} & \cdots & A_{pm} \end{pmatrix} \tag{13.6}$$

and that for all i,j, it makes sense to multiply $B_{is}A_{sj}$ for all $s\in\{1,\cdots,p\}$. (That is the two matrices, B_{is} and A_{sj} are conformable.) and that for fixed ij, it follows $B_{is}A_{sj}$ is the same size for each s so that it makes sense to write $\sum_s B_{is}A_{sj}$.

The following theorem says essentially that when you take the product of two matrices, you can do it two ways. One way is to simply multiply them forming BA. The other way is to partition both matrices, formally multiply the blocks to get another block matrix and this one will be BA partitioned. Before presenting this theorem, here is a simple lemma which is really a special case of the theorem.

Lemma 13.2.1 *Consider the following product.*

$$\begin{pmatrix}\mathbf{0}\\ I\\ \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{0} & I & \mathbf{0}\end{pmatrix}$$

where the first is $n\times r$ and the second is $r\times n$. The small identity matrix I is an $r\times r$ matrix and there are l zero rows above I and l zero columns to the left of I in the right matrix. Then the product of these matrices is a block matrix of the form

$$\begin{pmatrix}\mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & I & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\end{pmatrix}$$

Proof: From the definition of the way you multiply matrices, the product is

$$\left(\begin{pmatrix}\mathbf{0}\\ I\\ \mathbf{0}\end{pmatrix}\mathbf{0}\cdots\begin{pmatrix}\mathbf{0}\\ I\\ \mathbf{0}\end{pmatrix}\mathbf{0}\begin{pmatrix}\mathbf{0}\\ I\\ \mathbf{0}\end{pmatrix}\mathbf{e}_1\cdots\begin{pmatrix}\mathbf{0}\\ I\\ \mathbf{0}\end{pmatrix}\mathbf{e}_r\begin{pmatrix}\mathbf{0}\\ I\\ \mathbf{0}\end{pmatrix}\mathbf{0}\cdots\begin{pmatrix}\mathbf{0}\\ I\\ \mathbf{0}\end{pmatrix}\mathbf{0}\right)$$

which yields the claimed result. In the formula $\mathbf{e}_j$ refers to the column vector of length r which has a 1 in the j^{th} position. ■

Theorem 13.2.2 *Let B be a $q\times p$ block matrix as in 13.5 and let A be a $p\times n$ block matrix as in 13.6 such that B_{is} is conformable with A_{sj} and each product, $B_{is}A_{sj}$ for $s=1,\cdots,p$ is of the same size so they can be added. Then BA can be obtained as a block matrix such that the ij^{th} block is of the form*

$$\sum_s B_{is}A_{sj}. \tag{13.7}$$

Proof: From 13.4

$$B_{is}A_{sj}=\begin{pmatrix}\mathbf{0} & I_{r_i\times r_i} & \mathbf{0}\end{pmatrix}B\begin{pmatrix}\mathbf{0}\\ I_{p_s\times p_s}\\ \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{0} & I_{p_s\times p_s} & \mathbf{0}\end{pmatrix}A\begin{pmatrix}\mathbf{0}\\ I_{q_j\times q_j}\\ \mathbf{0}\end{pmatrix}$$

where here it is assumed B_{is} is $r_i\times p_s$ and A_{sj} is $p_s\times q_j$. The product involves the s^{th} block in the i^{th} row of blocks for B and the s^{th} block in the j^{th} column of A. Thus there are the same number of rows above the $I_{p_s\times p_s}$ as there are columns to the left of $I_{p_s\times p_s}$ in those two inside matrices. Then from Lemma 13.2.1

$$\begin{pmatrix} \mathbf{0} \\ I_{p_s\times p_s} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & I_{p_s\times p_s} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{p_s\times p_s} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Since the blocks of small identity matrices do not overlap,

$$\sum_s \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{p_s\times p_s} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} I_{p_1\times p_1} & & 0 \\ & \ddots & \\ 0 & & I_{p_p\times p_p} \end{pmatrix} = I$$

and so $\sum_s B_{is}A_{sj} =$

$$\sum_s \begin{pmatrix} \mathbf{0} & I_{r_i\times r_i} & \mathbf{0} \end{pmatrix} B \begin{pmatrix} \mathbf{0} \\ I_{p_s\times p_s} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & I_{p_s\times p_s} & \mathbf{0} \end{pmatrix} A \begin{pmatrix} \mathbf{0} \\ I_{q_j\times q_j} \\ \mathbf{0} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0} & I_{r_i\times r_i} & \mathbf{0} \end{pmatrix} B \sum_s \begin{pmatrix} \mathbf{0} \\ I_{p_s\times p_s} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & I_{p_s\times p_s} & \mathbf{0} \end{pmatrix} A \begin{pmatrix} \mathbf{0} \\ I_{q_j\times q_j} \\ \mathbf{0} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0} & I_{r_i\times r_i} & \mathbf{0} \end{pmatrix} BIA \begin{pmatrix} \mathbf{0} \\ I_{q_j\times q_j} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & I_{r_i\times r_i} & \mathbf{0} \end{pmatrix} BA \begin{pmatrix} \mathbf{0} \\ I_{q_j\times q_j} \\ \mathbf{0} \end{pmatrix}$$

which equals the ij^{th} block of BA. Hence the ij^{th} block of BA equals the formal multiplication according to matrix multiplication, $\sum_s B_{is}A_{sj}$. ■

Example 13.2.3 *Let an $n\times n$ matrix have the form*

$$A = \begin{pmatrix} a & \mathbf{b} \\ \mathbf{c} & P \end{pmatrix}$$

where P is $n-1\times n-1$. Multiply it by

$$B = \begin{pmatrix} p & \mathbf{q} \\ \mathbf{r} & Q \end{pmatrix}$$

where B is also an $n\times n$ matrix and Q is $n-1\times n-1$.

You use block multiplication

$$\begin{pmatrix} a & \mathbf{b} \\ \mathbf{c} & P \end{pmatrix} \begin{pmatrix} p & \mathbf{q} \\ \mathbf{r} & Q \end{pmatrix} = \begin{pmatrix} ap+\mathbf{br} & a\mathbf{q}+\mathbf{b}Q \\ p\mathbf{c}+P\mathbf{r} & \mathbf{cq}+PQ \end{pmatrix}$$

Note that this all makes sense. For example, $\mathbf{b} = 1\times n-1$ and $\mathbf{r} = n-1\times 1$ so $\mathbf{br}$ is a 1×1. Similar considerations apply to the other blocks.

Here is an interesting and significant application of block multiplication. In this theorem, $p_M(t)$ denotes the characteristic polynomial, $\det(tI-M)$. Thus the zeros of this polynomial are the eigenvalues of the matrix M.

Theorem 13.2.4 *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,

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} AB & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ B & BA \end{pmatrix}$$

Since the two matrices above are similar 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 $\det(tI - BA) = p_{BA}(t) = t^{n-m} \det(tI - AB) = t^{n-m} p_{AB}(t)$. ■

13.2.2 Orthonormal Bases, Gram Schmidt Process

Not all bases for $\mathbb{F}^n$ are created equal. Recall $\mathbb{F}$ equals either $\mathbb{C}$ or $\mathbb{R}$ and the dot product is given by

$$\mathbf{x} \cdot \mathbf{y} \equiv (\mathbf{x}, \mathbf{y}) \equiv \langle \mathbf{x}, \mathbf{y} \rangle = \sum_j x_j \overline{y_j}.$$

The best bases are orthonormal. Much of what follows will be for $\mathbb{F}^n$ in the interest of generality.

Definition 13.2.5 *Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is a set of vectors in $\mathbb{F}^n$. It is an orthonormal set if*

$$\mathbf{v}_i \cdot \mathbf{v}_j = \delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

Every orthonormal set of vectors is automatically linearly independent.

Proposition 13.2.6 *Suppose $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ is an orthonormal set of vectors. Then it is linearly independent.*

Proof: Suppose $\sum_{i=1}^{k} c_i \mathbf{v}_i = \mathbf{0}$. Then taking dot products with $\mathbf{v}_j$,

$$0 = \mathbf{0} \cdot \mathbf{v}_j = \sum_i c_i \mathbf{v}_i \cdot \mathbf{v}_j = \sum_i c_i \delta_{ij} = c_j.$$

Since j is arbitrary, this shows the set is linearly independent as claimed. ■

It turns out that if X is any subspace of $\mathbb{F}^m$, then there exists an orthonormal basis for X. This follows from the use of the next lemma applied to a basis for X.

Lemma 13.2.7 *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a linearly independent subset of $\mathbb{F}^p$, $p \geq n$. Then there exist orthonormal vectors $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ which have the property that for each $k \leq n$,*

$$\operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k).$$

Proof: Let $\mathbf{u}_1 \equiv \mathbf{x}_1 / |\mathbf{x}_1|$. Thus for $k = 1$, $\operatorname{span}(\mathbf{u}_1) = \operatorname{span}(\mathbf{x}_1)$ and $\{\mathbf{u}_1\}$ is an orthonormal set. Now suppose for some $k < n$, $\mathbf{u}_1, \cdots, \mathbf{u}_k$ have been chosen such that $(\mathbf{u}_j, \mathbf{u}_l) = \delta_{jl}$ and $\operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$. Then define

$$\mathbf{u}_{k+1} \equiv \frac{\mathbf{x}_{k+1} - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) \mathbf{u}_j}{\left| \mathbf{x}_{k+1} - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) \mathbf{u}_j \right|}, \tag{13.8}$$

where the denominator is not equal to zero because the $\mathbf{x}_j$ form a basis, and so

$$\mathbf{x}_{k+1} \notin \operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k)$$

Thus by induction,

$$\mathbf{u}_{k+1} \in \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{x}_{k+1}) = \operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{x}_{k+1}).$$

Also, $\mathbf{x}_{k+1} \in \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1})$ which is seen easily by solving 13.8 for $\mathbf{x}_{k+1}$ and it follows

$$\operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k, \mathbf{x}_{k+1}) = \operatorname{span}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1}).$$

If $l \leq k$,

$$(\mathbf{u}_{k+1} \cdot \mathbf{u}_l) = C\left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j)(\mathbf{u}_j \cdot \mathbf{u}_l) \right) =$$

$$C\left((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - \sum_{j=1}^{k} (\mathbf{x}_{k+1} \cdot \mathbf{u}_j) \delta_{lj} \right) = C((\mathbf{x}_{k+1} \cdot \mathbf{u}_l) - (\mathbf{x}_{k+1} \cdot \mathbf{u}_l)) = 0.$$

The vectors, $\{\mathbf{u}_j\}_{j=1}^{n}$, generated in this way are therefore orthonormal because each vector has unit length. ■

The process by which these vectors were generated is called the Gram Schmidt process. Note that from the construction, each $\mathbf{x}_k$ is in the span of $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$. In terms of matrices, this says

$$(\mathbf{x}_1 \cdots \mathbf{x}_n) = (\mathbf{u}_1 \cdots \mathbf{u}_n) R$$

where R is an upper triangular matrix. This is closely related to the QR factorization discussed earlier. It is called the thin QR factorization. If the Gram Schmidt process is used to enlarge $\{\mathbf{u}_1 \cdots \mathbf{u}_n\}$ to an orthonormal basis for $\mathbb{F}^m$, $\{\mathbf{u}_1 \cdots \mathbf{u}_n, \mathbf{u}_{n+1}, \cdots, \mathbf{u}_m\}$ then if Q is the matrix which has these vectors as columns and if R is also enlarged to R' by adding in rows of zeros, if necessary, to form an $m \times n$ matrix, then the above would be of the form

$$(\mathbf{x}_1 \cdots \mathbf{x}_n) = (\mathbf{u}_1 \cdots \mathbf{u}_m) R'$$

and you could read off the orthonormal basis for span$(\mathbf{x}_1 \cdots \mathbf{x}_n)$ by simply taking the first n columns of $Q = (\mathbf{u}_1 \cdots \mathbf{u}_m)$. This is convenient because computer algebra systems are set up to find QR factorizations.

Example 13.2.8 *Find an orthonormal basis for* span$\left(\begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}\right)$.

This is really easy to do using a computer algebra system.

$$\begin{pmatrix} 1 & 2 \\ 3 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{11}\sqrt{11} & \frac{19}{506}\sqrt{11}\sqrt{46} & \frac{3}{46}\sqrt{46} \\ \frac{3}{11}\sqrt{11} & -\frac{9}{506}\sqrt{11}\sqrt{46} & \frac{1}{46}\sqrt{46} \\ \frac{1}{11}\sqrt{11} & \frac{4}{253}\sqrt{11}\sqrt{46} & -\frac{3}{23}\sqrt{46} \end{pmatrix} \begin{pmatrix} \sqrt{11} & \frac{3}{11}\sqrt{11} \\ 0 & \frac{1}{11}\sqrt{11}\sqrt{46} \\ 0 & 0 \end{pmatrix}$$

and so the desired orthonormal basis is

$$\begin{pmatrix} \frac{1}{11}\sqrt{11} \\ \frac{3}{11}\sqrt{11} \\ \frac{1}{11}\sqrt{11} \end{pmatrix}, \begin{pmatrix} \frac{19}{506}\sqrt{11}\sqrt{46} \\ -\frac{9}{506}\sqrt{11}\sqrt{46} \\ \frac{4}{253}\sqrt{11}\sqrt{46} \end{pmatrix}$$

►

13.2.3 Schur's Theorem

Every matrix is related to an upper triangular matrix in a particularly significant way. This is Schur's theorem and it is the most important theorem in the spectral theory of matrices. The important result which makes this theorem possible is the Gram Schmidt procedure of Lemma 13.2.7.

Definition 13.2.9 *An $n \times n$ matrix U, is* ***unitary*** *if $UU^* = I = U^*U$ where U^* is defined to be the transpose of the conjugate of U. Thus $\overline{U_{ij}} = U^*_{ji}$. Note that every real orthogonal matrix is unitary. For A any matrix A^*, just defined as the conjugate of the transpose, is called the* ***adjoint.***

Note that if $U = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix}$ where the $\mathbf{v}_k$ are orthonormal vectors in $\mathbb{C}^n$, then U is unitary. This follows because the ij^{th} entry of U^*U is $\overline{\mathbf{v}_i^T}\mathbf{v}_j = \delta_{ij}$ since the $\mathbf{v}_i$ are assumed orthonormal.

Lemma 13.2.10 *The following holds.* $(AB)^* = B^*A^*$.

Proof: From the definition and remembering the properties of complex conjugation,

$$\left((AB)^*\right)_{ji} = \overline{(AB)_{ij}} = \overline{\sum_k A_{ik}B_{kj}} = \sum_k \overline{A_{ik}}\,\overline{B_{kj}} = \sum_k B^*_{jk}A^*_{ki} = (B^*A^*)_{ji} \blacksquare$$

Theorem 13.2.11 *Let A be an $n\times n$ matrix. Then there exists a unitary matrix U such that*

$$U^*AU = T, \tag{13.9}$$

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. If A is a real matrix having all real eigenvalues, then U can be chosen to be an orthogonal real matrix.

Proof: The theorem is clearly true if A is a 1×1 matrix. Just let $U = 1$, the 1×1 matrix which has entry 1. Suppose it is true for $(n-1)\times(n-1)$ matrices, $n\geq 2$ and let A be an $n\times n$ matrix. Then let $\mathbf{v}_1$ be a unit eigenvector for A. Then there exists λ_1 such that

$$A\mathbf{v}_1 = \lambda_1\mathbf{v}_1,\ |\mathbf{v}_1| = 1.$$

Extend $\{\mathbf{v}_1\}$ to a basis and then use the Gram - Schmidt process to obtain

$$\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$$

an orthonormal basis of $\mathbb{C}^n$. Let U_0 be a matrix whose i^{th} column is $\mathbf{v}_i$ so that U_0 is unitary. Consider $U_0^*AU_0$

$$U_0^*AU_0 = \begin{pmatrix} \mathbf{v}_1^* \\ \vdots \\ \mathbf{v}_n^* \end{pmatrix} \begin{pmatrix} A\mathbf{v}_1 & \cdots & A\mathbf{v}_n \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1^* \\ \vdots \\ \mathbf{v}_n^* \end{pmatrix} \begin{pmatrix} \lambda_1\mathbf{v}_1 & \cdots & A\mathbf{v}_n \end{pmatrix}$$

Thus $U_0^*AU_0$ is of the form

$$\begin{pmatrix} \lambda_1 & \mathbf{a} \\ \mathbf{0} & A_1 \end{pmatrix}$$

where A_1 is an $n-1\times n-1$ matrix. Now by induction, there exists an $(n-1)\times(n-1)$ unitary matrix $\widetilde{U}_1$ such that $\widetilde{U}_1^*A_1\widetilde{U}_1 = T_{n-1}$, an upper triangular matrix. Consider

$$U_1 \equiv \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix}.$$

Then

$$U_1^*U_1 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1^* \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & I_{n-1} \end{pmatrix}$$

Also

$$\begin{aligned} U_1^*U_0^*AU_0U_1 &= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1^* \end{pmatrix} \begin{pmatrix} \lambda_1 & * \\ \mathbf{0} & A_1 \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widetilde{U}_1 \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 & * \\ \mathbf{0} & T_{n-1} \end{pmatrix} \equiv T \end{aligned}$$

where T is upper triangular. Then let $U = U_0U_1$. It is clear that this is unitary because both matrices preserve distance. Therefore, so does the product and hence U. Alternatively,

$$I = U_0U_1U_1^*U_0^* = (U_0U_1)(U_0U_1)^*$$

and so, it follows that A is similar to T and that $U_0 U_1$ is unitary. Hence A and T have the same characteristic polynomials, and since the eigenvalues of T (A) are the diagonal entries listed with multiplicity, this proves the main conclusion of the theorem. In case A is real with all real eigenvalues, the above argument can be repeated word for word using only the real dot product to show that U can be taken to be real and orthogonal. ■

As a simple consequence of the above theorem, here is an interesting lemma.

Lemma 13.2.12 *Let A be of the form*

$$A = \begin{pmatrix} P_1 & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_s \end{pmatrix}$$

where P_k is an $m_k \times m_k$ matrix. Then

$$\det(A) = \prod_k \det(P_k).$$

Proof: Let U_k be an $m_k \times m_k$ unitary matrix such that

$$U_k^* P_k U_k = T_k$$

where T_k is upper triangular. Then letting U denote the block diagonal matrix, having the U_i as the blocks on the diagonal,

$$U = \begin{pmatrix} U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_s \end{pmatrix}, U^* = \begin{pmatrix} U_1^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_s^* \end{pmatrix}$$

and

$$\begin{pmatrix} U_1^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_s^* \end{pmatrix} \begin{pmatrix} P_1 & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_s \end{pmatrix} \begin{pmatrix} U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_s \end{pmatrix} = \begin{pmatrix} T_1 & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T_s \end{pmatrix}$$

and so

$$\det(A) = \prod_k \det(T_k) = \prod_k \det(P_k). \blacksquare$$

Definition 13.2.13 *An $n \times n$ matrix A is called* ***Hermitian*** *if $A = A^*$. Thus a real symmetric ($A = A^T$) matrix is Hermitian.*

Recall that from Theorem 13.2.14, the eigenvalues of a real symmetric matrix are all real.

Theorem 13.2.14 *If A is an $n \times n$ Hermitian matrix, there exists a unitary matrix U such that*

$$U^* A U = D \tag{13.10}$$

where D is a real diagonal matrix. That is, D has nonzero entries only on the main diagonal and these are real. Furthermore, the columns of U are an orthonormal basis of eigenvectors for $\mathbb{C}^n$. If A is real and symmetric, then U can be assumed to be a real orthogonal matrix and the columns of U form an orthonormal basis for $\mathbb{R}^n$.

Proof: From Schur's theorem above, there exists U unitary (real and orthogonal if A is real) such that

$$U^*AU = T$$

where T is an upper triangular matrix. Then from Lemma 13.2.10

$$T^* = (U^*AU)^* = U^*A^*U = U^*AU = T.$$

Thus $T = T^*$ and T is upper triangular. This can only happen if T is really a diagonal matrix having real entries on the main diagonal. (If $i \neq j$, one of T_{ij} or T_{ji} equals zero. But $T_{ij} = \overline{T_{ji}}$ and so they are both zero. Also $T_{ii} = \overline{T_{ii}}$.)

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. In case A is real and symmetric, simply ignore all complex conjugations in the above argument. ■

13.3 Least Square Approximation

A very important technique is that of the least square approximation.

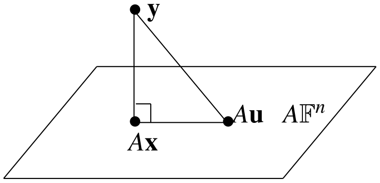
Lemma 13.3.1 *Let A be an $m \times n$ matrix and let $A(\mathbb{F}^n)$ denote the set of vectors in $\mathbb{F}^m$ which are of the form $A\mathbf{x}$ for some $\mathbf{x} \in \mathbb{F}^n$. Then $A(\mathbb{F}^n)$ is a subspace of $\mathbb{F}^m$.*

Proof: Let $A\mathbf{x}$ and $A\mathbf{y}$ be two points of $A(\mathbb{F}^n)$. It suffices to verify that if a, b are scalars, then $aA\mathbf{x} + bA\mathbf{y}$ is also in $A(\mathbb{F}^n)$. But $aA\mathbf{x} + bA\mathbf{y} = A(a\mathbf{x} + b\mathbf{y}) \in A(\mathbb{F}^n)$ because A is linear. ■

Lemma 13.3.2 *Suppose $b \geq 0$ and $c \in \mathbb{R}$ such that $a + bt^2 + ct \geq a$ for all $t \in \mathbb{R}$, then $c = 0$.*

Proof: You need $bt^2+ct\geq 0$ for all t. The slope of $t\mapsto bt^2+ct$ is c when $t=0$. Thus the inequality is violated unless $c=0$. ■

The following theorem gives the equivalence of an orthogonality condition with a minimization condition. The following picture illustrates the geometric meaning of this theorem



Theorem 13.3.3 *Let $\mathbf{y}\in\mathbb{F}^m$ and let A be an $m\times n$ matrix. Then there exists $\mathbf{x}\in\mathbb{F}^n$ minimizing the function $\mathbf{x}\mapsto\left|\mathbf{y}-A\mathbf{x}\right|^2$. Furthermore, $\mathbf{x}$ minimizes this function if and only if*

$$((\mathbf{y}-A\mathbf{x}),A\mathbf{u})=0$$

for all $\mathbf{u}\in\mathbb{F}^n$.

Proof: First consider the characterization of the minimizer. Let $\mathbf{u}\in\mathbb{F}^n$. Let $|\theta|=1$,

$$\bar{\theta}\left(\mathbf{y}-A\mathbf{x},A\mathbf{u}\right)=\left|\left(\mathbf{y}-A\mathbf{x},A\mathbf{u}\right)\right|$$

Now consider the function of $t\in\mathbb{R}$

$$p(t)\equiv\left|\mathbf{y}-\left(A\mathbf{x}+t\theta A\mathbf{u}\right)\right|^2=\left(\left(\mathbf{y}-A\mathbf{x}\right)-t\theta A\mathbf{u},\left(\mathbf{y}-A\mathbf{x}\right)-t\theta A\mathbf{u}\right)$$

$$=\left|\mathbf{y}-A\mathbf{x}\right|^2+t^2\left|A\mathbf{u}\right|^2-2t\operatorname{Re}\left(\mathbf{y}-A\mathbf{x},\theta A\mathbf{u}\right)\geq\left|\mathbf{y}-A\mathbf{x}\right|^2$$

$$=\left|\mathbf{y}-A\mathbf{x}\right|^2+t^2\left|A\mathbf{u}\right|^2-2t\operatorname{Re}\bar{\theta}\left(\mathbf{y}-A\mathbf{x},A\mathbf{u}\right)$$

$$=\left|\mathbf{y}-A\mathbf{x}\right|^2+t^2\left|A\mathbf{u}\right|^2-2t\left|\left(\mathbf{y}-A\mathbf{x},A\mathbf{u}\right)\right|$$

Then if $\left|\mathbf{y}-A\mathbf{x}\right|$ is as small as possible, this will occur when $t=0$ and so $p'(0)=0$. But this says

$$\left|\left(\mathbf{y}-A\mathbf{x},A\mathbf{u}\right)\right|=0$$

You could also use Lemma 13.3.2 to see this is 0. Since $\mathbf{u}$ was arbitrary, this proves one direction.

Conversely, if this quantity equals 0,

$$\begin{aligned}\left|\mathbf{y}-\left(A\mathbf{x}+A\mathbf{u}\right)\right|^2&=\left|\mathbf{y}-A\mathbf{x}\right|^2+\left|A\mathbf{x}-A\mathbf{u}\right|^2+2\operatorname{Re}\left(\mathbf{y}-A\mathbf{x},A\mathbf{u}\right)\\&=\left|\mathbf{y}-A\mathbf{x}\right|^2+\left|A\mathbf{x}-A\mathbf{u}\right|^2\end{aligned}$$

and so the minimum occurs at any point $\mathbf{z}$ such that $A\mathbf{x}=A\mathbf{z}$.

Does there exist an $\mathbf{x}$ which minimizes this function? From what was just shown, it suffices to show that there exists $\mathbf{x}$ such that $((\mathbf{y}-A\mathbf{x}),A\mathbf{u})$ for all $\mathbf{u}$. By the Gramm Schmidt process there exists an orthonormal basis $\{A\mathbf{x}_k\}$ for $A(\mathbb{F}^n)$. Then for a given $\mathbf{y}$,

$$\left(\mathbf{y}-\sum_{k=1}^{r}\left(\mathbf{y},A\mathbf{x}_k\right)A\mathbf{x}_k,A\mathbf{x}_j\right)=\left(\mathbf{y},A\mathbf{x}_j\right)-\sum_{k=1}^{r}\left(\mathbf{y},A\mathbf{x}_k\right)\overbrace{\left(A\mathbf{x}_k,A\mathbf{x}_j\right)}^{\delta_{kj}}=0.$$

In particular,

$$\left(\mathbf{y}-A\left(\sum_{k=1}^{r}(\mathbf{y},A\mathbf{x}_k)\,\mathbf{x}_k\right),\mathbf{w}\right)=0$$

for all $\mathbf{w}\in A(\mathbb{F}^n)$ since $\{A\mathbf{x}_k\}$ is a basis. Therefore,

$$\mathbf{x}=\sum_{k=1}^{r}(\mathbf{y},A\mathbf{x}_k)\,\mathbf{x}_k$$

is a minimizer. So is any $\mathbf{z}$ such that $A\mathbf{z}=A\mathbf{x}$. ■

Recall the definition of the adjoint of a matrix.

Definition 13.3.4 *Let A be an $m\times n$ matrix. Then*

$$A^*\equiv\overline{(A^T)}.$$

*This means you take the transpose of A and then replace each entry by its conjugate. This matrix is called the **adjoint**. Thus in the case of real matrices having only real entries, the adjoint is just the transpose.*

Lemma 13.3.5 *Let A be an $m\times n$ matrix. Then*

$$A\mathbf{x}\cdot\mathbf{y}=\mathbf{x}\cdot A^*\mathbf{y}$$

Proof: This follows from the definition.

$$A\mathbf{x}\cdot\mathbf{y}=\sum_{i,j}A_{ij}x_j\overline{y_i}=\sum_{i,j}x_j\overline{A^*_{ji}\overline{y_i}}=\mathbf{x}\cdot A^*\mathbf{y}.\ ■$$

The next corollary gives the technique of least squares.

Corollary 13.3.6 *A value of* $\mathbf{x}$ *which solves the problem of Theorem 13.3.3 is obtained by solving the equation*

$$A^*A\mathbf{x}=A^*\mathbf{y}$$

and furthermore, there exists a solution to this system of equations.

Proof: For $\mathbf{x}$ the minimizer of Theorem 13.3.3, $(\mathbf{y}-A\mathbf{x})\cdot A\mathbf{w}=0$ for all $\mathbf{w}\in\mathbb{F}^n$ and from Lemma 13.3.5, this is the same as saying

$$A^*(\mathbf{y}-A\mathbf{x})\cdot\mathbf{w}=0$$

for all $\mathbf{w}\in\mathbb{F}^n$. This implies

$$A^*\mathbf{y}-A^*A\mathbf{x}=\mathbf{0}.$$

Therefore, there is a solution to the equation of this corollary, and it solves the minimization problem of Theorem 13.3.3. ■

Note that $\mathbf{x}$ might not be unique but $A\mathbf{x}$, the closest point of $A(\mathbb{F}^n)$ to $\mathbf{y}$ is unique. This was shown in the above argument. Sometimes people like to consider the $\mathbf{x}$ such that $A\mathbf{x}$ is as close as possible to $\mathbf{y}$ and also $|\mathbf{x}|$ is as small as possible. It turns out that there exists a unique such $\mathbf{x}$ and it is denoted as $A^+\mathbf{y}$. However, this is as far as I will go with this in this part of the book.

There is also a useful observation about orthonormal sets of vectors which is stated in the next lemma.

Lemma 13.3.7 *Suppose* $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_r\}$ *is an orthonormal set of vectors. Then if* $c_1, \cdots, c_r$ *are scalars,*

$$\left|\sum_{k=1}^{r} c_k \mathbf{x}_k\right|^2 = \sum_{k=1}^{r} |c_k|^2 .$$

Proof: This follows from the definition. From the properties of the dot product and using the fact that the given set of vectors is orthonormal,

$$\left|\sum_{k=1}^{r} c_k \mathbf{x}_k\right|^2 = \left(\sum_{k=1}^{r} c_k \mathbf{x}_k, \sum_{j=1}^{r} c_j \mathbf{x}_j\right) = \sum_{k,j} c_k \overline{c_j} (\mathbf{x}_k, \mathbf{x}_j) = \sum_{k=1}^{r} |c_k|^2 . \blacksquare$$

13.3.1 The Least Squares Regression Line

For the situation of the least squares regression line discussed here I will specialize to the case of $\mathbb{R}^n$ rather than $\mathbb{F}^n$ because it seems this case is by far the most interesting and the extra details are not justified by an increase in utility. Thus, everywhere you see A^* it suffices to place A^T.

An important application of Corollary 13.3.6 is the problem of finding the least squares regression line in statistics. Suppose you are given points in xy plane

$$\{(x_i, y_i)\}_{i=1}^{n}$$

and you would like to find constants m and b such that the line $y = mx + b$ goes through all these points. Of course this will be impossible in general. Therefore, try to find m, b to get as close as possible. The desired system is

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} \equiv A \begin{pmatrix} m \\ b \end{pmatrix}$$

which is of the form $\mathbf{y} = A\mathbf{x}$ and it is desired to choose m and b to make

$$\left| A \begin{pmatrix} m \\ b \end{pmatrix} - \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right|^2$$

as small as possible. According to Theorem 13.3.3 and Corollary 13.3.6, the best values for m and b occur as the solution to

$$A^T A \begin{pmatrix} m \\ b \end{pmatrix} = A^T \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}.$$

Thus, computing $A^T A$,

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix}$$

Solving this system of equations for m and b,

$$m=\frac{-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)+\left(\sum_{i=1}^{n}x_iy_i\right)n}{\left(\sum_{i=1}^{n}x_i^2\right)n-\left(\sum_{i=1}^{n}x_i\right)^2}$$

and

$$b=\frac{-\left(\sum_{i=1}^{n}x_i\right)\sum_{i=1}^{n}x_iy_i+\left(\sum_{i=1}^{n}y_i\right)\sum_{i=1}^{n}x_i^2}{\left(\sum_{i=1}^{n}x_i^2\right)n-\left(\sum_{i=1}^{n}x_i\right)^2}.$$

One could clearly do a least squares fit for curves of the form $y=ax^2+bx+c$ in the same way. In this case you want to solve as well as possible for a,b, and c the system

$$\begin{pmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{pmatrix}\begin{pmatrix} a \\ b \\ c \end{pmatrix}=\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

and one would use the same technique as above. Many other similar problems are important, including many in higher dimensions and they are all solved the same way.

13.3.2 The Fredholm Alternative

The next major result is called the Fredholm alternative. It comes from Theorem 13.3.3 and Lemma 13.3.5.

Theorem 13.3.8 *Let A be an $m\times n$ matrix. Then there exists $\mathbf{x}\in\mathbb{F}^n$ such that $A\mathbf{x}=\mathbf{y}$ if and only if whenever $A^*\mathbf{z}=\mathbf{0}$ it follows that $\mathbf{z}\cdot\mathbf{y}=0$.*

Proof: First suppose that for some $\mathbf{x}\in\mathbb{F}^n$, $A\mathbf{x}=\mathbf{y}$. Then letting $A^*\mathbf{z}=\mathbf{0}$ and using Lemma 13.3.5

$$\mathbf{y}\cdot\mathbf{z}=A\mathbf{x}\cdot\mathbf{z}=\mathbf{x}\cdot A^*\mathbf{z}=\mathbf{x}\cdot\mathbf{0}=0.$$

This proves half the theorem.

To do the other half, suppose that whenever, $A^*\mathbf{z}=\mathbf{0}$ it follows that $\mathbf{z}\cdot\mathbf{y}=0$. It is necessary to show there exists $\mathbf{x}\in\mathbb{F}^n$ such that $\mathbf{y}=A\mathbf{x}$. From Theorem 13.3.3 there exists $\mathbf{x}$ minimizing $|\mathbf{y}-A\mathbf{x}|^2$ which therefore satisfies

$$(\mathbf{y}-A\mathbf{x})\cdot A\mathbf{w}=0 \tag{13.11}$$

for all $\mathbf{w}\in\mathbb{F}^n$. Therefore, for all $\mathbf{w}\in\mathbb{F}^n$,

$$A^*(\mathbf{y}-A\mathbf{x})\cdot\mathbf{w}=0$$

which shows that $A^*(\mathbf{y}-A\mathbf{x})=\mathbf{0}$. (Why?) Therefore, by assumption,

$$(\mathbf{y}-A\mathbf{x})\cdot\mathbf{y}=0.$$

Now by 13.11 with $\mathbf{w}=\mathbf{x}$,

$$(\mathbf{y}-A\mathbf{x})\cdot(\mathbf{y}–A\mathbf{x})=(\mathbf{y}-A\mathbf{x})\cdot\mathbf{y}–(\mathbf{y}-A\mathbf{x})\cdot A\mathbf{x}=0$$

showing that $\mathbf{y}=A\mathbf{x}$. ■

The following corollary is also called the Fredholm alternative.

Corollary 13.3.9 *Let A be an $m \times n$ matrix. Then A is onto if and only if A^* is one to one.*

Proof: Suppose first A is onto. Then by Theorem 13.3.8, it follows that for all $\mathbf{y} \in \mathbb{F}^m$, $\mathbf{y} \cdot \mathbf{z} = 0$ whenever $A^*\mathbf{z} = \mathbf{0}$. Therefore, let $\mathbf{y} = \mathbf{z}$ where $A^*\mathbf{z} = \mathbf{0}$ and conclude that $\mathbf{z} \cdot \mathbf{z} = 0$ whenever $A^*\mathbf{z} = 0$. If $A^*\mathbf{x} = A^*\mathbf{y}$, then $A^*(\mathbf{x} - \mathbf{y}) = \mathbf{0}$ and so $\mathbf{x} - \mathbf{y} = \mathbf{0}$. Thus A^* is one to one.

Now let $\mathbf{y} \in \mathbb{F}^m$ be given. $\mathbf{y} \cdot \mathbf{z} = 0$ whenever $A^*\mathbf{z} = \mathbf{0}$ because, since A^* is assumed to be one to one, and $\mathbf{0}$ is a solution to this equation, it must be the only solution. Therefore, by Theorem 13.3.8 there exists $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{y}$ therefore, A is onto. ■

13.4 The Right Polar Factorization*

The right polar factorization involves writing a matrix as a product of two other matrices, one which preserves distances and the other which stretches and distorts. First here are some lemmas which review and add to many of the topics discussed so far about adjoints and orthonormal sets and such things. This is of fundamental significance in geometric measure theory and also in continuum mechanics. Not surprisingly the stress should depend on the part which stretches and distorts. See [8].

Lemma 13.4.1 *Let A be a Hermitian matrix such that all its eigenvalues are nonnegative. Then there exists a Hermitian matrix $A^{1/2}$ such that $A^{1/2}$ has all nonnegative eigenvalues and $\left(A^{1/2}\right)^2 = A$.*

Proof: Since A is Hermitian, there exists a diagonal matrix D having all real nonnegative entries and a unitary matrix U such that $A = U^*DU$. Then denote by $D^{1/2}$ the matrix which is obtained by replacing each diagonal entry of D with its square root. Thus $D^{1/2}D^{1/2} = D$. Then define

$$A^{1/2} \equiv U^*D^{1/2}U.$$

Then

$$\left(A^{1/2}\right)^2 = U^*D^{1/2}UU^*D^{1/2}U = U^*DU = A.$$

Since $D^{1/2}$ is real,

$$\left(U^*D^{1/2}U\right)^* = U^*\left(D^{1/2}\right)^*(U^*)^* = U^*D^{1/2}U$$

so $A^{1/2}$ is Hermitian. ■

Next it is helpful to recall the Gram Schmidt algorithm and observe a certain property stated in the next lemma.

Lemma 13.4.2 *Suppose $\{\mathbf{w}_1, \cdots, \mathbf{w}_r, \mathbf{v}_{r+1}, \cdots, \mathbf{v}_p\}$ is a linearly independent set of vectors such that $\{\mathbf{w}_1, \cdots, \mathbf{w}_r\}$ is an orthonormal set of vectors. Then when the Gram Schmidt process is applied to the vectors in the given order, it will not change any of the $\mathbf{w}_1, \cdots, \mathbf{w}_r$.*

Proof: Let $\{\mathbf{u}_1, \cdots, \mathbf{u}_p\}$ be the orthonormal set delivered by the Gram Schmidt process. Then $\mathbf{u}_1 = \mathbf{w}_1$ because by definition, $\mathbf{u}_1 \equiv \mathbf{w}_1 / |\mathbf{w}_1| = \mathbf{w}_1$. Now suppose $\mathbf{u}_j = \mathbf{w}_j$ for all $j \leq k \leq r$. Then if $k < r$, consider the definition of $\mathbf{u}_{k+1}$.

$$\mathbf{u}_{k+1} \equiv \frac{\mathbf{w}_{k+1} - \sum_{j=1}^{k+1} (\mathbf{w}_{k+1}, \mathbf{u}_j)\, \mathbf{u}_j}{\left|\mathbf{w}_{k+1} - \sum_{j=1}^{k+1} (\mathbf{w}_{k+1}, \mathbf{u}_j)\, \mathbf{u}_j\right|}$$

By induction, $\mathbf{u}_j = \mathbf{w}_j$ and so this reduces to $\mathbf{w}_{k+1}/\left|\mathbf{w}_{k+1}\right| = \mathbf{w}_{k+1}$. ■

This lemma immediately implies the following lemma.

Lemma 13.4.3 *Let V be a subspace of dimension p and let $\{\mathbf{w}_1,\cdots,\mathbf{w}_r\}$ be an orthonormal set of vectors in V. Then this orthonormal set of vectors may be extended to an orthonormal basis for V,*

$$\left\{\mathbf{w}_1,\cdots,\mathbf{w}_r,\mathbf{y}_{r+1},\cdots,\mathbf{y}_p\right\}$$

Proof: First extend the given linearly independent set $\{\mathbf{w}_1,\cdots,\mathbf{w}_r\}$ to a basis for V and then apply the Gram Schmidt theorem to the resulting basis. Since $\{\mathbf{w}_1,\cdots,\mathbf{w}_r\}$ is orthonormal it follows from Lemma 13.4.2 the result is of the desired form, an orthonormal basis extending $\{\mathbf{w}_1,\cdots,\mathbf{w}_r\}$. ■

Here is another lemma about preserving distance.

Lemma 13.4.4 *Suppose R is an $m\times n$ matrix with $m\geq n$ and R preserves distances. Then $R^*R = I$.*

Proof: Since R preserves distances, $|R\mathbf{x}| = |\mathbf{x}|$ for every $\mathbf{x}$. Therefore from the axioms of the dot product,

$$\begin{aligned} & |\mathbf{x}|^2+|\mathbf{y}|^2+(\mathbf{x},\mathbf{y})+(\mathbf{y},\mathbf{x}) = |\mathbf{x}+\mathbf{y}|^2 = (R(\mathbf{x}+\mathbf{y}),R(\mathbf{x}+\mathbf{y})) \\ = \; & (R\mathbf{x},R\mathbf{x})+(R\mathbf{y},R\mathbf{y})+(R\mathbf{x},R\mathbf{y})+(R\mathbf{y},R\mathbf{x}) \\ = \; & |\mathbf{x}|^2+|\mathbf{y}|^2+(R^*R\mathbf{x},\mathbf{y})+(\mathbf{y},R^*R\mathbf{x}) \end{aligned}$$

and so for all $\mathbf{x},\mathbf{y}$,

$$(R^*R\mathbf{x}-\mathbf{x},\mathbf{y})+(\mathbf{y},R^*R\mathbf{x}-\mathbf{x}) = 0$$

Hence for all $\mathbf{x},\mathbf{y}$,

$$\operatorname{Re}(R^*R\mathbf{x}-\mathbf{x},\mathbf{y}) = 0$$

Now for a $\mathbf{x},\mathbf{y}$ given, choose $\alpha\in\mathbb{C}$ such that

$$\alpha(R^*R\mathbf{x}-\mathbf{x},\mathbf{y}) = |(R^*R\mathbf{x}-\mathbf{x},\mathbf{y})|$$

Then

$$0 = \operatorname{Re}(R^*R\mathbf{x}-\mathbf{x},\overline{\alpha}\mathbf{y}) = \operatorname{Re}\alpha(R^*R\mathbf{x}-\mathbf{x},\mathbf{y}) = |(R^*R\mathbf{x}-\mathbf{x},\mathbf{y})|$$

Thus $|(R^*R\mathbf{x}-\mathbf{x},\mathbf{y})| = 0$ for all $\mathbf{x},\mathbf{y}$ because the given $\mathbf{x},\mathbf{y}$ were arbitrary. Let $\mathbf{y} = R^*R\mathbf{x}-\mathbf{x}$ to conclude that for all $\mathbf{x}$,

$$R^*R\mathbf{x}-\mathbf{x} = \mathbf{0}$$

which says $R^*R = I$ since $\mathbf{x}$ is arbitrary. ■

With this preparation, here is the big theorem about the right polar factorization.

Theorem 13.4.5 *Let F be an $m\times n$ matrix where $m\geq n$. Then there exists a Hermitian $n\times n$ matrix U which has all nonnegative eigenvalues and an $m\times n$ matrix R which preserves distances and satisfies $R^*R = I$ such that*

$$F = RU.$$

Proof: Consider F^*F. This is a Hermitian matrix because

$$(F^*F)^* = F^*(F^*)^* = F^*F$$

Also the eigenvalues of the $n \times n$ matrix F^*F are all nonnegative. This is because if $\mathbf{x}$ is an eigenvalue,

$$\lambda(\mathbf{x},\mathbf{x}) = (F^*F\mathbf{x},\mathbf{x}) = (F\mathbf{x},F\mathbf{x}) \geq 0.$$

Therefore, by Lemma 13.4.1, there exists an $n \times n$ Hermitian matrix U having all nonnegative eigenvalues such that

$$U^2 = F^*F.$$

Consider the subspace $U(\mathbb{F}^n)$. Let $\{U\mathbf{x}_1,\cdots,U\mathbf{x}_r\}$ be an orthonormal basis for

$$U(\mathbb{F}^n) \subseteq \mathbb{F}^n.$$

Note that $U(\mathbb{F}^n)$ might not be all of $\mathbb{F}^n$. Using Lemma 13.4.3, extend to an orthonormal basis for all of $\mathbb{F}^n$,

$$\{U\mathbf{x}_1,\cdots,U\mathbf{x}_r,\mathbf{y}_{r+1},\cdots,\mathbf{y}_n\}.$$

Next observe that $\{F\mathbf{x}_1,\cdots,F\mathbf{x}_r\}$ is also an orthonormal set of vectors in $\mathbb{F}^m$. This is because

$$(F\mathbf{x}_k,F\mathbf{x}_j) = (F^*F\mathbf{x}_k,\mathbf{x}_j) = \left(U^2\mathbf{x}_k,\mathbf{x}_j\right) = (U\mathbf{x}_k,U^*\mathbf{x}_j) = (U\mathbf{x}_k,U\mathbf{x}_j) = \delta_{jk}$$

Therefore, from Lemma 13.4.3 again, this orthonormal set of vectors can be extended to an orthonormal basis for $\mathbb{F}^m$,

$$\{F\mathbf{x}_1,\cdots,F\mathbf{x}_r,\mathbf{z}_{r+1},\cdots,\mathbf{z}_m\}$$

Thus there are at least as many $\mathbf{z}_k$ as there are $\mathbf{y}_j$. Now for $\mathbf{x} \in \mathbb{F}^n$, since

$$\{U\mathbf{x}_1,\cdots,U\mathbf{x}_r,\mathbf{y}_{r+1},\cdots,\mathbf{y}_n\}$$

is an orthonormal basis for $\mathbb{F}^n$, there exist unique scalars,

$$c_1\cdots,c_r,d_{r+1},\cdots,d_n$$

such that

$$\mathbf{x} = \sum_{k=1}^{r} c_k U\mathbf{x}_k + \sum_{k=r+1}^{n} d_k\mathbf{y}_k$$

Define

$$R\mathbf{x} \equiv \sum_{k=1}^{r} c_k F\mathbf{x}_k + \sum_{k=r+1}^{n} d_k\mathbf{z}_k \tag{13.12}$$

Then also there exist scalars b_k such that

$$U\mathbf{x} = \sum_{k=1}^{r} b_k U\mathbf{x}_k$$

and so from 13.12,

$$RU\mathbf{x} = \sum_{k=1}^{r} b_k F\mathbf{x}_k = F\left(\sum_{k=1}^{r} b_k\mathbf{x}_k\right)$$

Is $F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) = F(\mathbf{x})$?

$$\left(F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) - F(\mathbf{x}), F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) - F(\mathbf{x})\right)$$

$$\begin{aligned}
&= \left((F^*F)\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right), \left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right)\right) \\
&= \left(U^2\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right), \left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right)\right) \\
&= \left(U\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right), U\left(\sum_{k=1}^{r} b_k \mathbf{x}_k - \mathbf{x}\right)\right) \\
&= \left(\sum_{k=1}^{r} b_k U\mathbf{x}_k - U\mathbf{x}, \sum_{k=1}^{r} b_k U\mathbf{x}_k - U\mathbf{x}\right) = 0
\end{aligned}$$

Therefore, $F\left(\sum_{k=1}^{r} b_k \mathbf{x}_k\right) = F(\mathbf{x})$ and this shows $RU\mathbf{x} = F\mathbf{x}$. From 13.12 and Lemma 13.3.7 R preserves distances. Therefore, by Lemma 13.4.4 $R^*R = I$. ■

13.5 The Singular Value Decomposition

In this section, A will be an $m \times n$ matrix. To begin with, here is a simple lemma.

Lemma 13.5.1 *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 13.5.2 *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 13.5.3 *Let A be an $m \times n$ matrix. Then there exist unitary matrices, U and V of the appropriate size such that*

$$U^*AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

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 13.2.14 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,

$$AA^*\mathbf{u}_i = AA^*\sigma_i^{-1}A\mathbf{v}_i = \sigma_i^{-1}AA^*A\mathbf{v}_i = \sigma_i^{-1}A\sigma_i^2\mathbf{v}_i = \sigma_i^2\mathbf{u}_i.$$

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

$$U^*AV = \begin{pmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_k^* \\ \vdots \\ \mathbf{u}_m^* \end{pmatrix} A(\mathbf{v}_1 \cdots \mathbf{v}_n) = \begin{pmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_k^* \\ \vdots \\ \mathbf{u}_m^* \end{pmatrix} (\sigma_1\mathbf{u}_1 \cdots \sigma_k\mathbf{u}_k, \mathbf{0} \cdots \mathbf{0}) = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

where σ is given in the statement of the theorem. ■

The singular value decomposition has as an immediate corollary the following interesting result.

Corollary 13.5.4 *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,

$$\operatorname{rank}(A^*) = \operatorname{rank}(V^*A^*U) = \operatorname{rank}\left((U^*AV)^*\right)$$

$$= \text{rank}\left(\left(\begin{array}{cc} \sigma & 0 \\ 0 & 0 \end{array}\right)^*\right) = \text{number of singular values. } \blacksquare$$

How could you go about computing the singular value decomposition? The proof of existence indicates how to do it. Here is an informal method. You have from the singular value decompositon,

$$A = U \left(\begin{array}{cc} \sigma & 0 \\ 0 & 0 \end{array}\right) V^*,\ A^* = V \left(\begin{array}{cc} \sigma & 0 \\ 0 & 0 \end{array}\right) U^*$$

Then it follows that

$$A^*A = V \left(\begin{array}{cc} \sigma & 0 \\ 0 & 0 \end{array}\right) U^*U \left(\begin{array}{cc} \sigma & 0 \\ 0 & 0 \end{array}\right) V^* = V \left(\begin{array}{cc} \sigma^2 & 0 \\ 0 & 0 \end{array}\right) V^*$$

and so $A^*AV = V \left(\begin{array}{cc} \sigma^2 & 0 \\ 0 & 0 \end{array}\right)$. Similarly, $AA^*U = U \left(\begin{array}{cc} \sigma^2 & 0 \\ 0 & 0 \end{array}\right)$. Therefore, you would find an orthonormal basis of eigenvectors for AA^* make them the columns of a matrix such that the corresponding eigenvalues are decreasing. This gives U. You could then do the same for A^*A to get V.

Example 13.5.5 *Find a singular value decomposition for the matrix*

$$A \equiv \left(\begin{array}{ccc} \frac{2}{5}\sqrt{2}\sqrt{5} & \frac{4}{5}\sqrt{2}\sqrt{5} & 0 \\ \frac{2}{5}\sqrt{2}\sqrt{5} & \frac{4}{5}\sqrt{2}\sqrt{5} & 0 \end{array}\right)$$

First consider A^*A

$$\left(\begin{array}{ccc} \frac{16}{5} & \frac{32}{5} & 0 \\ \frac{32}{5} & \frac{64}{5} & 0 \\ 0 & 0 & 0 \end{array}\right)$$

What are some eigenvalues and eigenvectors? Some computing shows these are

$$\left\{\left(\begin{array}{c} 0 \\ 0 \\ 1 \end{array}\right), \left(\begin{array}{c} -\frac{2}{5}\sqrt{5} \\ \frac{1}{5}\sqrt{5} \\ 0 \end{array}\right)\right\} \leftrightarrow 0, \left\{\left(\begin{array}{c} \frac{1}{5}\sqrt{5} \\ \frac{2}{5}\sqrt{5} \\ 0 \end{array}\right)\right\} \leftrightarrow 16$$

Thus the matrix V is given by

$$V = \left(\begin{array}{ccc} \frac{1}{5}\sqrt{5} & -\frac{2}{5}\sqrt{5} & 0 \\ \frac{2}{5}\sqrt{5} & \frac{1}{5}\sqrt{5} & 0 \\ 0 & 0 & 1 \end{array}\right)$$

Next consider $AA^* = \left(\begin{array}{cc} 8 & 8 \\ 8 & 8 \end{array}\right)$. Eigenvectors and eigenvalues are

$$\left\{\left(\begin{array}{c} -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{array}\right)\right\} \leftrightarrow 0, \left\{\left(\begin{array}{c} \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \end{array}\right)\right\} \leftrightarrow 16$$

In this case you can let U be given by

$$U = \begin{pmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix}$$

Lets check this. $U^*AV =$

$$\begin{pmatrix} \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \\ -\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix} \begin{pmatrix} \frac{2}{5}\sqrt{2}\sqrt{5} & \frac{4}{5}\sqrt{2}\sqrt{5} & 0 \\ \frac{2}{5}\sqrt{2}\sqrt{5} & \frac{4}{5}\sqrt{2}\sqrt{5} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{5}\sqrt{5} & -\frac{2}{5}\sqrt{5} & 0 \\ \frac{2}{5}\sqrt{5} & \frac{1}{5}\sqrt{5} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

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

This illustrates that if you have a good way to find the eigenvectors and eigenvalues for a Hermitian matrix which has nonnegative eigenvalues, then you also have a good way to find the singular value decomposition of an arbitrary matrix.

13.6 Approximation In The Frobenius Norm*

The Frobenius norm is one of many norms for a matrix. It is arguably the most obvious of all norms. First here is a short discussion of the trace.

Definition 13.6.1 *Let A be an $n \times n$ matrix. Then*

$$\text{trace}(A) \equiv \sum_i A_{ii}$$

just the sum of the entries on the main diagonal.

The fundamental property of the trace is in the next lemma.

Lemma 13.6.2 *Let $A = S^{-1}BS$. Then* $\text{trace}(A) = \text{trace}(B)$. *Also, for any two $n \times n$ matrices A, B*

$$\text{trace}(AB) = \text{trace}(BA)$$

Proof: Consider the displayed formula.

$$\text{trace}(AB) = \sum_i \sum_j A_{ij}B_{ji},\ \text{trace}(BA) = \sum_j \sum_i B_{ji}A_{ij}$$

they are the same thing. Thus if $A = S^{-1}BS$,

$$\text{trace}(A) = \text{trace}\left(S^{-1}(BS)\right) = \text{trace}\left(BSS^{-1}\right) = \text{trace}(B).\ \blacksquare$$

Here is the definition of the Frobenius norm.

Definition 13.6.3 *Let A be a complex $m \times n$ matrix. Then*

$$||A||_F \equiv (\text{trace}(AA^*))^{1/2}$$

Also this norm comes from the inner product

$$(A,B)_F \equiv \text{trace}\,(AB^*)$$

Thus $||A||_F^2$ *is easily seen to equal* $\sum_{ij} \left|a_{ij}\right|^2$ *so essentially, it treats the matrix as a vector in* $\mathbb{F}^{m\times n}$.

Lemma 13.6.4 *Let* A *be an* $m\times n$ *complex matrix with singular matrix*

$$\Sigma = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

with σ *as defined above. Then*

$$||\Sigma||_F^2 = ||A||_F^2 \tag{13.13}$$

and the following hold for the Frobenius norm. If U,V *are unitary and of the right size,*

$$||UA||_F = ||A||_F\,,\ ||UAV||_F = ||A||_F\,. \tag{13.14}$$

Proof: From the definition and letting U,V be unitary and of the right size,

$$||UA||_F^2 \equiv \text{trace}\,(UAA^*U^*) = \text{trace}\,(AA^*) = ||A||_F^2$$

Also,

$$||AV||_F^2 \equiv \text{trace}\,(AVV^*A^*) = \text{trace}\,(AA^*) = ||A||_F^2\,.$$

It follows

$$||UAV||_F^2 = ||AV||_F^2 = ||A||_F^2\,.$$

Now consider 13.13. From what was just shown,

$$||A||_F^2 = ||U\Sigma V^*||_F^2 = ||\Sigma||_F^2\,. \blacksquare$$

Of course, this shows that

$$||A||_F^2 = \sum_i \sigma_i^2,$$

the sum of the squares of the singular values of A.

Why is the singular value decomposition important? It implies

$$A = U\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}V^*$$

where σ is the diagonal matrix having the singular values down the diagonal. Now sometimes A is a huge matrix, 1000×2000 or something like that. This happens in applications to situations where the entries of A describe a picture. What also happens is that most of the singular values are very small. What if you deleted those which were very small, say for all $i\geq l$ and got a new matrix,

$$A' \equiv U\begin{pmatrix} \sigma' & 0 \\ 0 & 0 \end{pmatrix}V^*?$$

Then the entries of A' would end up being close to the entries of A but there is much less information to keep track of. This turns out to be very useful. More precisely, letting

$$\sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix}, \; U^*AV = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix},$$

$$\left|\left|A - A'\right|\right|_F^2 = \left|\left| U \begin{pmatrix} \sigma - \sigma' & 0 \\ 0 & 0 \end{pmatrix} V^* \right|\right|_F^2 = \sum_{k=l+1}^{r} \sigma_k^2$$

Thus A is approximated by A' where A' has rank $l < r$. In fact, it is also true that out of all matrices of rank l, this A' is the one which is closest to A in the Frobenius norm.

Thus A is approximated by A' where A' has rank $l < r$. In fact, it is also true that out of all matrices of rank l, this A' is the one which is closest to A in the Frobenius norm. Here is roughly why this is so. First consider approximating

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

as well as possible with a rank 2 matrix. It seems clear that the one which will work best is

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

More generally if σ is a $r \times r$ diagonal matrix in which the positive diagonal entries are decreasing from upper left to lower right, then the best rank l approximation to

$$\begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix}$$

would be

$$\begin{pmatrix} \sigma' & 0 \\ 0 & 0 \end{pmatrix}$$

where σ' is the upper left $l \times l$ corner of σ as in the above example.

Now suppose A is an $m \times n$ matrix. Let U, V be unitary and of the right size such that

$$U^*AV = \begin{pmatrix} \sigma_{r\times r} & 0 \\ 0 & 0 \end{pmatrix}$$

Then suppose B approximates A as well as possible in the Frobenius norm. Then you would want

$$\|A - B\| = \|U^*AV - U^*BV\| = \left\| \begin{pmatrix} \sigma_{r\times r} & 0 \\ 0 & 0 \end{pmatrix} - U^*BV \right\|$$

to be as small as possible. Therefore, from the above discussion, you should have

$$U^*BV = \begin{pmatrix} \sigma' & 0 \\ 0 & 0 \end{pmatrix}, B = U \begin{pmatrix} \sigma' & 0 \\ 0 & 0 \end{pmatrix} V^*$$

whereas

$$A = U \begin{pmatrix} \sigma_{r\times r} & 0 \\ 0 & 0 \end{pmatrix} V^*$$

13.7 Moore Penrose Inverse*

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 13.3.3 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^*.

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix} V^*\mathbf{x} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y}. \tag{13.15}$$

One solution to this equation which is very easy to spot is

$$\mathbf{x} = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y}. \tag{13.16}$$

This special $\mathbf{x}$ is denoted by $A^+\mathbf{y}$. The matrix $V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*$ is denoted by A^+. Thus $\mathbf{x}$ just defined is a solution to the least squares problem of finding the $\mathbf{x}$ such that $A\mathbf{x}$ is as close as possible to $\mathbf{y}$. Suppose now that $\mathbf{z}$ is some other solution to this least squares problem. Thus from the above,

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix} V^*\mathbf{z} = \begin{pmatrix} \sigma & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y}$$

and so, multiplying both sides by $\begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 0 \end{pmatrix}$,

$$\begin{pmatrix} I_{r\times r} & 0 \\ 0 & 0 \end{pmatrix} V^*\mathbf{z} = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y}$$

To make $V^*\mathbf{z}$ as small as possible, you would have only the first r entries of $V^*\mathbf{z}$ be nonzero since the later ones will be zeroed out anyway so they are unnecessary. Hence

$$V^*\mathbf{z} = \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y}$$

and consequently

$$\mathbf{z} = V \begin{pmatrix} \sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*\mathbf{y} \equiv A^+\mathbf{y}$$

However, minimizing $|V^*\mathbf{z}|$ is the same as minimizing $|\mathbf{z}|$ because V is unitary. Hence $A^+\mathbf{y}$ is the solution to the least squares problem which has smallest norm.

13.8 MATLAB And Singular Value Decomposition

MATLAB can find this very well. The syntax is [U,S,V]=svd(A) and it will give you the unitary matrices U,V such that $U^*AV = S$ where S is the singular value matrix. Here is an example.

```
A=[1,2,5;3,-2,-1];
[U,S,V]=svd(A)
```

Then press return to get the desired matrices. Check your work by typing at >> U'*A*V and press enter to see S.

MATLAB can also find the Moore Penrose inverse or pseudoinverse as follows. First enter A followed by ; and then type B=pinv(A) and press return. It will give the pseudoinverse. Here is an example where A does not have an inverse.

```
A=[1,2,3;2,4,6;-3,-2,1];
B=pinv(A)
```

13.9 Exercises

1. Here are some matrices. Label according to whether they are symmetric, skew symmetric, or orthogonal. If the matrix is orthogonal, determine whether it is proper or improper.

   (a) $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$

   (b) $\begin{pmatrix} 1 & 2 & -3 \\ 2 & 1 & 4 \\ -3 & 4 & 7 \end{pmatrix}$

   (c) $\begin{pmatrix} 0 & -2 & -3 \\ 2 & 0 & -4 \\ 3 & 4 & 0 \end{pmatrix}$

2. Show that every real matrix may be written as the sum of a skew symmetric and a symmetric matrix. **Hint:** If A is an $n \times n$ matrix, show that $B \equiv \frac{1}{2}\left(A - A^T\right)$ is skew symmetric.

3. Let $\mathbf{x}$ be a vector in $\mathbb{R}^n$ and consider the matrix $I - \frac{2\mathbf{x}\mathbf{x}^T}{||\mathbf{x}||^2}$. Show this matrix is both symmetric and orthogonal.

4. For U an orthogonal matrix, explain why $||U\mathbf{x}|| = ||\mathbf{x}||$ for any vector $\mathbf{x}$. Next explain why if U is an $n \times n$ matrix with the property that $||U\mathbf{x}|| = ||\mathbf{x}||$ for all vectors, $\mathbf{x}$, then U must be orthogonal. Thus the orthogonal matrices are exactly those which preserve distance.

5. A quadratic form in three variables is an expression of the form $a_1x^2 + a_2y^2 + a_3z^2 + a_4xy + a_5xz + a_6yz$. Show that every such quadratic form may be written as

$$\begin{pmatrix} x & y & z \end{pmatrix} A \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

where A is a symmetric matrix.

6. Given a quadratic form in three variables, x, y, and z, show there exists an orthogonal matrix U and variables x', y', z' such that $\begin{pmatrix} x & y & z \end{pmatrix}^T = U \begin{pmatrix} x' & y' & z' \end{pmatrix}^T$ with the property that in terms of the new variables, the quadratic form is

$$\lambda_1 (x')^2 + \lambda_2 (y')^2 + \lambda_3 (z')^2$$

where the numbers, λ_1, λ_2, and λ_3 are the eigenvalues of the matrix A in Problem 5.

7. If A is a symmetric invertible matrix, is it always the case that A^{-1} must be symmetric also? How about A^k for k a positive integer? Explain.

8. If A, B are symmetric matrices, does it follow that AB is also symmetric?

9. Suppose A, B are symmetric and $AB = BA$. Does it follow that AB is symmetric?

10. Here are some matrices. What can you say about the eigenvalues of these matrices just by looking at them?

(a) $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$

(b) $\begin{pmatrix} 1 & 2 & -3 \\ 2 & 1 & 4 \\ -3 & 4 & 7 \end{pmatrix}$

(c) $\begin{pmatrix} 0 & -2 & -3 \\ 2 & 0 & -4 \\ 3 & 4 & 0 \end{pmatrix}$

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

11. Find the eigenvalues and eigenvectors of the matrix $\begin{pmatrix} c & 0 & 0 \\ 0 & 0 & -b \\ 0 & b & 0 \end{pmatrix}$. Here b, c are real numbers.

12. Find the eigenvalues and eigenvectors of the matrix $\begin{pmatrix} c & 0 & 0 \\ 0 & a & -b \\ 0 & b & a \end{pmatrix}$. Here a, b, c are real numbers.

13. Find the eigenvalues and an orthonormal basis of eigenvectors for A.

$$A = \begin{pmatrix} 11 & -1 & -4 \\ -1 & 11 & -4 \\ -4 & -4 & 14 \end{pmatrix}.$$

Hint: Two eigenvalues are 12 and 18.

14. Find the eigenvalues and an orthonormal basis of eigenvectors for A.

$$A = \begin{pmatrix} 4 & 1 & -2 \\ 1 & 4 & -2 \\ -2 & -2 & 7 \end{pmatrix}.$$

Hint: One eigenvalue is 3.

15. Show that if A is a real symmetric matrix and λ and μ are two different eigenvalues, then if $\mathbf{x}$ is an eigenvector for λ and $\mathbf{y}$ is an eigenvector for μ, then $\mathbf{x} \cdot \mathbf{y} = 0$. Also all eigenvalues are real. Supply reasons for each step in the following argument. First

$$\lambda \mathbf{x}^T \overline{\mathbf{x}} = (A\mathbf{x})^T \overline{\mathbf{x}} = \mathbf{x}^T A \overline{\mathbf{x}} = \mathbf{x}^T \overline{A\mathbf{x}} = \mathbf{x}^T \overline{\lambda} \overline{\mathbf{x}} = \overline{\lambda} \mathbf{x}^T \overline{\mathbf{x}}$$

and so $\lambda = \overline{\lambda}$. This shows that all eigenvalues are real. It follows all the eigenvectors are real. Why? Now let $\mathbf{x}, \mathbf{y}, \mu$ and λ be given as above.

$$\lambda (\mathbf{x} \cdot \mathbf{y}) = \lambda \mathbf{x} \cdot \mathbf{y} = A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A\mathbf{y} = \mathbf{x} \cdot \mu \mathbf{y} = \mu (\mathbf{x} \cdot \mathbf{y}) = \mu (\mathbf{x} \cdot \mathbf{y})$$

and so

$$(\lambda - \mu) \mathbf{x} \cdot \mathbf{y} = 0.$$

Since $\lambda \neq \mu$, it follows $\mathbf{x} \cdot \mathbf{y} = 0$.

16. Suppose U is an orthogonal $n \times n$ matrix. Explain why $\operatorname{rank}(U) = n$.

17. Show that if A is an Hermitian matrix and λ and μ are two different eigenvalues, then if $\mathbf{x}$ is an eigenvector for λ and $\mathbf{y}$ is an eigenvector for μ, then $\mathbf{x} \cdot \mathbf{y} = 0$. Also all eigenvalues are real. Supply reasons for each step in the following argument. First

$$\lambda \mathbf{x} \cdot \mathbf{x} = A\mathbf{x} \cdot \mathbf{x} = \mathbf{x} \cdot A\mathbf{x} = \mathbf{x} \cdot \lambda \mathbf{x} = \overline{\lambda} \mathbf{x} \cdot \mathbf{x}$$

and so $\lambda = \overline{\lambda}$. This shows that all eigenvalues are real. Now let $\mathbf{x}, \mathbf{y}, \mu$ and λ be given as above.

$$\lambda (\mathbf{x} \cdot \mathbf{y}) = \lambda \mathbf{x} \cdot \mathbf{y} = A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A\mathbf{y} = \mathbf{x} \cdot \mu \mathbf{y} = \overline{\mu} (\mathbf{x} \cdot \mathbf{y}) = \mu (\mathbf{x} \cdot \mathbf{y})$$

and so $(\lambda - \mu) \mathbf{x} \cdot \mathbf{y} = 0$. Since $\lambda \neq \mu$, it follows $\mathbf{x} \cdot \mathbf{y} = 0$.

18. Show that the eigenvalues and eigenvectors of a real matrix occur in conjugate pairs.

19. If a real matrix A has all real eigenvalues, does it follow that A must be symmetric. If so, explain why and if not, give an example to the contrary.

20. Suppose A is a 3×3 symmetric matrix and you have found two eigenvectors which form an orthonormal set. Explain why their cross product is also an eigenvector.

21. Study the definition of an orthonormal set of vectors. Write it from memory.

22. Determine which of the following sets of vectors are orthonormal sets. Justify your answer.

    (a) $\{(1,1),(1,-1)\}$

    (b) $\left\{\left(\frac{1}{\sqrt{2}},\frac{-1}{\sqrt{2}}\right),(1,0)\right\}$

    (c) $\left\{\left(\frac{1}{3},\frac{2}{3},\frac{2}{3}\right),\left(\frac{-2}{3},\frac{-1}{3},\frac{2}{3}\right),\left(\frac{2}{3},\frac{-2}{3},\frac{1}{3}\right)\right\}$

23. Show that if $\{\mathbf{u}_1,\cdots,\mathbf{u}_n\}$ is an orthonormal set of vectors in $\mathbb{F}^n$, then it is a basis. **Hint:** It was shown earlier that this is a linearly independent set. If you wish, replace $\mathbb{F}^n$ with $\mathbb{R}^n$. Do this version if you do not know the dot product for vectors in $\mathbb{C}^n$.

24. Fill in the missing entries to make the matrix orthogonal.

$$\begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \_ & \_ \\ \_ & \frac{\sqrt{6}}{3} & \_ \end{pmatrix}.$$

25. Fill in the missing entries to make the matrix orthogonal.

$$\begin{pmatrix} \frac{2}{3} & \frac{\sqrt{2}}{2} & \frac{1}{6}\sqrt{2} \\ \frac{2}{3} & \_ & \_ \\ \_ & 0 & \_ \end{pmatrix}$$

26. Fill in the missing entries to make the matrix orthogonal.

$$\begin{pmatrix} \frac{1}{3} & -\frac{2}{\sqrt{5}} & \_ \\ \frac{2}{3} & 0 & \_ \\ \_ & \_ & \frac{4}{15}\sqrt{5} \end{pmatrix}$$

27. Find the eigenvalues and an orthonormal basis of eigenvectors for A. Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T AU = D$.

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

Hint: One eigenvalue is -2.

28. Find the eigenvalues and an orthonormal basis of eigenvectors for A. Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T AU = D$.

$$A = \begin{pmatrix} 17 & -7 & -4 \\ -7 & 17 & -4 \\ -4 & -4 & 14 \end{pmatrix}.$$

Hint: Two eigenvalues are 18 and 24.

29. Find the eigenvalues and an orthonormal basis of eigenvectors for A. Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T AU = D$.

$$A = \begin{pmatrix} 13 & 1 & 4 \\ 1 & 13 & 4 \\ 4 & 4 & 10 \end{pmatrix}.$$

Hint: Two eigenvalues are 12 and 18.

30. Find the eigenvalues and an orthonormal basis of eigenvectors for A. Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T AU = D$.

$$A = \begin{pmatrix} -\frac{5}{3} & \frac{1}{15}\sqrt{6}\sqrt{5} & \frac{8}{15}\sqrt{5} \\ \frac{1}{15}\sqrt{6}\sqrt{5} & -\frac{14}{5} & -\frac{1}{15}\sqrt{6} \\ \frac{8}{15}\sqrt{5} & -\frac{1}{15}\sqrt{6} & \frac{7}{15} \end{pmatrix}$$

Hint: The eigenvalues are $-3, -2, 1$.

31. Find the eigenvalues and an orthonormal basis of eigenvectors for A. Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T AU = D$.

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

32. Find the eigenvalues and an orthonormal basis of eigenvectors for A. Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T AU = D$.

$$A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 5 & 1 \\ 0 & 1 & 5 \end{pmatrix}.$$

33. Find the eigenvalues and an orthonormal basis of eigenvectors for A. Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^T AU = D$.

$$A = \begin{pmatrix} \frac{4}{3} & \frac{1}{3}\sqrt{3}\sqrt{2} & \frac{1}{3}\sqrt{2} \\ \frac{1}{3}\sqrt{3}\sqrt{2} & 1 & -\frac{1}{3}\sqrt{3} \\ \frac{1}{3}\sqrt{2} & -\frac{1}{3}\sqrt{3} & \frac{5}{3} \end{pmatrix}$$

Hint: The eigenvalues are $0,2,2$ where 2 is listed twice because it is a root of multiplicity 2.

34. Find the eigenvalues and an orthonormal basis of eigenvectors for A. Diagonalize A by finding an orthogonal matrix U and a diagonal matrix D such that $U^TAU = D$.

$$A = \begin{pmatrix} 1 & \frac{1}{6}\sqrt{3}\sqrt{2} & \frac{1}{6}\sqrt{3}\sqrt{6} \\ \frac{1}{6}\sqrt{3}\sqrt{2} & \frac{3}{2} & \frac{1}{12}\sqrt{2}\sqrt{6} \\ \frac{1}{6}\sqrt{3}\sqrt{6} & \frac{1}{12}\sqrt{2}\sqrt{6} & \frac{1}{2} \end{pmatrix}$$

Hint: The eigenvalues are $2,1,0$.

35. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{6}\sqrt{3}\sqrt{2} & -\frac{7}{18}\sqrt{3}\sqrt{6} \\ \frac{1}{6}\sqrt{3}\sqrt{2} & \frac{3}{2} & -\frac{1}{12}\sqrt{2}\sqrt{6} \\ -\frac{7}{18}\sqrt{3}\sqrt{6} & -\frac{1}{12}\sqrt{2}\sqrt{6} & -\frac{5}{6} \end{pmatrix}$$

Hint: The eigenvalues are $1,2,-2$.

36. Find the eigenvalues and an orthonormal basis of eigenvectors for the matrix

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

Hint: The eigenvalues are $-1,2,-1$ where -1 is listed twice because it has multiplicity 2 as a zero of the characteristic equation.

37. Explain why a matrix A is symmetric if and only if there exists an orthogonal matrix U such that $A = U^TDU$ for D a diagonal matrix.

38. The proof of Theorem 13.3.3 concluded with the following observation. If $-ta+t^2b \geq 0$ for all $t \in \mathbb{R}$ and $b \geq 0$, then $a=0$. Why is this so?

39. Using Schur's theorem, show that whenever A is an $n \times n$ matrix, $\det(A)$ equals the product of the eigenvalues of A.

40. In the proof of Theorem 13.3.8 the following argument was used. If $\mathbf{x} \cdot \mathbf{w} = 0$ for all $\mathbf{w} \in \mathbb{R}^n$, then $\mathbf{x} = \mathbf{0}$. Why is this so?

41. Using Corollary 13.3.9 show that a real $m \times n$ matrix is onto if and only if its transpose is one to one.

42. Suppose A is a 3×2 matrix. Is it possible that A^T is one to one? What does this say about A being onto? Prove your answer.

43. Find the least squares solution to the system $x+2y=1, 2x+3y=2, 3x+5y=4$.

44. You are doing experiments and have obtained the ordered pairs,

$$(0,1),(1,2),(2,3.5),(3,4)$$

Find m and b such that $y=mx+b$ approximates these four points as well as possible. Now do the same thing for $y=ax^2+bx+c$, finding a,b, and c to give the best approximation.

45. Suppose you have several ordered triples, (x_i,y_i,z_i). Describe how to find a polynomial,

$$z=a+bx+cy+dxy+ex^2+fy^2$$

for example giving the best fit to the given ordered triples. Is there any reason you have to use a polynomial? Would similar approaches work for other combinations of functions just as well?

46. Find an orthonormal basis for the spans of the following sets of vectors.

    (a) $(3,-4,0),(7,-1,0),(1,7,1)$.

    (b) $(3,0,-4),(11,0,2),(1,1,7)$

    (c) $(3,0,-4),(5,0,10),(-7,1,1)$

47. Using the Gram Schmidt process or the QR factorization, find an orthonormal basis for the span of the vectors, $(1,2,1),(2,-1,3)$, and $(1,0,0)$.

48. Using the Gram Schmidt process or the QR factorization, find an orthonormal basis for the span of the vectors, $(1,2,1,0),(2,-1,3,1)$, and $(1,0,0,1)$.

49. The set, $V \equiv \{(x,y,z) : 2x+3y-z=0\}$ is a subspace of $\mathbb{R}^3$. Find an orthonormal basis for this subspace.

50. The two level surfaces, $2x+3y-z+w=0$ and $3x-y+z+2w=0$ intersect in a subspace of $\mathbb{R}^4$, find a basis for this subspace. Next find an orthonormal basis for this subspace.

51. Let A,B be a $m\times n$ matrices. Define an inner product on the set of $m\times n$ matrices by
$$(A,B)_F \equiv \text{trace}\,(AB^*)\,.$$
Show this is an inner product satisfying all the inner product axioms. Recall for M an $n\times n$ matrix, 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.

52. 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.

53. The trace of an $n\times n$ matrix M is defined as $\sum_i M_{ii}$. In other words it is the sum of the entries on the main diagonal. If A,B are $n\times n$ matrices, show trace $(AB) =$ trace (BA). Now explain why if $A=S^{-1}BS$ it follows trace $(A)=$ trace (B). **Hint:** For the first part, write these in terms of components of the matrices and it just falls out.

54. Using Problem 53 and Schur's theorem, show that the trace of an $n\times n$ matrix equals the sum of the eigenvalues.

55. If A is a general $n\times n$ matrix having possibly repeated eigenvalues, show there is a sequence $\{A_k\}$ of $n\times n$ matrices having distinct eigenvalues which has the property that the ij^{th} entry of A_k converges to the ij^{th} entry of A for all ij. **Hint:** Use Schur's theorem.

56. Prove the Cayley Hamilton theorem as follows. First suppose A has a basis of eigenvectors $\{\mathbf{v}_k\}_{k=1}^n, A\mathbf{v}_k=\lambda_k\mathbf{v}_k$. Let $p(\lambda)$ be the characteristic polynomial. Show $p(A)\mathbf{v}_k = p(\lambda_k)\mathbf{v}_k=\mathbf{0}$. Then since $\{\mathbf{v}_k\}$ is a basis, it follows $p(A)\mathbf{x}=\mathbf{0}$ for all $\mathbf{x}$ and so $p(A)=0$. Next in the general case, use Problem 55 to obtain a sequence $\{A_k\}$ of matrices whose entries converge to the entries of A such that A_k has n distinct eigenvalues and therefore by Theorem 12.1.15 A_k has a basis of eigenvectors. Therefore, from the first part and for $p_k(\lambda)$ the characteristic polynomial for A_k, it follows $p_k(A_k)=0$. Now explain why and the sense in which $\lim_{k\to\infty} p_k(A_k)=p(A)$.

57. Show that the Moore Penrose inverse A^+ satisfies the following conditions.
$$AA^+A=A,\ \ A^+AA^+=A^+,\ \ A^+A,\ AA^+ \text{ are Hermitian.}$$
Next show that if A_0 satisfies the above conditions, then it must be the Moore Penrose inverse and that if A is an $n\times n$ invertible matrix, then A^{-1} satisfies the above conditions. Thus the Moore Penrose inverse generalizes the usual notion of inverse but does not contradict it. **Hint:** Let
$$U^*AV=\Sigma\equiv\begin{pmatrix}\sigma & 0\\ 0 & 0\end{pmatrix}$$
and suppose
$$V^+A_0U=\begin{pmatrix}P & Q\\ R & S\end{pmatrix}$$
where P is the same size as σ. Now use the conditions to identify $P=\sigma, Q=0$ etc.

58. Find the least squares solution to

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1+\varepsilon \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

Next suppose ε is so small that all ε^2 terms are ignored by the computer but the terms of order ε are not ignored. Show the least squares equations in this case reduce to

$$\begin{pmatrix} 3 & 3+\varepsilon \\ 3+\varepsilon & 3+2\varepsilon \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a+b+c \\ a+b+(1+\varepsilon)c \end{pmatrix}.$$

Find the solution to this and compare the y values of the two solutions. Show that one of these is -2 times the other. This illustrates a problem with the technique for finding least squares solutions presented as the solutions to $A^*A\mathbf{x} = A^*\mathbf{y}$. One way of dealing with this problem is to use the QR factorization. This is illustrated in the next problem. It turns out that this helps alleviate some of the round off difficulties of the above.

59. Show that the equations $A^*A\mathbf{x} = A^*\mathbf{y}$ can be written as $R^*R\mathbf{x} = R^*Q^*\mathbf{y}$ where R is upper triangular and R^* is lower triangular. Explain how to solve this system efficiently. **Hint:** You first find $R\mathbf{x}$ and then you find $\mathbf{x}$ which will not be hard because R is upper triangular.

60. Show that $A^+ = (A^*A)^+ A^*$. **Hint:** You might use the description of A^+ in terms of the singular value decomposition.

61. Let $A = \begin{pmatrix} 1 & -3 & 0 \\ 3 & -1 & 0 \end{pmatrix}$. Then

$$\begin{pmatrix} -\sqrt{2}/2 & \sqrt{2}/2 & 0 \\ \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}^T A^T A \begin{pmatrix} -\sqrt{2}/2 & \sqrt{2}/2 & 0 \\ \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} 16 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$AA^T = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$. A matrix U with

$$U^T AA^T U = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}$$

is $\begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix}$. However,

$$\begin{aligned} & \begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix}^T \begin{pmatrix} 1 & -3 & 0 \\ 3 & -1 & 0 \end{pmatrix} \begin{pmatrix} -\sqrt{2}/2 & \sqrt{2}/2 & 0 \\ \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ = & \begin{pmatrix} -4 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix}. \end{aligned}$$

How can this be fixed so that you get $\begin{pmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix}$?

Chapter 14

Numerical Solutions of Linear Systems

14.1 Iterative Methods For Linear Systems

Consider the problem of solving the equation

$$A\mathbf{x} = \mathbf{b} \tag{14.1}$$

where A is an $n \times n$ matrix. In many applications, the matrix A is huge and composed mainly of zeros. For such matrices, the method of Gauss elimination (row operations) is not a good way to solve the system because the row operations can destroy the zeros and storing all those zeros takes a lot of room in a computer. These systems are called sparse. To solve them it is common to use an iterative technique. The idea is to obtain a sequence of approximate solutions which get close to the true solution after a sufficient number of iterations.

Definition 14.1.1 *Let $\{\mathbf{x}_k\}_{k=1}^{\infty}$ be a sequence of vectors in $\mathbb{F}^n$. Say*

$$\mathbf{x}_k = \left(x_1^k, \cdots, x_n^k \right).$$

Then this sequence is said to converge to the vector $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{F}^n$, written as

$$\lim_{k \to \infty} \mathbf{x}_k = \mathbf{x}$$

if for each $j = 1, 2, \cdots, n$,

$$\lim_{k \to \infty} x_j^k = x_j.$$

In words, the sequence converges if the entries of the vectors in the sequence converge to the corresponding entries of $\mathbf{x}$.

Example 14.1.2 *Consider $\mathbf{x}_k = \left(\sin(1/k), \frac{k^2}{1+k^2}, \ln\left(\frac{1+k^2}{k^2}\right) \right)$. Find $\lim_{k \to \infty} \mathbf{x}_k$.*

From the above definition, this limit is the vector $(0,1,0)$ because

$$\lim_{k\to\infty}\sin(1/k)=0,\ \lim_{k\to\infty}\frac{k^2}{1+k^2}=1,\ \text{ and }\lim_{k\to\infty}\ln\left(\frac{1+k^2}{k^2}\right)=0.$$

A more complete mathematical explanation is given in [13].

14.1.1 The Jacobi Method

The first technique to be discussed here is the Jacobi method which is described in the following definition. In this technique, you have a sequence of vectors, $\left\{\mathbf{x}^k\right\}$ which converge to the solution to the linear system of equations and to get the i^{th} component of the $\mathbf{x}^{k+1}$, you use all the components of $\mathbf{x}^k$ except for the i^{th}. The precise description follows.

Definition 14.1.3 *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 14.1. The vectors in this list are called* ***iterates*** *and they are obtained according to the following procedure. Letting* $A=(a_{ij})$,

$$a_{ii}x_i^{r+1}=-\sum_{j\neq i}a_{ij}x_j^r+b_i. \tag{14.2}$$

In terms of matrices, letting

$$A=\begin{pmatrix} a_{11} & \cdots & a_{1n}\\ \vdots & \ddots & \vdots\\ a_{n1} & \cdots & a_{nn}\end{pmatrix}$$

The iterates are defined as

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0\\ 0 & a_{22} & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & a_{nn}\end{pmatrix}\begin{pmatrix} x_1^{r+1}\\ x_2^{r+1}\\ \vdots\\ x_n^{r+1}\end{pmatrix}$$

$$= \ -\begin{pmatrix} 0 & a_{12} & \cdots & a_{1n}\\ a_{21} & 0 & \ddots & \vdots\\ \vdots & \ddots & \ddots & a_{n-1n}\\ a_{n1} & \cdots & a_{nn-1} & 0\end{pmatrix}\begin{pmatrix} x_1^r\\ x_2^r\\ \vdots\\ x_n^r\end{pmatrix}+\begin{pmatrix} b_1\\ b_2\\ \vdots\\ b_n\end{pmatrix} \tag{14.3}$$

If these iterates do converge, then the vector to which they converge will be a solution to the original system of equations.

The matrix on the left in 14.3 is obtained by retaining the main diagonal of A and setting every other entry equal to zero. The matrix on the right in 14.3 is obtained from A by setting every diagonal entry equal to zero and retaining all the other entries unchanged.

Example 14.1.4 *Use the Jacobi method to solve the system*

$$\begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

In terms of the matrices, the Jacobi iteration is of the form

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ x_3^{r+1} \\ x_4^{r+1} \end{pmatrix} = - \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 0 \end{pmatrix} \begin{pmatrix} x_1^r \\ x_2^r \\ x_3^r \\ x_4^r \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

14.2 Using MATLAB To Iterate

The syntax you can use to accomplish this iteration is as follows. This is for the purposes of illustration. In fact, you would not take an inverse of one of the matrices in practice.

```
D=[3 0 0 0;0 4 0 0;0 0 5 0;0 0 0 4];
O=[0 1 0 0;1 0 1 0;0 2 0 1;0 0 2 0];
d=1; x=[0;0;0;0]; b=[1;2;3;4]; k=1; F=inv(D);
while d >.0000001 & k <1000
y=-F*O*x+F*b; d=(y-x)’*(y-x); k=k+1;
x=y;
end
x
k
(((D+O)*x-b)’*((D+O)*x-b))^(1/2)
```

It is going to iterate till $|y-x|^2$ is smaller than 10^{-7} or 1000 iterations, whichever comes first. Of course $y = \mathbf{x}_{r+1}$ and $x = \mathbf{x}_r$. The next to last line which has k tells you how many iterations it took to get there and the bottom line tells you how close $\mathbf{x}$ is to solving the equation. This yields

$$x = \begin{pmatrix} .2069 \\ .3793 \\ .2759 \\ .8621 \end{pmatrix}, \; k = 14, \; 6.1753 \times 10^{-4}$$

14.2.1 The Gauss Seidel Method

The Gauss Seidel method differs from the Jacobi method in using x_j^{k+1} for all $j < i$ in going from $\mathbf{x}^k$ to $\mathbf{x}^{k+1}$. This is why it is called the method of successive corrections. The precise description of this method is in the following definition.

Definition 14.2.1 *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*

$$\sum_{j=1}^{i} a_{ij} x_j^{r+1} = -\sum_{j=i+1}^{n} a_{ij} x_j^{r} + b_i. \tag{14.4}$$

In terms of matrices, letting

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

The iterates are defined as

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n1} & \cdots & a_{nn-1} & a_{nn} \end{pmatrix} \begin{pmatrix} x_1^{r+1} \\ x_2^{r+1} \\ \vdots \\ x_n^{r+1} \end{pmatrix}$$

$$= \quad -\begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1n} \\ 0 & \cdots & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^{r} \\ x_2^{r} \\ \vdots \\ x_n^{r} \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{14.5}$$

If $\mathbf{x}_r$ converges to some $\mathbf{x}$ then this will be a solution to the original equation.

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 14.4 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 14.2.2 *Use the Gauss Seidel method to solve the system*

$$\begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 2 & 5 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

You can use MATLAB in the same way. You just use different matrices. As is the case with the Jacobi method, you would not invert the matrix but would use the description of the method given above. The following works fine for small systems of equations however.

```
L=[3 0 0 0;1 4 0 0;0 2 5 0;0 0 2 4];
U=[0 1 0 0;0 0 1 0;0 0 0 1;0 0 0 0];
d=1; x=[0;0;0;0]; b=[1;2;3;4]; k=1; F=inv(L);
while d >.0000001 & k <1000
y=-F*U*x+F*b; d=(y-x)’*(y-x); k=k+1;
x=y;
end
x
k
(((L+U)*x-b)’*((L+U)*x-b))ˆ(1/2)
```

This yields

$$\begin{pmatrix} .207 \\ .3793 \\ .2759 \\ .8621 \end{pmatrix}, \; k = 8, \; 1.581 \times 10^{-4}$$

Thus it took only 8 iterations rather than 14. This is typical. The Gauss Seidel method is more complicated but tends to converge more quickly.

Now consider the following example.

Example 14.2.3 *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

$$\begin{pmatrix} 6 \\ -\frac{5}{4} \\ 1 \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 6.0 \\ -1.25 \\ 1.0 \\ .5 \end{pmatrix}$$

Using the same iteration scheme, you get the following.

$$1.0\times 10^{45}\begin{pmatrix}-8.2309\\ 2.2839\\ -1.004\\ .502\end{pmatrix},\ k=1000,\ 9.114\times 10^{44}$$

Thus the answer is totally useless, this after 1000 iterations. The error in approximating the solution is larger than 10^{44}. In other words, the method failed spectacularly to converge to anything. Row operations worked fine but this iterative procedure failed.

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? A complete answer to this question is given in more advanced linear algebra books. You can also see it in [13].

Both iterative procedures for solving

$$A\mathbf{x}=\mathbf{b} \tag{14.6}$$

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

$$\mathbf{x}^{r+1}=-B^{-1}C\mathbf{x}^{r}+B^{-1}\mathbf{b}. \tag{14.7}$$

Theorem 14.2.4 *Let $A=B+C$ and suppose all eigenvalues of $B^{-1}C$ have absolute value less than 1 where $A=B+C$. Then the iterates in 14.7 converge to the unique solution of 14.6.*

A complete explanation of this important result is found in more advanced linear algebra books. You can also see it in [13]. It depends on a theorem of Gelfand which is completely proved in this reference. Theorem 14.2.4 is very remarkable because it gives an algebraic condition for convergence, which is essentially an analytical question.

14.3 The Operator Norm*

Recall that for $\mathbf{x}\in\mathbb{C}^n$,

$$|\mathbf{x}|\equiv\sqrt{\langle\mathbf{x},\mathbf{x}\rangle}$$

Also recall Theorem 3.2.17 which says that

$$|\mathbf{z}|\geq 0 \text{ and } |\mathbf{z}|=0 \text{ if and only if } \mathbf{z}=\mathbf{0} \tag{14.8}$$

$$\text{If } \alpha \text{ is a scalar, } |\alpha \mathbf{z}| = |\alpha| |\mathbf{z}| \tag{14.9}$$

$$|\mathbf{z}+\mathbf{w}| \leq |\mathbf{z}| + |\mathbf{w}|. \tag{14.10}$$

If you have the above axioms holding for $||\cdot||$ replacing $|\cdot|$, then $||\cdot||$ is called a norm. For example, you can easily verify that

$$||\mathbf{x}|| \equiv \max \{|x_i|, i = 1, \cdots, n : \mathbf{x} = (x_1, \cdots, x_n)\}$$

is a norm. However, there are many other norms.

One important observation is that $\mathbf{x} \mapsto ||\mathbf{x}||$ is a continuous function. This follows from the observation that from the triangle inequality,

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

Hence

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

and so

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

This section will involve some analysis. If you want to talk about norms, this is inevitable. It will need some of the theorems of calculus which are usually neglected. In particular, it needs the following result which is a case of the Heine Borel theorem. To see this proved, see any good calculus book, not most of the ones which are used in beginning courses on calculus.

Theorem 14.3.1 *Let S denote the points $\mathbf{x} \in \mathbb{F}^n$ such that $|\mathbf{x}| = 1$. Then if $\{\mathbf{x}_k\}_{k=1}^{\infty}$ is any sequence of points of S, there exists a subsequence which converges to a point of S.*

Definition 14.3.2 *Let A be an $m \times n$ matrix. Let $||\cdot||_k$ denote a norm on $\mathbb{C}^k$. Then the operator norm is defined as follows.*

$$||A|| \equiv \max \{||A\mathbf{x}||_m : ||\mathbf{x}||_n \leq 1\}$$

Lemma 14.3.3 *The operator norm is well defined and is in fact a norm on the vector space of $m \times n$ matrices.*

Proof: It has already been observed that the $m \times n$ matrices form a vector space starting on Page 85. Why is $||A|| < \infty$?

claim: There exists $c > 0$ such that whenever $||\mathbf{x}|| \leq 1$, it follows that $|\mathbf{x}| \leq c$.

Proof of the claim: If not, then there exists $\{\mathbf{x}_k\}$ such that $||\mathbf{x}_k|| \leq 1$ but $|\mathbf{x}_k| > k$ for $k = 1, 2, \cdots$. Then $|\mathbf{x}_k / |\mathbf{x}_k|| = 1$ and so by the Heine Borel theorem from calculus, there exists a further subsequence, still denoted by k such that

$$\left| \frac{\mathbf{x}_k}{|\mathbf{x}_k|} - \mathbf{y} \right| \to 0, \ |\mathbf{y}| = 1.$$

Letting

$$\frac{\mathbf{x}_k}{|\mathbf{x}_k|} = \sum_{i=1}^n a_i^k \mathbf{e}_i,\ \mathbf{y} = \sum_{i=1}^n a_i \mathbf{e}_i,$$

It follows that $\mathbf{a}^k \to \mathbf{a}$ in $\mathbb{F}^n$. Hence

$$\left\| \frac{\mathbf{x}_k}{|\mathbf{x}_k|} - \mathbf{y} \right\| \leq \sum_{i=1}^n \left| a_i^k - a_i \right| \|\mathbf{e}_i\|$$

which converges to 0. However,

$$\left\| \frac{\mathbf{x}_k}{|\mathbf{x}_k|} \right\| \leq \frac{1}{k}$$

and so, by continuity of $\|\cdot\|$ mentioned above,

$$\|\mathbf{y}\| = \lim_{k\to\infty} \left\| \frac{\mathbf{x}_k}{|\mathbf{x}_k|} \right\| = 0$$

Therefore, $\mathbf{y} = \mathbf{0}$ but also $|\mathbf{y}| = 1$, a contradiction. This proves the claim.

Now consider why $\|A\| < \infty$. Let c be as just described in the claim.

$$\sup\{\|A\mathbf{x}\|_m : \|\mathbf{x}\|_n \leq 1\} \leq \sup\{\|A\mathbf{x}\|_m : |\mathbf{x}| \leq c\}$$

Consider for $\mathbf{x}, \mathbf{y}$ with $|\mathbf{x}|, |\mathbf{y}| \leq c$

$$\|A\mathbf{x} - A\mathbf{y}\| = \left\| \sum_i A_{ij} (x_j - y_j) \mathbf{e}_i \right\|$$

$$\leq \sum_i |A_{ij}| \left| x_j - y_j \right| \|\mathbf{e}_i\| \leq C |\mathbf{x} - \mathbf{y}|$$

for some constant C. So $\mathbf{x} \mapsto A\mathbf{x}$ is continuous. Since the norm $\|\cdot\|_m$ is continuous also, it follows from the extreme value theorem of calculus that $\|A\mathbf{x}\|_m$ achieves its maximum on the compact set $\{\mathbf{x} : |\mathbf{x}| \leq c\}$. Thus $\|A\|$ is well defined. The only other issue of significance is the triangle inequality. However,

$$\begin{aligned} \|A+B\| &\equiv \max\{\|(A+B)\mathbf{x}\|_m : \|\mathbf{x}\|_n \leq 1\} \\ &\leq \max\{\|A\mathbf{x}\|_m + \|B\mathbf{x}\|_m : \|\mathbf{x}\|_n \leq 1\} \end{aligned}$$

$$\leq \max\{\|A\mathbf{x}\|_m : \|\mathbf{x}\|_n \leq 1\} + \max\{\|B\mathbf{x}\|_m : \|\mathbf{x}\|_n \leq 1\}$$

$$= \|A\| + \|B\|$$

Obviously $\|A\| = 0$ if and only if $A = 0$. The rule for scalars is also immediate. ■

The operator norm is one way to describe the magnitude of a matrix. Earlier the Frobenius norm was discussed. The Frobenius norm is actually not used as much as the operator norm. Recall that the Frobenius norm involved considering the $m \times n$ matrix as a vector in $\mathbb{F}^{mn}$ and using the usual Euclidean norm. It can be shown that it really doesn't matter which norm you use in terms of estimates because they are all equivalent. This is discussed in Problem 25 below for those who have had a legitimate calculus course, not just the usual undergraduate offering.

14.4 The Condition Number*

Let A be an $m\times n$ matrix and consider the problem $A\mathbf{x}=\mathbf{b}$ 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 $\mathbf{b}$ is changed a little bit? This is clearly an interesting question because you often do not know A and $\mathbf{b}$ exactly. If a small change in these quantities results in a large change in the solution $\mathbf{x}$, then it seems clear this would be undesirable. In what follows $||\cdot||$ when applied to a matrix will always refer to the operator norm.

Lemma 14.4.1 *Let A,B be $m\times n$ matrices. Then for $||\cdot||$ denoting the operator norm,*

$$||AB||\leq ||A||\,||B||.$$

Proof: This follows from the definition. Letting $||\mathbf{x}||\leq 1$, it follows from the definition of the operator norm that

$$||AB\mathbf{x}||\leq ||A||\,||B\mathbf{x}||\leq ||A||\,||B||\,||\mathbf{x}||\leq ||A||\,||B||$$

and so

$$||AB||\equiv \sup_{||\mathbf{x}||\leq 1}||AB\mathbf{x}||\leq ||A||\,||B||. \blacksquare$$

Lemma 14.4.2 *Let A,B be $m\times n$ matrices such that A^{-1} exists, and suppose $||B||<1/\left|\left|A^{-1}\right|\right|$. Then $(A+B)^{-1}$ exists and*

$$\left|\left|(A+B)^{-1}\right|\right|\leq \left|\left|A^{-1}\right|\right|\left|\frac{1}{1-||A^{-1}B||}\right|.$$

The above formula makes sense because $\left|\left|A^{-1}B\right|\right|<1$.

Proof: By Lemma 14.4.1,

$$\left\|A^{-1}B\right\|\leq \left\|A^{-1}\right\|\|B\|<\left\|A^{-1}\right\|\frac{1}{\|A^{-1}\|}=1$$

Suppose $(A+B)\mathbf{x}=0$. Then $0=A\left(I+A^{-1}B\right)\mathbf{x}$ and so since A is one to one,

$$\left(I+A^{-1}B\right)\mathbf{x}=0.$$

Therefore,

$$\begin{aligned} 0 &= \left\|\left(I+A^{-1}B\right)\mathbf{x}\right\|\geq \|\mathbf{x}\|-\left\|A^{-1}B\mathbf{x}\right\| \\ &\geq \|\mathbf{x}\|-\left\|A^{-1}B\right\|\|\mathbf{x}\|=\left(1-\left\|A^{-1}B\right\|\right)\|\mathbf{x}\|>0 \end{aligned}$$

a contradiction. This also shows $\left(I+A^{-1}B\right)$ is one to one. Therefore, both $(A+B)^{-1}$ and $\left(I+A^{-1}B\right)^{-1}$ exist. Hence

$$(A+B)^{-1}=\left(A\left(I+A^{-1}B\right)\right)^{-1}=\left(I+A^{-1}B\right)^{-1}A^{-1}$$

Now if

$$\mathbf{x}=\left(I+A^{-1}B\right)^{-1}\mathbf{y}$$

for $\|\mathbf{y}\| \leq 1$, then

$$\left(I + A^{-1}B\right)\mathbf{x} = \mathbf{y}$$

and so

$$\|\mathbf{x}\|\left(1 - \left\|A^{-1}B\right\|\right) \leq \left\|\mathbf{x} + A^{-1}B\mathbf{x}\right\| \leq \|\mathbf{y}\| = 1$$

and so

$$||\mathbf{x}|| = \left\|\left(I + A^{-1}B\right)^{-1}\mathbf{y}\right\| \leq \frac{1}{1 - \|A^{-1}B\|}$$

Since $||\mathbf{y}|| \leq 1$ is arbitrary, this shows

$$\left\|\left(I + A^{-1}B\right)^{-1}\right\| \leq \frac{1}{1 - \|A^{-1}B\|}$$

Therefore,

$$\begin{aligned} \left\|(A+B)^{-1}\right\| &= \left\|\left(I + A^{-1}B\right)^{-1}A^{-1}\right\| \\ &\leq \left\|A^{-1}\right\|\left\|\left(I + A^{-1}B\right)^{-1}\right\| \leq \left\|A^{-1}\right\|\frac{1}{1 - \|A^{-1}B\|} \ \blacksquare \end{aligned}$$

Proposition 14.4.3 *Suppose A is invertible, $\mathbf{b} \neq 0$, $A\mathbf{x} = \mathbf{b}$, and $A_1\mathbf{x}_1 = \mathbf{b}_1$ where*

$$\|A - A_1\| < 1/\left\|A^{-1}\right\|$$

Then

$$\frac{\|\mathbf{x}_1 - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{1}{(1 - \|A^{-1}(A_1 - A)\|)}\|A\|\left\|A^{-1}\right\|\left(\frac{\|A_1 - A\|}{\|A\|} + \frac{\|\mathbf{b} - \mathbf{b}_1\|}{\|\mathbf{b}\|}\right). \tag{14.11}$$

Proof: It follows from the assumptions that

$$A\mathbf{x} - A_1\mathbf{x} + A_1\mathbf{x} - A_1\mathbf{x}_1 = \mathbf{b} - \mathbf{b}_1.$$

Hence

$$A_1(\mathbf{x} - \mathbf{x}_1) = (A_1 - A)\mathbf{x} + \mathbf{b} - \mathbf{b}_1.$$

Now $A_1 = (A + (A_1 - A))$ and so by the above lemma, A_1^{-1} exists and so

$$(\mathbf{x} - \mathbf{x}_1) = A_1^{-1}(A_1 - A)\mathbf{x} + A_1^{-1}(\mathbf{b} - \mathbf{b}_1)$$

$$= (A + (A_1 - A))^{-1}(A_1 - A)\mathbf{x} + (A + (A_1 - A))^{-1}(\mathbf{b} - \mathbf{b}_1).$$

By the estimate in Lemma 14.4.2,

$$\|\mathbf{x} - \mathbf{x}_1\| \leq \frac{\left\|A^{-1}\right\|}{1 - \|A^{-1}(A_1 - A)\|}(\|A_1 - A\|\|\mathbf{x}\| + \|\mathbf{b} - \mathbf{b}_1\|).$$

Dividing by $\|\mathbf{x}\|$,

$$\frac{\|\mathbf{x} - \mathbf{x}_1\|}{\|\mathbf{x}\|} \leq \frac{\left\|A^{-1}\right\|}{1 - \|A^{-1}(A_1 - A)\|}\left(\|A_1 - A\| + \frac{\|\mathbf{b} - \mathbf{b}_1\|}{\|\mathbf{x}\|}\right) \tag{14.12}$$

Now $\mathbf{b} = A\mathbf{x} = A\left(A^{-1}\mathbf{b}\right)$ and so $\|\mathbf{b}\| \leq \|A\| \left\|A^{-1}\mathbf{b}\right\|$ and so

$$\|\mathbf{x}\| = \left\|A^{-1}\mathbf{b}\right\| \geq \|\mathbf{b}\| / \|A\|.$$

Therefore, from 14.12,

$$\begin{aligned}\frac{\|\mathbf{x}-\mathbf{x}_1\|}{\|\mathbf{x}\|} &\leq \frac{\left\|A^{-1}\right\|}{1-\|A^{-1}(A_1-A)\|}\left(\frac{\|A\|\,\|A_1-A\|}{\|A\|}+\frac{\|A\|\,\|\mathbf{b}-\mathbf{b}_1\|}{\|\mathbf{b}\|}\right) \\ &\leq \frac{\left\|A^{-1}\right\|\,\|A\|}{1-\|A^{-1}(A_1-A)\|}\left(\frac{\|A_1-A\|}{\|A\|}+\frac{\|\mathbf{b}-\mathbf{b}_1\|}{\|\mathbf{b}\|}\right) \quad \blacksquare\end{aligned}$$

This shows that the number, $\left\|A^{-1}\right\| \|A\|$, controls how sensitive the relative change in the solution of $A\mathbf{x} = \mathbf{b}$ is to small changes in A and $\mathbf{b}$. This number is called the condition number. It is bad when it is large because a small relative change in $\mathbf{b}$, for example, could yield a large relative change in $\mathbf{x}$.

14.5 Exercises

1. Solve the system
$$\begin{pmatrix} 4 & 1 & 1 \\ 1 & 5 & 2 \\ 0 & 2 & 6 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$$
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
$$\begin{pmatrix} 5 & 1 & 1 \\ 1 & 7 & 2 \\ 0 & 2 & 4 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

4. Solve the system
$$\begin{pmatrix} 7 & 1 & 0 \\ 1 & 5 & 2 \\ 0 & 2 & 6 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

5. Solve the system
$$\begin{pmatrix} 5 & 0 & 1 \\ 1 & 7 & 1 \\ 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 7 \\ 3 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

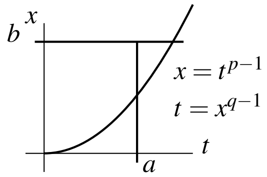
6. Solve the system
$$\begin{pmatrix} 5 & 0 & 1 \\ 1 & 7 & 1 \\ 0 & 2 & 9 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$
using the Gauss Seidel method and the Jacobi method. Check your answer by also solving it using row operations.

7. 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 using either of these methods for finding eigenvectors for a given eigenvalue?

8. Verify that
$$\|\mathbf{x}\|_\infty \equiv \max\{|x_i|, i = 1, \cdots, n : \mathbf{x} = (x_1, \cdots, x_n)\}$$
is a norm. Next verify that
$$\|\mathbf{x}\|_1 \equiv \sum_{i=1}^n |x_i|, \ \mathbf{x} = (x_1, \cdots, x_n)$$
is also a norm on $\mathbb{F}^n$.

9. Let A be an $n \times n$ matrix. Denote by $\|A\|_2$ the operator norm taken with respect to the usual norm on $\mathbb{F}^n$. Show that
$$\|A\|_2 = \sigma_1$$
where σ_1 is the largest singular value. Next explain why $\left\|A^{-1}\right\|_2 = 1/\sigma_n$ where σ_n is the smallest singular value of A. Explain why the condition number reduces to σ_1/σ_n if the operator norm is defined in terms of the usual norm, $|\mathbf{x}| = \left(\sum_{j=1}^n |x_j|^2\right)^{1/2}$.

10. Let $p, q > 1$ and $1/p + 1/q = 1$. Consider the following picture.



Using elementary calculus, verify that for $a, b > 0$,
$$\frac{a^p}{p} + \frac{b^q}{q} \geq ab.$$

11. ↑For $p > 1$, the p norm on $\mathbb{F}^n$ is defined by

$$\|\mathbf{x}\|_p \equiv \left(\sum_{k=1}^n |x_k|^p\right)^{1/p}$$

In fact, this is a norm and this will be shown in this and the next problem. Using the above problem in the context stated there where $p, q > 1$ and $1/p + 1/q = 1$, verify Holder's inequality

$$\sum_{k=1}^n |x_k| |y_k| \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q$$

Hint: You ought to consider the following.

$$\sum_{k=1}^n \frac{|x_k|}{\|\mathbf{x}\|_p} \frac{|y_k|}{\|\mathbf{y}\|_q}$$

Now use the result of the above problem.

12. ↑ Now for $p > 1$, verify that $\|\mathbf{x}+\mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$. Then verify the other axioms of a norm. This will give an infinite collection of norms for $\mathbb{F}^n$. **Hint:** You might do the following.

$$\begin{aligned} \|\mathbf{x}+\mathbf{y}\|_p^p &\leq \sum_{k=1}^n |x_k + y_k|^{p-1} (|x_k| + |y_k|) \\ &= \sum_{k=1}^n |x_k + y_k|^{p-1} |x_k| + \sum_{k=1}^n |x_k + y_k|^{p-1} |y_k| \end{aligned}$$

Now explain why $p - 1 = p/q$ and use the Holder inequality.

13. This problem will reveal the best kept **secret** in undergraduate mathematics, the definition of the derivative of a function of n variables. Let $\|\cdot\|$ be a norm on $\mathbb{F}^n$ and also denote by $\|\cdot\|$ a norm on $\mathbb{F}^m$. If you like, just use the standard norm on both $\mathbb{F}^n$ and $\mathbb{F}^m$. It can be shown that this doesn't matter at all (See Problem 25 on 429.) but to avoid possible confusion, you can be specific about the norm. A set $U \subseteq \mathbb{F}^n$ is said to be open if for every $\mathbf{x} \in U$, there exists some $r_{\mathbf{x}} > 0$ such that $B(\mathbf{x}, r_{\mathbf{x}}) \subseteq U$ where

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

This just says that if U contains a point $\mathbf{x}$ then it contains all the other points sufficiently near to $\mathbf{x}$. Let $\mathbf{f} : U \mapsto \mathbb{F}^m$ be a function defined on U having values in $\mathbb{F}^m$. Then $\mathbf{f}$ is differentiable at $\mathbf{x} \in U$ means that there exists an $m \times n$ matrix A such that for every $\varepsilon > 0$, there exists a $\delta > 0$ such that whenever $0 < \|\mathbf{v}\| < \delta$, it follows that

$$\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x}) - A\mathbf{v}\|}{\|\mathbf{v}\|} < \varepsilon$$

Stated more simply,

$$\lim_{\|\mathbf{v}\| \to 0} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{v}) - \mathbf{f}(\mathbf{x}) - A\mathbf{v}\|}{\|\mathbf{v}\|} = 0$$

Show that A is unique and verify that the i^{th} column of A is

$$\frac{\partial \mathbf{f}}{\partial x_i}(\mathbf{x})$$

so in particular, all partial derivatives exist. This unique $m \times n$ matrix is called the derivative of $\mathbf{f}$. It is written as $D\mathbf{f}(\mathbf{x}) = A$.

Chapter 15

Numerical Methods, Eigenvalues

15.1 The Power Method For Eigenvalues

This chapter presents some simple ways to find eigenvalues and eigenvectors. It is only an introduction to this important subject. However, I hope to convey some of the ideas which are used. As indicated earlier, the eigenvalue eigenvector problem is extremely difficult. Consider for example what happens if you find an eigenvalue approximately. Then you can't find an approximate eigenvector by the straight forward approach because $A-\lambda I$ is invertible whenever λ is not exactly equal to an eigenvalue.

Of course computer algebra systems allow you to ask for eigenvalues and eigenvectors and get the answer with no effort. This chapter is going to describe some of the ideas which lead to software which is able to give such answers.

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 a real eigenvalue of algebraic multiplicity 1, λ_n which has the property that $|\lambda_k| < |\lambda_n|$ for all $k \neq n$. 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

$$\mathbf{u}_m = A^m \mathbf{u}_1 = \sum_{k=1}^{n-1} c_k \lambda_k^m \mathbf{x}_k + \lambda_n^m c_n \mathbf{x}_n . \tag{15.1}$$

For large m the last term, $\lambda_n^m c_n \mathbf{x}_n$, determines quite well the direction of the vector on the right. This is because $|\lambda_n|$ is larger than $|\lambda_k|$ for $k < n$ 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 15.1. Indeed it amounts to nothing more than changing the units of length. Also note that from this scaling procedure, the absolute value of the largest element of $\mathbf{u}_k$ is always equal to 1. Therefore, for large m,

$$\mathbf{u}_m = \frac{\lambda_n^m c_n \mathbf{x}_n}{S_2 S_3 \cdots S_m} + (\text{relatively insignificant term}).$$

Therefore, the entry of $A\mathbf{u}_m$ which has the largest absolute value is essentially equal to the entry having largest absolute value of

$$A\left(\frac{\lambda_n^m c_n \mathbf{x}_n}{S_2 S_3 \cdots S_m}\right) = \frac{\lambda_n^{m+1} c_n \mathbf{x}_n}{S_2 S_3 \cdots S_m} \approxeq \lambda_n \mathbf{u}_m$$

and so for large m, it must be the case that $\lambda_n \approxeq S_{m+1}$. Here $\approxeq$ means "approximately equal". 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.

In finding an initial vector, it is clear that if you start with a vector which isn't too far from an eigenvector, the process will work faster. Also, the computer is able to raise the matrix to a power quite easily. You might start with $A^p\mathbf{x}$ for large p. As explained above, this will point in roughly the right direction. Then normalize it by dividing by the largest entry and use the resulting vector as your initial approximation. This ought to be close to an eigenvector and so the process would be expected to converge rapidly for this as an initial choice.

Example 15.1.1 *Find the largest eigenvalue of* $A = \begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}$.

▶

I will use the above suggestion.

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix}^{15} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1.0271 \times 10^{16} \\ -5.1357 \times 10^{15} \\ 4.7018 \times 10^{11} \end{pmatrix}$$

Now divide by the largest entry to get the initial approximation for an eigenvector

$$\begin{pmatrix} 1.0271\times 10^{16} \\ -5.1357\times 10^{15} \\ 4.7018\times 10^{11} \end{pmatrix} \frac{1}{1.0271\times 10^{16}} = \begin{pmatrix} 1.0 \\ -0.50002 \\ 4.5777\times 10^{-5} \end{pmatrix} = \mathbf{u}_1$$

The power method will now be applied to find the largest eigenvalue for the above matrix beginning with this vector.

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -0.50002 \\ 4.5777\times 10^{-5} \end{pmatrix} = \begin{pmatrix} 12.001 \\ -6.0003 \\ -2.5733\times 10^{-4} \end{pmatrix}.$$

Scaling this vector by dividing by the largest entry gives

$$\begin{pmatrix} 12.001 \\ -6.0003 \\ -2.5733\times 10^{-4} \end{pmatrix} \frac{1}{12.001} = \begin{pmatrix} 1.0 \\ -0.49998 \\ -2.1442\times 10^{-5} \end{pmatrix} \equiv \mathbf{u}_2$$

Now lets do it again.

$$\begin{pmatrix} 5 & -14 & 11 \\ -4 & 4 & -4 \\ 3 & 6 & -3 \end{pmatrix} \begin{pmatrix} 1.0 \\ -0.49998 \\ -2.1442\times 10^{-5} \end{pmatrix} = \begin{pmatrix} 11.999 \\ -5.9998 \\ 1.8433\times 10^{-4} \end{pmatrix}$$

The new scaling factor is very close to the one just encountered. Therefore, it seems this is a good place to stop. The eigenvalue is approximately 11.999 and the eigenvector is close to the one obtained above. How well does it work? With the above equation, consider

$$11.999 \begin{pmatrix} 1.0 \\ -0.49998 \\ -2.1442\times 10^{-5} \end{pmatrix} = \begin{pmatrix} 11.999 \\ -5.9993 \\ -2.5728\times 10^{-4} \end{pmatrix}$$

These are clearly very close so this is a good approximation. In fact, the exact eigenvalue is 12 and an eigenvector is

$$\begin{pmatrix} 1.0 \\ -0.5 \\ 0 \end{pmatrix}$$

15.2 The Shifted Inverse Power Method

This method can find various eigenvalues and eigenvectors. It is a significant generalization of the above simple procedure and yields very good results. The situation is this: You have a number α which is close to λ, some eigenvalue of an $n\times n$ matrix A. You don't know λ but you know that α is closer to λ than to any other eigenvalue. Your problem is to find both λ and an eigenvector which goes with λ. Another way to look at this is to start with α and seek the eigenvalue λ, which is closest to α along with an eigenvector associated with λ. If α is an eigenvalue of A, then you have what you want. Therefore, we will always assume α is not an eigenvalue of A and so $(A-\alpha I)^{-1}$ exists. When using this method it is nice to choose α fairly close to an eigenvalue. Otherwise, the method will converge slowly. In order to get some idea where to start, you could use Gerschgorin's theorem to get a rough idea where to look. The method is based on the following lemma.

Lemma 15.2.1 *Let $\{\lambda_k\}_{k=1}^n$ be the eigenvalues of A, α not an eigenvalue. Then $\mathbf{x}_k$ is an eigenvector of A for the eigenvalue λ_k, if and only if $\mathbf{x}_k$ is an eigenvector for $(A-\alpha I)^{-1}$ corresponding to the eigenvalue $\frac{1}{\lambda_k-\alpha}$.*

Proof: Let λ_k and $\mathbf{x}_k$ be as described in the statement of the lemma. Then

$$(A-\alpha I)\mathbf{x}_k = (\lambda_k-\alpha)\mathbf{x}_k$$

if and only if

$$\frac{1}{\lambda_k-\alpha}\mathbf{x}_k = (A-\alpha I)^{-1}\mathbf{x}_k. \blacksquare$$

In explaining why the method works, we will assume A is nondefective. **This is not necessary!** One can use Gelfand's theorem on the spectral radius which is presented in [13] and invariance of $(A-\alpha I)^{-1}$ on generalized eigenspaces to prove more general results. It suffices to assume that the eigenspace for λ_k has dimension equal to the multiplicity of the eigenvalue λ_k but even this is not necessary to obtain convergence of the method. This method is better than might be supposed from the following explanation.

Pick $\mathbf{u}_1$, an initial vector and let $A\mathbf{x}_k = \lambda_k\mathbf{x}_k$, where $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ is a basis of eigenvectors which exists from the assumption that A is nondefective. Assume α is closer to λ_n than to any other eigenvalue. Since A is nondefective, there exist constants, a_k such that

$$\mathbf{u}_1 = \sum_{k=1}^n a_k\mathbf{x}_k.$$

Possibly λ_n is a repeated eigenvalue. Then combining the terms in the sum which involve eigenvectors for λ_n, a simpler description of $\mathbf{u}_1$ is

$$\mathbf{u}_1 = \sum_{j=1}^m a_j\mathbf{x}_j + \mathbf{y}$$

where $\mathbf{y}$ is an eigenvector for λ_n which is assumed not equal to $\mathbf{0}$. (If you are unlucky in your choice for $\mathbf{u}_1$, this might not happen and things won't work.) Now the iteration procedure is defined as

$$\mathbf{u}_{k+1} \equiv \frac{(A-\alpha I)^{-1}\mathbf{u}_k}{S_{k+1}}$$

where S_{k+1} is the element of $(A-\alpha I)^{-1}\mathbf{u}_k$ which has largest absolute value. From Lemma

15.2.1,

$$\begin{aligned}\mathbf{u}_{k+1} &= \frac{\sum_{j=1}^{m} a_j \left(\frac{1}{\lambda_j-\alpha}\right)^k \mathbf{x}_j + \left(\frac{1}{\lambda_n-\alpha}\right)^k \mathbf{y}}{S_2\cdots S_{k+1}} \\ &= \frac{\left(\frac{1}{\lambda_n-\alpha}\right)^k}{S_2\cdots S_{k+1}} \left(\sum_{j=1}^{m} a_j \left(\frac{\lambda_n-\alpha}{\lambda_j-\alpha}\right)^k \mathbf{x}_j + \mathbf{y}\right).\end{aligned}$$

Now it is being assumed that λ_n is the eigenvalue which is closest to α and so for large k, the term,

$$\sum_{j=1}^{m} a_j \left(\frac{\lambda_n-\alpha}{\lambda_j-\alpha}\right)^k \mathbf{x}_j \equiv \mathbf{E}_k$$

is very small, while for every $k \geq 1$, $\mathbf{u}_k$ is a moderate sized vector because every entry has absolute value less than or equal to 1. Thus

$$\mathbf{u}_{k+1} = \frac{\left(\frac{1}{\lambda_n-\alpha}\right)^k}{S_2\cdots S_{k+1}} (\mathbf{E}_k + \mathbf{y}) \equiv C_k (\mathbf{E}_k + \mathbf{y})$$

where $\mathbf{E}_k \to \mathbf{0}$, $\mathbf{y}$ is some eigenvector for λ_n, and C_k is of moderate size, remaining bounded as $k \to \infty$ due to the fact that from the construction, $\mathbf{u}_{k+1}$ has all entries no larger than 1. Therefore, for large k, and letting $\approxeq$ denote "approximately equal",

$$\mathbf{u}_{k+1} - C_k\mathbf{y} = C_k\mathbf{E}_k \approxeq \mathbf{0}$$

and multiplying by $(A-\alpha I)^{-1}$ yields

$$\begin{aligned}(A-\alpha I)^{-1}\mathbf{u}_{k+1} - (A-\alpha I)^{-1} C_k\mathbf{y} &= (A-\alpha I)^{-1}\mathbf{u}_{k+1} - C_k \left(\frac{1}{\lambda_n-\alpha}\right)\mathbf{y} \\ &\approxeq (A-\alpha I)^{-1}\mathbf{u}_{k+1} - \left(\frac{1}{\lambda_n-\alpha}\right)\mathbf{u}_{k+1} \approxeq \mathbf{0}.\end{aligned}$$

Therefore, for large k, $\mathbf{u}_k$ is approximately equal to an eigenvector of $(A-\alpha I)^{-1}$. Therefore,

$$(A-\alpha I)^{-1}\mathbf{u}_k \approxeq \frac{1}{\lambda_n-\alpha}\mathbf{u}_k$$

and so you could take the dot product of both sides with $\mathbf{u}_k$ and approximate λ_n by solving the following for λ_n.

$$\frac{(A-\alpha I)^{-1}\mathbf{u}_k \cdot \mathbf{u}_k}{|\mathbf{u}_k|^2} = \frac{1}{\lambda_n-\alpha}$$

How else can you find the eigenvalue from this? Suppose $\mathbf{u}_k = (w_1, \cdots, w_n)^T$ and from the construction $|w_i| \leq 1$ and $w_k = 1$ for some k. Then

$$S_{k+1}\mathbf{u}_{k+1} = (A-\alpha I)^{-1}\mathbf{u}_k \approxeq (A-\alpha I)^{-1}(C_{k-1}\mathbf{y}) = \frac{1}{\lambda_n-\alpha}(C_{k-1}\mathbf{y}) \approxeq \frac{1}{\lambda_n-\alpha}\mathbf{u}_k.$$

Hence the entry of $(A-\alpha I)^{-1}\mathbf{u}_k$ which has largest absolute value is approximately $\frac{1}{\lambda_n-\alpha}$ and so it is likely that you can estimate λ_n using the formula

$$S_{k+1}=\frac{1}{\lambda_n-\alpha}.$$

Of course this would fail if $(A-\alpha I)^{-1}\mathbf{u}_k$ had consistently more than one entry having equal absolute value, but this is unlikely.

Here is how you use the shifted inverse power method to find the eigenvalue and eigenvector closest to α.

1. Find $(A-\alpha I)^{-1}$.

2. Pick $\mathbf{u}_1$. It is important that $\mathbf{u}_1=\sum_{j=1}^{m}a_j\mathbf{x}_j+\mathbf{y}$ where $\mathbf{y}$ is an eigenvector which goes with the eigenvalue closest to α and the sum is in an "invariant subspace corresponding to the other eigenvalues". Of course you have no way of knowing whether this is so but it typically is so. If things don't work out, just start with a different $\mathbf{u}_1$. You were phenomenally 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 element of $(A-\alpha I)^{-1}\mathbf{u}_k$ which has largest absolute value.

4. When the scaling factors, S_{k+1} 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.

Also note that this is just the power method applied to $(A-\lambda I)^{-1}$. The eigenvalue you want is the one which makes $\frac{1}{\lambda-\alpha}$ as large as possible for all $\lambda\in\sigma(A)$. This is because making $\lambda-\alpha$ small is the same as making $(\lambda-\alpha)^{-1}$ large.

15.3 Automation With MATLAB

You can do the above example and other examples using MATLAB. Here are some commands which will do this. It is done here for a 3 × 3 matrix but you adapt for any size.

```
a=[5 -8 6;1 0 0;0 1 0]; b=i; F=inv(a-b*eye(3));
S=1; u=[1;1;1]; d=1; k=1;
while d > .00001 & k <1000
w=F*u; [M,I]=max(abs(w)); T=w(I); u=w/T;
d=abs(T-S); S=T; k=k+1;
end
u
b+1/T
k
a*u-(b+1/T)*u
```

Note how the "while loop" is limited to 1000 iterations. That way it won't go on forever if there is something wrong. This asks for the eigenvalue closest to $b = i$. When MATLAB stalls, to get it to quit, you type control c. The last line checks the answer and the line with k tells the number of iterations used. Also, the funny notation [M,I]=max(abs(w)); T=w(I); gets it to pick out the entry which has largest absolute value w(I) and keep that entry unchanged. The above iteration finds the eigenvalue closest to i along with the corresponding eigenvector. When the procedure does not work well for b real, you might imagine that there are complex eigenvalues and so, since the above procedure is going to give you real approximations, it can't find the complex eigenvalues. Thus you should take b to be complex as done above.

If you have MATLAB work the above iteration, you get the following for the eigenvector eigenvalue and number of iterations, and error .

$$\begin{pmatrix} 1 \\ .5-.5i \\ -.5i \end{pmatrix},\ 1+i,\ k=18,\ 10^{-5}\begin{pmatrix} 0 \\ -0.1321+0.1862i \\ -0.1325+0.1863i \end{pmatrix}$$

In fact, this eigenvector is exactly right as is the eigenvalue $1+i$.

Thus this method will find eigenvalues real or complex along with an eigenvector associated with the eigenvalue. Note that the characteristic polynomial of the above matrix is $\lambda^3 - 5\lambda^2 + 8\lambda - 6$ and the above finds a complex root to this polynomial. More generally, if you have a polynomial $\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0$, a matrix which has this as its characteristic polynomial is called a companion matrix and you can show a matrix which works for this polynomial is of the form

$$\begin{pmatrix} -a_{n-1} & -a_{n-2} & \cdots & a_0 \\ 1 & 0 & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{pmatrix}$$

Thus this method is capable of finding roots to a polynomial equation which are close to a given complex number. Of course there is a problem with determining which number you should pick. A way to determine this will be discussed later. It involves something called the QR algorithm.

Example 15.3.1 *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.*

We use the algorithm described above.

```
a=[5 -14 11;-4 4 -4;3 6 -3]; b=-7; F=inv(a-b*eye(3));
S=1; u=[1;1;1]; d=1; k=1;
while d > .0001 & k < 1000
w=F*u; [M,I]=max(abs(w)); T=w(I); u=w/T;
d=abs(T-S); S=T; k=k+1;
end
u
b+1/T
a*u-(b+1/T)*u
```

This yields

$$\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, -6$$

for the eigenvector and eigenvalue. In fact, this is exactly correct.

Example 15.3.2 *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

$$\boxed{\textbf{\underline{Eigenvectors}\ \underline{are}\ \underline{never}\ \underline{equal}\ \underline{to}\ \underline{zero}!}}$$

However, there exists such an eigenvector and you can find it using the shifted inverse power method. Pick $\alpha = -.855$ in the above algorithm. Then entering the matrix and running the algorithm yields the eigenvector and eigenvalue

$$\begin{pmatrix} -.0111 \\ -.2776 \\ .2470 \end{pmatrix}, \ -.8569$$

In fact the error is on the order of 10^{-14}.

There is an easy to use trick which will eliminate some of the fuss and bother in using the shifted inverse power method. If you have

$$(A - \alpha I)^{-1} \mathbf{x} = \mu \mathbf{x}$$

then multiplying through by $(A - \alpha I)$, one finds that $\mathbf{x}$ will be an eigenvector for A with eigenvalue $\alpha + \mu^{-1}$. Hence you could simply take $(A - \alpha I)^{-1}$ to a high power and multiply by a vector to get a vector which points in the direction of an eigenvalue of A. Then divide by the largest entry and identify the eigenvalue directly by multiplying the eigenvector by A. This is illustrated in the next example.

Example 15.3.3 *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 we 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

$$(A - \alpha I)^{-1} = \begin{pmatrix} -25.357143 & -33.928571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214286 & 30.357143 & -45.0 \end{pmatrix}.$$

Then

$$\begin{pmatrix} -25.357143 & -33.928571 & 50.0 \\ 12.5 & 17.5 & -25.0 \\ 23.214286 & 30.357143 & -45.0 \end{pmatrix}^{17} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} -4.9432 \times 10^{28} \\ 2.4312 \times 10^{28} \\ 4.4928 \times 10^{28} \end{pmatrix}$$

The initial approximation for an eigenvector will then be the above divided by its largest entry.

$$\begin{pmatrix} -4.9432\times 10^{28} \\ 2.4312\times 10^{28} \\ 4.4928\times 10^{28} \end{pmatrix} \frac{1}{-4.9432\times 10^{28}} = \begin{pmatrix} 1.0 \\ -0.49183 \\ -0.90888 \end{pmatrix}$$

How close is this to being an eigenvector?

$$\begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1.0 \\ -0.49183 \\ -0.90888 \end{pmatrix} = \begin{pmatrix} -1.2185 \\ 0.59929 \\ 1.1075 \end{pmatrix}$$

$$-1.2185 \begin{pmatrix} 1.0 \\ -0.49183 \\ -0.90888 \end{pmatrix} = \begin{pmatrix} -1.2185 \\ 0.59929 \\ 1.1075 \end{pmatrix}$$

For all practical purposes, this has found the eigenvector for the eigenvalue -1.2185.

Next we shall find the eigenvector and a more precise value for the eigenvalue near $-.4$. In this case,

$$(A-\alpha I)^{-1} = \begin{pmatrix} 8.0645161\times 10^{-2} & -9.2741935 & 6.4516129 \\ -.40322581 & 11.370968 & -7.2580645 \\ .40322581 & 3.6290323 & -2.7419355 \end{pmatrix}.$$

The first approximation to an eigenvector can be obtained as before.

$$\begin{pmatrix} 8.0645161\times 10^{-2} & -9.2741935 & 6.4516129 \\ -.40322581 & 11.370968 & -7.2580645 \\ .40322581 & 3.6290323 & -2.7419355 \end{pmatrix}^{17} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} -1.8535\times 10^{16} \\ 2.3724\times 10^{16} \\ 6.2874\times 10^{15} \end{pmatrix}$$

The first choice for an approximate eigenvector is

$$\begin{pmatrix} -1.8535\times 10^{16} \\ 2.3724\times 10^{16} \\ 6.2874\times 10^{15} \end{pmatrix} \frac{1}{2.3724\times 10^{16}} = \begin{pmatrix} -0.78128 \\ 1.0 \\ 0.26502 \end{pmatrix}$$

Lets see how well this works as an eigenvector.

$$\begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} -0.78128 \\ 1.0 \\ 0.26502 \end{pmatrix} = \begin{pmatrix} 0.2325 \\ -0.29754 \\ -0.07882 \end{pmatrix}$$

$$(-0.29754)\begin{pmatrix} -0.78128 \\ 1.0 \\ 0.26502 \end{pmatrix} = \begin{pmatrix} 0.23246 \\ -0.29754 \\ -7.8854 \times 10^{-2} \end{pmatrix}$$

Thus this works as an eigenvector with the eigenvalue (-0.29754).

Next we 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 to avoid giving the impression that you always need to start with the vector $(1,1,1)^T$, let $\mathbf{u}_1 = (1,2,3)^T$. I will use the same shortcut to get this eigenvector as in the above case.

$$\begin{pmatrix} 29.2 & 16.8 & 23.2 \\ 19.2 & 10.8 & 15.2 \\ 28.0 & 16.0 & 22.0 \end{pmatrix}^{16} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 1.0987 \times 10^{29} \\ 7.1868 \times 10^{28} \\ 1.0482 \times 10^{29} \end{pmatrix}$$

Then dividing by the largest entry, a good guess for the eigenvector is

$$\begin{pmatrix} 1.0 \\ 0.65412 \\ 0.95404 \end{pmatrix}$$

To see if more iteration would be needed, check this.

$$\begin{pmatrix} 2 & 1 & 3 \\ 2 & 1 & 1 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1.0 \\ 0.65412 \\ 0.95404 \end{pmatrix} = \begin{pmatrix} 5.5162 \\ 3.6082 \\ 5.2623 \end{pmatrix}$$

and

$$5.5162 \begin{pmatrix} 1.0 \\ 0.65412 \\ 0.95404 \end{pmatrix} = \begin{pmatrix} 5.5162 \\ 3.6083 \\ 5.2627 \end{pmatrix}$$

Thus this is essentially an eigenvector with eigenvalue equal to 5.5162.

15.4 The Rayleigh Quotient

There are many specialized results concerning the eigenvalues and eigenvectors for Hermitian matrices. 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 $\mathbf{x} \in \mathbb{F}^n$,

$$|\mathbf{x}|^2 = \mathbf{x}^*\mathbf{x} = \sum_{j=1}^{n} |x_j|^2 .$$

The following corollary gives the theoretical foundation for the spectral theory of Hermitian matrices. This is a corollary of a theorem which is proved Corollary 13.2.14 and Theorem 13.2.14 on Page 310.

Corollary 15.4.1 *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,

$$|\mathbf{x}|^2 = \sum_{i=1}^n \overline{a}_i \mathbf{x}_i^* \sum_{j=1}^n a_j \mathbf{x}_j = \sum_{ij} \overline{a}_i a_j \mathbf{x}_i^* \mathbf{x}_j = \sum_{ij} \overline{a}_i a_j \delta_{ij} = \sum_{i=1}^n |a_i|^2.$$

Therefore,

$$\frac{\mathbf{x}^* A \mathbf{x}}{|\mathbf{x}|^2} = \frac{\left(\sum_{i=1}^n \overline{a}_i \mathbf{x}_i^*\right)\left(\sum_{j=1}^n a_j \lambda_j \mathbf{x}_j\right)}{\sum_{i=1}^n |a_i|^2}$$

$$= \frac{\sum_{ij} \overline{a}_i a_j \lambda_j \mathbf{x}_i^* \mathbf{x}_j}{\sum_{i=1}^n |a_i|^2} = \frac{\sum_{ij} \overline{a}_i a_j \lambda_j \delta_{ij}}{\sum_{i=1}^n |a_i|^2} = \frac{\sum_{i=1}^n |a_i|^2 \lambda_i}{\sum_{i=1}^n |a_i|^2} \in [\lambda_1, \lambda_n].$$

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 15.4.2 *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

$$|\lambda_q - q| \leq \frac{|A\mathbf{x} - q\mathbf{x}|}{|\mathbf{x}|}. \tag{15.2}$$

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.

$$\begin{aligned}
|A\mathbf{x} - q\mathbf{x}|^2 &= (A\mathbf{x} - q\mathbf{x})^* (A\mathbf{x} - q\mathbf{x}) \\
&= \left(\sum_{k=1}^{n} a_k \lambda_k \mathbf{x}_k - q a_k \mathbf{x}_k \right)^* \left(\sum_{k=1}^{n} a_k \lambda_k \mathbf{x}_k - q a_k \mathbf{x}_k \right) \\
&= \left(\sum_{j=1}^{n} (\lambda_j - q) \overline{a}_j \mathbf{x}_j^* \right) \left(\sum_{k=1}^{n} (\lambda_k - q) a_k \mathbf{x}_k \right) \\
&= \sum_{j,k} (\lambda_j - q) \overline{a}_j (\lambda_k - q) a_k \mathbf{x}_j^* \mathbf{x}_k \\
&= \sum_{k=1}^{n} |a_k|^2 (\lambda_k - q)^2
\end{aligned}$$

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 15.2. ■

Example 15.4.3 *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

$$\begin{aligned}\left|\lambda_q-\frac{19}{3}\right| &\leq \frac{\left|\begin{pmatrix}1&2&3\\2&2&1\\3&1&4\end{pmatrix}\begin{pmatrix}1\\1\\1\end{pmatrix}-\frac{19}{3}\begin{pmatrix}1\\1\\1\end{pmatrix}\right|}{\sqrt{3}}\\ &= \frac{1}{\sqrt{3}}\begin{pmatrix}-\frac{1}{3}\\-\frac{4}{3}\\\frac{5}{3}\end{pmatrix}\\ &= \frac{\sqrt{\frac{1}{9}+\left(\frac{4}{3}\right)^2+\left(\frac{5}{3}\right)^2}}{\sqrt{3}}=1.2472\end{aligned}$$

Could you find this eigenvalue and associated eigenvector? Of course you could. This is what the inverse shifted 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, 3.8707, to get

$$\mathbf{u}_2=\begin{pmatrix}.69925\\.49389\\1.0\end{pmatrix}$$

Now solve

$$\begin{pmatrix}-\frac{16}{3}&2&3\\2&-\frac{13}{3}&1\\3&1&-\frac{7}{3}\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}.69925\\.49389\\1.0\end{pmatrix}$$

and divide by the entry with largest absolute value, 2.9979 to get

$$\mathbf{u}_3=\begin{pmatrix}.71473\\.52263\\1.0\end{pmatrix}$$

Now solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .71473 \\ .52263 \\ 1.0 \end{pmatrix}$$

and divide by the entry with largest absolute value, 3.0454, to get

$$\mathbf{u}_4 = \begin{pmatrix} .7137 \\ .52056 \\ 1.0 \end{pmatrix}$$

Solve

$$\begin{pmatrix} -\frac{16}{3} & 2 & 3 \\ 2 & -\frac{13}{3} & 1 \\ 3 & 1 & -\frac{7}{3} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} .7137 \\ .52056 \\ 1.0 \end{pmatrix}$$

and divide by the largest entry, 3.0421 to get

$$\mathbf{u}_5 = \begin{pmatrix} .71378 \\ .52073 \\ 1.0 \end{pmatrix}$$

You can see these scaling factors are not changing much. The predicted eigenvalue is obtained by solving

$$\frac{1}{\lambda - \frac{19}{3}} = 3.0421$$

to obtain $\lambda = 6.6621$. How close is this?

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 4 \end{pmatrix} \begin{pmatrix} .71378 \\ .52073 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 4.7552 \\ 3.469 \\ 6.6621 \end{pmatrix}$$

while

$$6.6621 \begin{pmatrix} .71378 \\ .52073 \\ 1.0 \end{pmatrix} = \begin{pmatrix} 4.7553 \\ 3.4692 \\ 6.6621 \end{pmatrix}.$$

You see that for practical purposes, this has found the eigenvalue and an eigenvector.

15.5 The *QR* Algorithm

15.5.1 Basic Considerations

The *QR* algorithm is one of the most remarkable techniques for finding eigenvalues. In this section, I will discuss this method. To see more on this algorithm, consult Golub and Van

Loan [6]. For an explanation of why the algorithm works see Wilkinson [18]. There is also more discussion in [13]. This will only discuss real matrices for the sake of simplicity. Also, there is a lot more to this algorithm than will be presented here. First here is an introductory lemma.

Lemma 15.5.1 *Suppose A is a block upper triangular matrix,*

$$A = \begin{pmatrix} B_1 & & * \\ & \ddots & \\ 0 & & B_r \end{pmatrix}$$

This means that the B_i are $r_i \times r_i$ matrices whose diagonals are subsets of the main diagonal of A. Then $\sigma(A) = \cup_{i=1}^{r} \sigma(B_i)$.

Proof: Say $Q_i^* B_i Q_i = T_i$ where T_i is upper triangular. Such unitary matrices exist by Schur's theorem. Then consider the similarity transformation,

$$\begin{pmatrix} Q_1^* & & 0 \\ & \ddots & \\ 0 & & Q_r^* \end{pmatrix} \begin{pmatrix} B_1 & & * \\ & \ddots & \\ 0 & & B_r \end{pmatrix} \begin{pmatrix} Q_1 & & 0 \\ & \ddots & \\ 0 & & Q_r \end{pmatrix}$$

By block multiplication this equals

$$\begin{pmatrix} Q_1^* & & 0 \\ & \ddots & \\ 0 & & Q_r^* \end{pmatrix} \begin{pmatrix} B_1 Q_1 & & * \\ & \ddots & \\ 0 & & B_r Q_r \end{pmatrix}$$

$$= \begin{pmatrix} Q_1^* B_1 Q_1 & & * \\ & \ddots & \\ 0 & & Q_r^* B_r Q_r \end{pmatrix} = \begin{pmatrix} T_1 & & * \\ & \ddots & \\ 0 & & T_r \end{pmatrix}$$

Now this is a real upper triangular matrix and the eigenvalues of A consist of the union of the eigenvalues of the T_i which is the same as the union of the eigenvalues of the B_i. ■

Here is the description of the great and glorious QR algorithm.

The QR Algorithm

Let A be an $n \times n$ real matrix. Let $A_0 = A$. Suppose that A_{k-1} has been found. To find A_k let

$$A_{k-1} = Q_k R_k,\ A_k = R_k Q_k,$$

where $Q_k R_k$ is a QR factorization of A_{k-1}. Thus R is upper triangular with nonnegative entries on the main diagonal and Q is real and unitary (orthogonal).

15.6 MATLAB And The QR Algorithm

This is most easily done in MATLAB. Given H you would then just do the QR algorithm on this matrix to get eigenvalues. The syntax for doing this is as follows. Here 50 iterations are being used.

```
H=[enter H here]
hold on
for k=1:50
[Q,R]=qr(H);
H=R*Q;
end
Q
R
H
```

Of course if MATLAB already knows H then you don't need to re-enter it. This happens when you use MATLAB to find an upper Hessenberg matrix similar to the original matrix. This is discussed later.

The main significance of this algorithm is in the following easy but important theorem.

Theorem 15.6.1 *Let A be any $n \times n$ complex matrix and let $\{A_k\}$ be the sequence of matrices described above by the QR algorithm. Then each of these matrices is unitarily similar to A.*

Proof: Clearly A_0 is orthogonally similar to A because they are the same thing. Suppose then that

$$A_{k-1} = Q^* A Q$$

Then from the algorithm,

$$A_{k-1} = Q_k R_k,\ R_k = Q_k^* A_{k-1}$$

Therefore, from the algorithm,

$$A_k \equiv R_k Q_k = Q_k^* A_{k-1} Q_k = Q_k^* Q^* A Q Q_k = (Q Q_k)^* A Q Q_k,$$

and so A_k is unitarily similar to A also. ■

Although the sequence $\{A_k\}$ may fail to converge, it is nevertheless often the case that for large k, A_k is of the form

$$A_k = \begin{pmatrix} B_k & & * \\ & \ddots & \\ e & & B_r \end{pmatrix}$$

where the B_i are blocks which run down the diagonal of the matrix, and all of the entries below this block diagonal are very small. Then letting T_B denote the matrix obtained by setting all of these small entries equal to zero, one can argue, using methods of analysis, that the eigenvalues of A_k are close to the eigenvalues of T_B. From Lemma 15.5.1 the eigenvalues of T_B are the eigenvalues of the blocks B_i. Thus, the eigenvalues of A are the same as those of A_k and these are close to the eigenvalues of T_B.

In proving things about this algorithm and also for the sake of convenience, here is a technical result.

Corollary 15.6.2 *For Q_k, R_k, A_k given in the QR algorithm,*

$$A = Q_1 \cdots Q_k A_k Q_k^* \cdots Q_1^* \tag{15.3}$$

For $Q^{(k)} \equiv Q_1 \cdots Q_k$ and $R^{(k)} \equiv R_k \cdots R_1$, it follows that

$$A^k = Q^{(k)} R^{(k)}$$

Here A^k is the usual thing, A raised to the k^{th} power.

Proof: From the algorithm,

$$A = A_0 = Q_1 R_1,\ Q_1^* A_0 = R_1,\ A_1 \equiv R_1 Q_1 = Q_1^* A Q_1$$

Hence

$$Q_1 A_1 Q_1^* = A$$

Suppose the formula 15.3 holds for k. Then from the algorithm,

$$A_k = Q_{k+1} R_{k+1},\ R_{k+1} = Q_{k+1}^* A_k,\ A_{k+1} = R_{k+1} Q_{k+1} = Q_{k+1}^* A_k Q_{k+1}$$

Hence $Q_{k+1} A_{k+1} Q_{k+1}^* = A_k$ and so

$$A = Q_1 \cdots Q_k A_k Q_k^* \cdots Q_1^* = Q_1 \cdots Q_k Q_{k+1} A_{k+1} Q_{k+1}^* Q_k^* \cdots Q_1^*$$

This shows the first part.

The second part is clearly true from the algorithm if $k = 1$. Then from the first part and the algorithm,

$$A = Q_1 \cdots Q_k Q_{k+1} A_{k+1} Q_{k+1}^* Q_k^* \cdots Q_1^* = Q_1 \cdots Q_k Q_{k+1} R_{k+1} \overbrace{Q_{k+1} Q_{k+1}^*}^{I} Q_k^* \cdots Q_1^*$$

It follows that

$$\begin{aligned} A^{k+1} &= AA^k = Q_1 \cdots Q_k Q_{k+1} R_{k+1} Q_k^* \cdots Q_1^* Q^{(k)} R^{(k)} \\ &= Q^{(k+1)} R_{k+1} \left(Q^{(k)}\right)^* Q^{(k)} R^{(k)} \end{aligned}$$

Hence

$$A^{k+1} = Q^{(k+1)} R_{k+1} R^{(k)} = Q^{(k+1)} R^{(k+1)} \blacksquare$$

Now suppose that A^{-1} exists. How do two QR factorizations compare? Since A^{-1} exists, it would require that if $A = QR$, then R^{-1} must exist. Now an upper triangular matrix has inverse which is also upper triangular. This follows right away from the algorithm presented early in the book for finding the inverse. If $A = Q_1 R_1 = Q_2 R_2$, then $Q_1^* Q_2 = R_1 R_2^{-1}$ and so $R_1 R_2^{-1}$ is an upper triangular matrix which is also unitary and in addition has all positive entries down the diagonal. For simplicity, call it R. Thus R is upper triangular and $RR^* = R^*R = I$. It follows easily that R must equal I and so $R_1 = R_2$ which requires $Q_1 = Q_2$.

Now in the above corollary, you know that

$$A = Q_1 \cdots Q_k A_k Q_k^* \cdots Q_1^* = Q^{(k)} A_k \left(Q^{(k)} \right)^*$$

Also, from this corollary, you know that

$$A^k = Q^{(k)} R^{(k)}$$

You could also simply take the QR factorization of A^k to obtain $A^k = QR$. Then from what was just pointed out, if A^{-1} exists,

$$Q^{(k)} = Q$$

Thus from the above corollary,

$$A_k = \left(Q^{(k)} \right)^* A Q^{(k)} = Q^* A Q$$

Therefore, in using the QR algorithm in the case where A has an inverse, it suffices to take

$$A^k = QR$$

and then consider the matrix

$$Q^* A Q = A_k.$$

This is so theoretically. In practice it might not work out all that well because of round off errors.

There is also an interesting relation to the power method. Let

$$A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix}$$

Then from the way we multiply matrices,

$$A^{k+1} = \begin{pmatrix} A^k \mathbf{a}_1 & \cdots & A^k \mathbf{a}_n \end{pmatrix}$$

and for large k, $A^k \mathbf{a}_i$ would be expected to point roughly in the direction of the eigenvector corresponding to the largest eigenvalue. Then if you form the QR factorization,

$$A^{k+1} = QR$$

the columns of Q are an orthonormal basis obtained essentially from the Gram Schmidt procedure. Thus the first column of Q has roughly the direction of an eigenvector associated with the largest eigenvalue of A. It follows that the first column of Q^*AQ is approximately equal to $\lambda_1 \mathbf{q}_1$ and so the top entry will be close to $\lambda_1 \mathbf{q}_1^* \mathbf{q}_1 = \lambda_1$ and the entries below it are close to 0. Thus the eigenvalues of the matrix should be close to this top entry of the first column along with the eigenvalues of the $(n-1) \times (n-1)$ matrix in the lower right corner. If this is a 2×2 you can find the eigenvalues using the quadratic formula. If it is larger, you could just use the same procedure for finding its eigenvalues but now you are dealing with a smaller matrix.

15.6.1 The Upper Hessenberg Form

Actually, when using the QR algorithm, contrary to what is discussed above, you should always deal with a matrix which is similar to the given matrix which is in upper Hessenberg form. This means all the entries below the sub diagonal equal 0. Here is an easy lemma.

Lemma 15.6.3 *Let A be an $n\times n$ matrix. Then it is unitarily similar to a matrix in upper Hessenberg form and this similarity can be computed.*

Proof: Let A be an $n\times n$ matrix. Suppose $n>2$. There is nothing to show otherwise.

$$A=\begin{pmatrix} a & \mathbf{b} \\ \mathbf{d} & A_1 \end{pmatrix}$$

where A_1 is $n-1\times n-1$. Consider the $n-1\times 1$ matrix $\mathbf{d}$. Then let Q be a Householder reflection such that

$$Q\mathbf{b}=\begin{pmatrix} c \\ \mathbf{0} \end{pmatrix}\equiv \mathbf{c}$$

Then

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q \end{pmatrix}\begin{pmatrix} a & \mathbf{b} \\ \mathbf{d} & A_1 \end{pmatrix}\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q^* \end{pmatrix}=\begin{pmatrix} a & \mathbf{b}Q^* \\ \mathbf{c} & QA_1Q^* \end{pmatrix}$$

By similar reasoning, there exists an $n-1\times n-1$ matrix

$$U=\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_1 \end{pmatrix}$$

such that

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_1 \end{pmatrix}QA_1Q^*\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Q_1^* \end{pmatrix}=\begin{pmatrix} * & * & \cdots & * \\ * & * & \ddots & \vdots \\ & \ddots & \ddots & * \\ \mathbf{0} & & * & * \end{pmatrix}$$

Thus

$$\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & U \end{pmatrix}\begin{pmatrix} a & \mathbf{b}Q^* \\ \mathbf{c} & QA_1Q^* \end{pmatrix}\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & U^* \end{pmatrix}$$

will have all zeros below the first two entries on the sub diagonal. Continuing this way shows the result. ■

Not surprisingly, MATLAB can find a Hessenberg matrix for a given square matrix. Here is the syntax.

```
A=[2 1 3;-5,3,-2;1,2,3];
[P,H]=hess(A)
P’*A*P
```

The output will be first P a unitary matrix, then H, a Hessenberg matrix and finally, the last line just verifies that $P^*AP = H$. Then you can do the QR algorithm on H.

The reason you should use a matrix which is upper Hessenberg and similar to A in the QR algorithm is that the algorithm keeps returning a matrix in upper Hessenberg form and if you are looking for block upper triangular matrices, this will force the size of the blocks to be no larger than 2×2 which are easy to handle using the quadratic formula. This is in the following lemma.

Lemma 15.6.4 *Let $\{A_k\}$ be the sequence of iterates from the QR algorithm, A^{-1} exists. Then if A_k is upper Hessenberg, so is A_{k+1}.*

Proof: The matrix is upper Hessenberg means that $A_{ij} = 0$ whenever $i - j \geq 2$.

$$A_{k+1} = R_k Q_k$$

where $A_k = Q_k R_k$. Therefore $A_k R_k^{-1} = Q_k$ and so

$$A_{k+1} = R_k Q_k = R_k A_k R_k^{-1}$$

Let the ij^{th} entry of A_k be a_{ij}^k. Then if $i - j \geq 2$

$$a_{ij}^{k+1} = \sum_{p=i}^{n} \sum_{q=1}^{j} r_{ip} a_{pq}^k r_{qj}^{-1}$$

It is given that $a_{pq}^k = 0$ whenever $p - q \geq 2$. However, from the above sum,

$$p - q \geq i - j \geq 2,$$

and so the sum equals 0. ■

Example 15.6.5 *Find the solutions to the equation $x^4 - 4x^3 + 8x^2 - 8x + 4 = 0$ using the QR algorithm.*

This is the characteristic equation of the matrix

$$H = \begin{pmatrix} 4 & -8 & 8 & -4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Since the constant term in the equation is not 0, it follows that the matrix has an inverse. It is already in upper Hessenberg form. Lets apply the QR algorithm above with 100 iterations. This yields the following matrix which is similar to H.

$$\begin{pmatrix} .6761 & -.7372 & .0469 & .1593 \\ 1.5344 & 1.3435 & 3.1593 & 10.7276 \\ 0 & 0.0001 & -1.3435 & -4.1885 \\ 0 & 0 & 1.5344 & 3.3239 \end{pmatrix}$$

The number in the third row and second column is so small that we neglect it. All that remains is to find the eigenvalues are the two blocks

$$\begin{pmatrix} .6761 & -.7372 \\ 1.5344 & 1.3435 \end{pmatrix}, \begin{pmatrix} -1.3435 & -4.1885 \\ 1.5344 & 3.3239 \end{pmatrix}$$

Thus the eigenvalues of the original matrix are those which result from these two blocks. Since these are 2×2 matrices, you can find the answer from the quadratic formula. Thus the eigenvalues are

$$\begin{gathered} 1.0098 + 1.0099i, 1.0098 - 1.0099i, \\ 0.9902 + 0.99029i, 0.9902 - 0.99029i \end{gathered}$$

In fact, the eigenvalues are exactly $1+i, 1+i, 1-i, 1-i$ listed according to multiplicity. Now you can use the shifted inverse power method to find much better approximations as well as eigenvectors. For example, you would use that algorithm to find the eigenvalue close to $0.9902 + 0.99029i$ along with the associated eigenvector. It yields $1+i$ as the eigenvalue closest to $0.99 + 0.99i$ along with the eigenvector

$$\begin{pmatrix} 1 & 0.5 - 0.5i & -0.5i & -0.25 - 0.25i \end{pmatrix}^T$$

Of course we didn't care about the eigenvector but there it is anyway. It took 844 iterations even though $0.99 + 0.99i$ was very close to the true eigenvalue.

Example 15.6.6 *Find the eigenvalues for the symmetric matrix*

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

Also find an eigenvector.

A Hessenberg matrix similar to the above matrix is

$$H = \begin{pmatrix} .0888 & -.6421 & 0 & 0 \\ -.6421 & 3.8398 & -3.348 & 0 \\ 0 & -3.348 & .0714 & -3.7417 \\ 0 & 0 & -3.7417 & 1 \end{pmatrix}$$

Thus you could use the QR algorithm on this to identify the eigenvalues. In using this algorithm, MATLAB already knows H unless you did clear all or close all. Thus you don't need to enter the matrix in the QR algorithm. You just need to refer to H.

This yields

$$\begin{pmatrix} 6.643 & 0 & 0 & 0 \\ 0 & -4.1018 & 0 & 0 \\ 0 & 0 & 2.4589 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Thus the eigenvalues are those diagonal entries. Now if you want to find eigenvectors, there is a way to keep track of things and get them from the above, but you could also simply go back to the matrix A and use the shifted inverse power method. Starting with one of these approximate eigenvalues or a number close to one as α. I shall pick the eigenvalue 6.643 and obtain an eigenvector and possibly a better approximation to this eigenvalue using the shifted inverse power method using the iterative procedure given above. This yields the eigenvector

$$\mathbf{u} = \begin{pmatrix} 0.6442 & 0.5961 & 1 & 0.5572 \end{pmatrix}^T$$

which works extremely well, along with the eigenvalue 6.643. In fact, the error between $A\mathbf{u}$ and $6.643\mathbf{u}$ is on the order of 10^{-14}.

15.7 Exercises

1. Using the power method, find the eigenvalue correct to one decimal place having largest absolute value for the matrix $A = \begin{pmatrix} 0 & -4 & -4 \\ 7 & 10 & 5 \\ -2 & 0 & 6 \end{pmatrix}$ along with an eigenvector associated with this eigenvalue.

2. Using the power method, find the eigenvalue correct to one decimal place having largest absolute value for the matrix $A = \begin{pmatrix} 15 & 6 & 1 \\ -5 & 2 & 1 \\ 1 & 2 & 7 \end{pmatrix}$ along with an eigenvector associated with this eigenvalue.

3. Using the power method, find the eigenvalue correct to one decimal place having largest absolute value for the matrix $A = \begin{pmatrix} 10 & 4 & 2 \\ -3 & 2 & -1 \\ 0 & 0 & 4 \end{pmatrix}$ along with an eigenvector associated with this eigenvalue.

4. Using the power method, find the eigenvalue correct to one decimal place having largest absolute value for the matrix $A = \begin{pmatrix} 15 & 14 & -3 \\ -13 & -18 & 9 \\ 5 & 10 & -1 \end{pmatrix}$ along with an eigenvector associated with this eigenvalue.

5. In Example 15.4.3 an eigenvalue was found correct to several decimal places along with an eigenvector. Find the other eigenvalues along with their eigenvectors.

6. 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.

7. 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?

8. Find the eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically. We don't know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

9. Find the eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 3 \\ 1 & 3 & 2 \end{pmatrix}$ numerically. We don't know the exact eigenvalues in this case. Check your answers by multiplying your numerically computed eigenvectors by the matrix.

10. Consider the matrix $A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 0 \end{pmatrix}$ and the vector $(1,1,1)^T$. Estimate the distance between the Rayleigh quotient determined by this vector and some eigenvalue of A.

11. Consider the matrix $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 1 & 4 \\ 1 & 4 & 5 \end{pmatrix}$ and the vector $(1,1,1)^T$. Estimate the distance between the Rayleigh quotient determined by this vector and some eigenvalue of A.

12. Using Gerschgorin's theorem, find upper and lower bounds for the eigenvalues of

$$A = \begin{pmatrix} 3 & 2 & 3 \\ 2 & 6 & 4 \\ 3 & 4 & -3 \end{pmatrix}.$$

13. The *QR* algorithm works very well on general matrices. Try the *QR* algorithm on the following matrix which happens to have some complex eigenvalues.

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

Use the *QR* algorithm to get approximate eigenvalues and then use the shifted inverse power method on one of these to get an approximate eigenvector for one of the complex eigenvalues.

14. Use the *QR* algorithm to approximate the eigenvalues of the symmetric matrix

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

15. Try to find the eigenvalues of the matrix $\begin{pmatrix} 3 & 3 & 1 \\ -2 & -2 & -1 \\ 0 & 1 & 0 \end{pmatrix}$ using the *QR* algorithm. It has eigenvalues $1, i, -i$. You will see the algorithm won't work well. ▶

16. Let $q(\lambda) = a_0 + a_1\lambda + \cdots + a_{n-1}\lambda^{n-1} + \lambda^n$. Now consider the **companion matrix**,

$$C \equiv \begin{pmatrix} 0 & \cdots & 0 & -a_0 \\ 1 & 0 & & -a_1 \\ & \ddots & \ddots & \vdots \\ 0 & & 1 & -a_{n-1} \end{pmatrix}$$

Show that $q(\lambda)$ is the characteristic equation for C. Thus the roots of $q(\lambda)$ are the eigenvalues of C. You can prove something similar for

$$C = \begin{pmatrix} -a_{n-1} & -a_{n-2} & \cdots & -a_0 \\ 1 & & & \\ & \ddots & & \\ & & 1 & \end{pmatrix}$$

Hint: The characteristic equation is

$$\det \begin{pmatrix} \lambda & \cdots & 0 & a_0 \\ -1 & \lambda & & a_1 \\ & \ddots & \ddots & \vdots \\ 0 & & -1 & \lambda + a_{n-1} \end{pmatrix}$$

Expand along the first column. Thus

$$\lambda \det \begin{pmatrix} \lambda & \cdots & 0 & a_1 \\ -1 & \lambda & & a_2 \\ & \ddots & \ddots & \vdots \\ 0 & & -1 & \lambda + a_{n-1} \end{pmatrix} + \det \begin{pmatrix} 0 & 0 & \cdots & a_0 \\ -1 & \lambda & \cdots & a_2 \\ \vdots & \ddots & & \vdots \\ 0 & & -1 & \lambda + a_3 \end{pmatrix}$$

Now use induction on the first term and for the second, note that you can expand along the top row to get

$$(-1)^{n-2} a_0 (-1)^n = a_0.$$

17. Suppose A is a real symmetric, invertible, matrix, or more generally one which has real eigenvalues. Then as described above, it is typically the case that

$$A^p = Q_1 R$$

and

$$Q_1^T A Q_1 = \begin{pmatrix} a_1 & \mathbf{b}_1^T \\ \mathbf{e}_1 & A_1 \end{pmatrix}$$

where $\mathbf{e}$ is very small. Then you can do the same thing with A_1 to obtain another smaller orthogonal matrix Q_2 such that

$$Q_2^T A_1 Q_2 = \begin{pmatrix} a_2 & \mathbf{b}_2^T \\ \mathbf{e}_2 & A_2 \end{pmatrix}$$

Explain why

$$\begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & Q_2 \end{pmatrix}^T Q_1^T A Q_1 \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & Q_2 \end{pmatrix} = \begin{pmatrix} a_1 & & * \\ \vdots & a_2 & \\ \mathbf{e}_1 & \mathbf{e}_2 & A_3 \end{pmatrix}$$

where the $\mathbf{e}_i$ are very small. Explain why one can construct an orthogonal matrix Q such that

$$Q^T A Q = (T + E)$$

where T is an upper triangular matrix and E is very small. In case A is symmetric, explain why T is actually a diagonal matrix. Next explain why, in the case of A symmetric, that the columns of Q are an orthonormal basis of vectors, each of which is close to an eigenvector. Thus this will compute, not just the eigenvalues but also the eigenvectors.

18. Explain how one could use the QR algorithm or the above procedure to compute the singular value decomposition of an arbitrary real $m \times n$ matrix.

Chapter 16

Vector Spaces

16.1 Algebraic Considerations

It is time to consider the idea of an abstract Vector space which is something which has two operations satisfying the following vector space axioms.

Definition 16.1.1 *A vector space is an Abelian group of "vectors" denoted here by bold face letters, 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.)

$$\alpha(\mathbf{v}+\mathbf{w})=\alpha\mathbf{v}+\alpha\mathbf{v}, \tag{16.1}$$

$$(\alpha+\beta)\mathbf{v}=\alpha\mathbf{v}+\beta\mathbf{v}, \tag{16.2}$$

$$\alpha(\beta\mathbf{v})=\alpha\beta(\mathbf{v}), \tag{16.3}$$

$$1\mathbf{v}=\mathbf{v}. \tag{16.4}$$

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. These axioms do not tell us anything about what is being considered. Nevertheless, one can prove some fundamental properties just based on these vector space axioms.

Proposition 16.1.2 *In any vector space,* $\mathbf{0}$ *is unique, and it follows that* $-\mathbf{x}$ *is unique,* $0\mathbf{x} = \mathbf{0}$, *and* $(-1)\mathbf{x} = -\mathbf{x}$.

Proof: Suppose $\mathbf{0}'$ is also an additive identity. Then for $\mathbf{0}$ the additive identity in the axioms,

$$\mathbf{0}' = \mathbf{0}' + \mathbf{0} = \mathbf{0}$$

Next suppose $\mathbf{x}+\mathbf{y} = \mathbf{0}$. Then add $-\mathbf{x}$ to both sides.

$$-\mathbf{x} = -\mathbf{x} + (\mathbf{x}+\mathbf{y}) = (-\mathbf{x}+\mathbf{x}) + \mathbf{y} = \mathbf{0} + \mathbf{y} = \mathbf{y}$$

Thus if $\mathbf{y}$ acts like the additive inverse, it is the additive inverse.

$$0\mathbf{x} = (0+0)\mathbf{x} = 0\mathbf{x} + 0\mathbf{x}$$

Now add $-0\mathbf{x}$ to both sides. This gives $\mathbf{0} = 0\mathbf{x}$. Finally,

$$(-1)\mathbf{x} + \mathbf{x} = (-1)\mathbf{x} + 1\mathbf{x} = (-1+1)\mathbf{x} = 0\mathbf{x} = \mathbf{0}$$

By the uniqueness of the additive inverse shown earlier, $(-1)\mathbf{x} = -\mathbf{x}$. ■

If you are interested in considering other fields, you should have some examples other than $\mathbb{C}$, $\mathbb{R}$, $\mathbb{Q}$. Some of these are discussed in the following exercises. If you are happy with only considering $\mathbb{R}$ and $\mathbb{C}$, skip these exercises. Here is an important example which gives the typical vector space.

Example 16.1.3 *Let* Ω *be a nonempty set and define* V *to be the set of functions defined on* Ω. *Letting* a, b, c *be scalars and* f, g, h *functions, the vector operations are defined as*

$$\begin{aligned} (f+g)(x) &\equiv f(x) + g(x) \\ (af)(x) &\equiv a(f(x)) \end{aligned}$$

Then this is an example of a vector space.

To verify this, we check the axioms.

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

Since x is arbitrary, $f+g = g+f$.

$$((f+g)+h)(x) \equiv (f+g)(x) + h(x) = (f(x)+g(x)) + h(x)$$

$$= f(x) + (g(x)+h(x)) = (f(x) + (g+h)(x)) = (f+(g+h))(x)$$

and so $(f+g)+h = f+(g+h)$. Let 0 denote the function which is given by $0(x) = 0$. Then this is an additive identity because

$$(f+0)(x) = f(x) + 0(x) = f(x)$$

and so $f+0 = f$. Let $-f$ be the function which satisfies $(-f)(x) \equiv -f(x)$. Then

$$(f+(-f))(x) \equiv f(x) + (-f)(x) \equiv f(x) + -f(x) = 0$$

Hence $f+(-f)=0$.

$$((a+b)f)(x)\equiv(a+b)f(x)=af(x)+bf(x)\equiv(af+bf)(x)$$

and so $(a+b)f=af+bf$.

$$\begin{aligned}(a(f+g))(x)&\equiv a(f+g)(x)\equiv a(f(x)+g(x))\\&=af(x)+bg(x)\equiv(af+bg)(x)\end{aligned}$$

and so $a(f+g)=af+bg$.

$$((ab)f)(x)\equiv(ab)f(x)=a(bf(x))\equiv(a(bf))(x)$$

so $(abf)=a(bf)$. Finally $(1f)(x)\equiv 1f(x)=f(x)$ so $1f=f$.

Note that $\mathbb{R}^n$ can be considered the set of real valued functions defined on $(1,2,\cdots,n)$. Thus everything up till now was just a special case of this more general situation.

16.2 Exercises

1. Prove the Euclidean algorithm: If m,n are positive integers, then there exist integers $q,r\geq 0$ such that $r<m$ and
$$n=qm+r$$
Hint: You might try considering
$$S\equiv\{n-km:k\in\mathbb{N}\text{ and }n-km<0\}$$
and picking the smallest integer in S or something like this.

2. ↑The greatest common divisor of two positive integers m,n, denoted as q is a positive number which divides both m and n and if p is any other positive number which divides both m,n, then p divides q. Recall what it means for p to divide q. It means that $q=pk$ for some integer k. Show that the greatest common divisor of m,n is the smallest positive integer in the set S
$$S\equiv\{xm+yn:x,y\in\mathbb{Z}\text{ and }xm+yn>0\}$$
Two positive integers are called relatively prime if their greatest common divisor is 1.

3. ↑A positive integer larger than 1 is called a prime number if the only positive numbers which divide it are 1 and itself. Thus 2,3,5,7, etc. are prime numbers. If m is a positive integer and p does not divide m where p is a prime number, show that p and m are relatively prime.

4. ↑There are lots of fields. This will give an example of a finite field. Let $\mathbb{Z}$ denote the set of integers. Thus $\mathbb{Z}=\{\cdots,-3,-2,-1,0,1,2,3,\cdots\}$. Also let p be a prime number. We will say that two integers, a,b are equivalent and write $a\sim b$ if $a-b$ is divisible by p. Thus they are equivalent if $a-b=px$ for some integer x. First show that $a\sim a$. Next show that if $a\sim b$ then $b\sim a$. Finally show that if $a\sim b$ and $b\sim c$ then $a\sim c$. For a an integer, denote by $[a]$ the set of all integers which is equivalent

to a, the equivalence class of a. Show first that is suffices to consider only $[a]$ for $a=0,1,2,\cdots,p-1$ and that for $0\leq a<b\leq p-1$, $[a]\neq[b]$. That is, $[a]=[r]$ where $r\in\{0,1,2,\cdots,p-1\}$. Thus there are exactly p of these equivalence classes. **Hint:** Recall the Euclidean algorithm. For $a>0$, $a=mp+r$ where $r<p$. Next define the following operations.

$$[a]+[b]\equiv[a+b]$$

$$[a][b]=[ab]$$

Show these operations are well defined. That is, if $[a]=[a']$ and $[b]=[b']$, then $[a]+[b]=[a']+[b']$ with a similar conclusion holding for multiplication. Thus for addition you need to verify $[a+b]=[a'+b']$ and for multiplication you need to verify $[ab]=[a'b']$. For example, if $p=5$ you have $[3]=[8]$ and $[2]=[7]$. Is $[2\times3]=[8\times7]$? Is $[2+3]=[8+7]$? Clearly so in this example because when you subtract, the result is divisible by 5. So why is this so in general? Now verify that $\{[0],[1],\cdots,[p-1]\}$ with these operations is a Field. This is called the integers modulo a prime and is written $\mathbb{Z}_p$. Since there are infinitely many primes p, it follows there are infinitely many of these finite fields. **Hint:** Most of the axioms are easy once you have shown the operations are well defined. The only two which are tricky are the ones which give the existence of the additive inverse and the multiplicative inverse. Of these, the first is not hard. $-[x]=[-x]$. Since p is prime, there exist integers x,y such that $1=px+ky$ and so $1-ky=px$ which says $1\sim ky$ and so $[1]=[ky]$. Now you finish the argument. What is the multiplicative identity in this collection of equivalence classes?

16.3 Linear Independence And Bases

Just as in the case of $\mathbb{F}^n$ one has a concept of subspace, linear independence, and bases.

Definition 16.3.1 *If $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}\subseteq V$, a vector space, then*

$$\operatorname{span}(\mathbf{v}_1,\cdots,\mathbf{v}_n)\equiv\left\{\sum_{i=1}^n\alpha_i\mathbf{v}_i:\alpha_i\in\mathbb{F}\right\}.$$

A non empty subset, $W\subseteq V$ is said to be a subspace if $a\mathbf{x}+b\mathbf{y}\in W$ whenever $a,b\in\mathbb{F}$ and $\mathbf{x},\mathbf{y}\in W$. The span of a set of vectors as just described is an example of a subspace.

Then the following fundamental result says that subspaces are subsets of a vector space which are themselves vector spaces.

Proposition 16.3.2 *Let W be a nonempty collection of vectors in V, a vector space. Then W is a subspace if and only if W is itself a vector space having the same operations as those defined on V.*

Proof: Suppose first that W is a subspace. It is obvious that all the algebraic laws hold on W because it is a subset of V and they hold on V. Thus $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$ along with the other axioms. Does W contain $\mathbf{0}$? Yes because it contains $0\mathbf{u}=\mathbf{0}$. See Proposition 16.1.2. Are the operations defined on W? That is, when you add vectors of W do you get a vector in W? When you multiply a vector in W by a scalar, do you get a vector in W? Yes. This is contained in the definition. Does every vector in W have an additive inverse? Yes by Proposition 16.1.2 because $-\mathbf{v} = (-1)\mathbf{v}$ which is given to be in W provided $\mathbf{v} \in W$.

Next suppose W is a vector space. Then by definition, it is closed with respect to linear combinations. Hence it is a subspace. ■

Next is the definition of linear independence.

Definition 16.3.3 *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.

The next theorem is called the exchange theorem. It is very important that you understand this theorem. There are two kinds of people who go further in linear algebra, those who understand this theorem and its corollary presented later and those who don't. Those who do understand these theorems are able to proceed and learn more linear algebra while those who don't are doomed to wander in the wilderness of confusion and sink into the swamp of despair. Therefore, I am giving multiple proofs. Try to understand at least one of them. Several amount to the same thing, just worded differently. Before giving the proof, here is some important notation.

Notation 16.3.4 *Let $\mathbf{w}_{ij} \in V$, a vector space and let $1 \leq i \leq r$ while $1 \leq j \leq s$. Thus these vectors can be listed in a rectangular array.*

$$\begin{array}{cccc} \mathbf{w}_{11} & \mathbf{w}_{12} & \cdots & \mathbf{w}_{1s} \\ \mathbf{w}_{21} & \mathbf{w}_{22} & \cdots & \mathbf{w}_{2s} \\ \vdots & \vdots & & \vdots \\ \mathbf{w}_{r1} & \mathbf{w}_{r2} & \cdots & \mathbf{w}_{rs} \end{array}$$

Then $\sum_{j=1}^{s}\sum_{i=1}^{r} \mathbf{w}_{ij}$ means to sum the vectors in each column and then to add the s sums which result while $\sum_{i=1}^{r}\sum_{j=1}^{s} \mathbf{w}_{ij}$ means to sum the vectors in each row and then to add the r sums which result. Either way you simply get the sum of all the vectors in the above array. This is because, from the vector space axioms, you can add vectors in any order and you get the same answer.

Theorem 16.3.5 *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 1: Let

$$\mathbf{x}_k = \sum_{j=1}^{s} a_{jk} \mathbf{y}_j$$

If $r>s$, then the matrix $A=\left(a_{jk}\right)$ has more columns than rows. By Corollary 8.2.8 one of these columns is a linear combination of the others. This implies there exist scalars $c_1,\cdots,c_r$ not all zero such that

$$\sum_{k=1}^{r} a_{jk}c_k=0,\ j=1,\cdots,r$$

Then

$$\sum_{k=1}^{r} c_k\mathbf{x}_k=\sum_{k=1}^{r} c_k\sum_{j=1}^{s} a_{jk}\mathbf{y}_j=\sum_{j=1}^{s}\left(\overbrace{\sum_{k=1}^{r} c_k a_{jk}}^{=0}\right)\mathbf{y}_j=\mathbf{0}$$

which contradicts the assumption that $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$ is linearly independent. Hence $r\leq s$. ■

Proof 2: Define span$(\mathbf{y}_1,\cdots,\mathbf{y}_s)\equiv V$, it follows there exist scalars, $c_1,\cdots,c_s$ such that

$$\mathbf{x}_1=\sum_{i=1}^{s} c_i\mathbf{y}_i. \tag{16.5}$$

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

Say $c_k\neq 0$. Then solve (16.5) for 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$\left(\mathbf{x}_1,\mathbf{z}_1,\cdots,\mathbf{z}_{s-1}\right)=V$ because if $\mathbf{v}\in V$, there exist constants $c_1,\cdots,c_s$ such that

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

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 $\mathbf{v}\in$ span$\left(\mathbf{x}_1,\mathbf{z}_1,\cdots,\mathbf{z}_{s-1}\right)$. 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$\left(\mathbf{x}_1,\cdots,\mathbf{x}_l,\mathbf{z}_1,\cdots,\mathbf{z}_p\right)=V$, 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$\left(\mathbf{x}_1,\cdots,\mathbf{x}_l,\mathbf{z}_1,\cdots,\mathbf{z}_p\right)=V$ there exist scalars, c_i and d_j such that

$$\mathbf{x}_{l+1}=\sum_{i=1}^{l} c_i\mathbf{x}_i+\sum_{j=1}^{p} d_j\mathbf{z}_j. \tag{16.6}$$

Not all the d_j can equal zero because if this were so, it would follow that $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$ would be a linearly dependent set because one of the vectors would equal a linear combination of the others. Therefore, (16.6) 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. ■

Proof 3: Let $V\equiv\text{span}\,(\mathbf{y}_1,\cdots,\mathbf{y}_s)$ and suppose $r>s$. Let

$$A_l\equiv\{\mathbf{x}_1,\cdots,\mathbf{x}_l\},A_0=\emptyset,$$

and let B_{s-l} denote a subset of the vectors, $\{\mathbf{y}_1,\cdots,\mathbf{y}_s\}$ which contains $s-l$ vectors and has the property that $\text{span}\,(A_l,B_{s-l})=V$. Note that the assumption of the theorem says $\text{span}\,(A_0,B_s)=V$.

Now an exchange operation is given for $\text{span}\,(A_l,B_{s-l})=V$. Since $r>s$, it follows $l<r$. Letting

$$B_{s-l}\equiv\{\mathbf{z}_1,\cdots,\mathbf{z}_{s-l}\}\subseteq\{\mathbf{y}_1,\cdots,\mathbf{y}_s\},$$

it follows there exist constants, c_i and d_i such that

$$\mathbf{x}_{l+1}=\sum_{i=1}^{l}c_i\mathbf{x}_i+\sum_{i=1}^{s-l}d_i\mathbf{z}_i,$$

and not all the d_i can equal zero. (If they were all equal to zero, it would follow that the set, $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$ would be dependent since one of the vectors in it would be a linear combination of the others.)

Let $d_k\neq 0$. Then z_k can be solved for as follows.

$$\mathbf{z}_k=\frac{1}{d_k}\mathbf{x}_{l+1}-\sum_{i=1}^{l}\frac{c_i}{d_k}\mathbf{x}_i-\sum_{i\neq k}\frac{d_i}{d_k}\mathbf{z}_i.$$

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}. The process exchanged a vector in B_{s-l} with one from $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$ 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 $\mathbf{z}\in\{\mathbf{y}_1,\cdots,\mathbf{y}_s\}$. Then one more application of the exchange operation yields $V=\text{span}\,(A_s)$. But this implies $\mathbf{x}_r\in\text{span}\,(A_s)=\text{span}\,(\mathbf{x}_1,\cdots,\mathbf{x}_s)$, contradicting the linear independence of $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$. It follows that $r\leq s$ as claimed. ■

Proof 4: Suppose $r>s$. Let $\mathbf{z}_k$ denote a vector of $\{\mathbf{y}_1,\cdots,\mathbf{y}_s\}$. Thus there exists j as small as possible such that

$$\text{span}\,(\mathbf{y}_1,\cdots,\mathbf{y}_s)=\text{span}\,(\mathbf{x}_1,\cdots,\mathbf{x}_m,\mathbf{z}_1,\cdots,\mathbf{z}_j)$$

where $m+j=s$. It is given that $m=0$, corresponding to no vectors of $\{\mathbf{x}_1,\cdots,\mathbf{x}_m\}$ and $j=s$, corresponding to all the $\mathbf{y}_k$ results in the above equation holding. If $j>0$ then $m<s$ and so

$$\mathbf{x}_{m+1}=\sum_{k=1}^{m}a_k\mathbf{x}_k+\sum_{i=1}^{j}b_i\mathbf{z}_i$$

Not all the b_i can equal 0 and so you can solve for one of the $\mathbf{z}_i$ in terms of

$$\mathbf{x}_{m+1},\mathbf{x}_m,\cdots,\mathbf{x}_1,$$

and the other z_k. Therefore, there exists

$$\left\{\mathbf{z}_1,\cdots,\mathbf{z}_{j-1}\right\}\subseteq\{\mathbf{y}_1,\cdots,\mathbf{y}_s\}$$

such that

$$\text{span}\,(\mathbf{y}_1,\cdots,\mathbf{y}_s)=\text{span}\left(\mathbf{x}_1,\cdots,\mathbf{x}_{m+1},\mathbf{z}_1,\cdots,\mathbf{z}_{j-1}\right)$$

contradicting the choice of j. Hence $j=0$ and

$$\text{span}\,(\mathbf{y}_1,\cdots,\mathbf{y}_s)=\text{span}\,(\mathbf{x}_1,\cdots,\mathbf{x}_s)$$

It follows that

$$\mathbf{x}_{s+1}\in\text{span}\,(\mathbf{x}_1,\cdots,\mathbf{x}_s)$$

contrary to the assumption the $\mathbf{x}_k$ are linearly independent. Therefore, $r\leq s$ as claimed. ■

Corollary 16.3.6 *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 16.3.5, $m\leq n$ and $n\leq m$. ■

This corollary is very important so here is another proof of it given independent of the exchange theorem above.

Theorem 16.3.7 *Let V be a vector space and $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ and $\{\mathbf{v}_1,\cdots,\mathbf{v}_m\}$ are two bases for V. Then $k=m$.*

Proof: Suppose $k>m$. Then since the vectors, $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ span V, there exist scalars, c_{ij} such that

$$\sum_{i=1}^{m}c_{ij}\mathbf{v}_i=\mathbf{u}_j.$$

Therefore,

$$\sum_{j=1}^{k}d_j\mathbf{u}_j=\mathbf{0}\text{ if and only if }\sum_{j=1}^{k}\sum_{i=1}^{m}c_{ij}d_j\mathbf{v}_i=\mathbf{0}$$

if and only if

$$\sum_{i=1}^{m}\left(\sum_{j=1}^{k}c_{ij}d_j\right)\mathbf{v}_i=\mathbf{0}$$

Now since$\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is independent, this happens if and only if

$$\sum_{j=1}^{k}c_{ij}d_j=0,\ i=1,2,\cdots,m.$$

However, this is a system of m equations in k variables, $d_1,\cdots,d_k$ and $m<k$. Therefore, there exists a solution to this system of equations in which not all the d_j are equal to zero. Recall why this is so. The augmented matrix for the system is of the form $\left(\begin{array}{cc} C & \mathbf{0} \end{array}\right)$ where C is a matrix which has more columns than rows. Therefore, there are free variables and hence nonzero solutions to the system of equations. However, this contradicts the linear independence of $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ because, as explained above, $\sum_{j=1}^k d_j\mathbf{u}_j=\mathbf{0}$. Similarly it cannot happen that $m>k$. ■

Definition 16.3.8 *A vector space V is of dimension n if it has a basis consisting of n vectors. This is well defined thanks to Corollary 16.3.6. It is always assumed here that $n<\infty$ and in this case, such a vector space is said to be finite dimensional.*

The following says that if you add a vector which is not in the span of a linearly independent set of vectors to this set of vectors, then the resulting list is linearly independent.

Lemma 16.3.9 *Suppose $\mathbf{v}\notin \operatorname{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$ and $\{\mathbf{u}_1,\cdots,\mathbf{u}_k\}$ is linearly independent. Then $\{\mathbf{u}_1,\cdots,\mathbf{u}_k,\mathbf{v}\}$ is also linearly independent.*

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

Theorem 16.3.10 *If $V=\operatorname{span}(\mathbf{u}_1,\cdots,\mathbf{u}_n)$ 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

$$S=\{E\subseteq\{\mathbf{u}_1,\cdots,\mathbf{u}_n\}\text{ such that }\operatorname{span}(E)=V\}.$$

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

$$\operatorname{span}(\mathbf{v}_1,\cdots,\mathbf{v}_m)=V$$

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,

$$V = \text{span}\left(\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_{k+1}, \cdots, \mathbf{v}_m\right)$$

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

$$\text{span}\left(\mathbf{u}_1, \cdots, \mathbf{u}_k\right) = V,$$

then $k = n$. If not, there exists a vector

$$\mathbf{u}_{k+1} \notin \text{span}\left(\mathbf{u}_1, \cdots, \mathbf{u}_k\right).$$

Then from Lemma 16.3.9, $\{\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{u}_{k+1}\}$ is also linearly independent. Continue adding vectors in this way until n linearly independent vectors have been obtained. Then

$$\text{span}\left(\mathbf{u}_1, \cdots, \mathbf{u}_n\right) = V$$

because if not, there would exist $\mathbf{u}_{n+1}$ as just described and $\{\mathbf{u}_1, \cdots, \mathbf{u}_{n+1}\}$ would be linearly independent having $n+1$ elements even though $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ is a basis. This would contradict Theorem 16.3.5. Therefore, this list is a basis. ■

Theorem 16.3.11 *Let V be a nonzero subspace of a finite dimensional 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 $\text{span}(\mathbf{v}_1) = V$, stop. $\{\mathbf{v}_1\}$ is a basis for V. Otherwise, there exists $\mathbf{v}_2 \in V$ which is not in $\text{span}(\mathbf{v}_1)$. By Lemma 16.3.9 $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a linearly independent set of vectors. If $\text{span}(\mathbf{v}_1, \mathbf{v}_2) = V$ stop, $\{\mathbf{v}_1, \mathbf{v}_2\}$ is a basis for V. If $\text{span}(\mathbf{v}_1, \mathbf{v}_2) \neq V$, then there exists $\mathbf{v}_3 \notin \text{span}(\mathbf{v}_1, \mathbf{v}_2)$ 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, Theorem 16.3.5. ■

Example 16.3.12 *Let V be the polynomials of degree no more than 2. Thus, abusing notation, $V = \text{span}\left(1, x, x^2\right)$. Is $\left\{1, x, x^2\right\}$ a basis for V?*

It will be a basis if it is linearly independent. Suppose then that

$$a + bx + cx^2 = 0$$

First let $x = 0$ and this shows that $a = 0$. Therefore, $bx + cx^2 = 0$. Take a derivative of both sides. Then

$$b + 2cx = 0$$

Let $x = 0$. Then $b = 0$. Hence $2cx = 0$ and this must hold for all x. Therefore, $c = 0$. This shows that these functions are linearly independent. Since they also span, these must yield a basis.

Example 16.3.13 *Let V be the polynomials of degree no more than 2. Is*

$$\left\{x^2+x+1, 2x+1, 3x^2+1\right\}$$

a basis for V?

If these vectors are linearly independent, then they must be a basis since otherwise, you could obtain four functions in V which are linearly independent which is impossible because there is a spanning set of only three vectors, namely $\left\{1, x, x^2\right\}$. Suppose then that

$$a\left(x^2+x+1\right)+b(2x+1)+c\left(3x^2+1\right)=0$$

Then

$$(a+3c)x^2+(a+2b)x+(a+b+c)=0$$

From the above example, you know that

$$a+3c=0,\ a+2b=0,\ a+b+c=0$$

and there is only one solution to this system of equations, $a=b=c=0$. Therefore, these are linearly independent and hence are a basis for this vector space.

16.4 Vector Spaces And Fields*

16.4.1 Irreducible Polynomials

I mentioned earlier that most things hold for arbitrary fields. However, I have not bothered to give any examples of other fields. This is the point of this section. It also turns out that showing the algebraic numbers are a field can be understood using vector space concepts and it gives a very convincing application of the abstract theory presented earlier in this chapter.

Here I will give some basic algebra relating to polynomials. This is interesting for its own sake but also provides the basis for constructing many different kinds of fields. The first is the Euclidean algorithm for polynomials.

Definition 16.4.1 *A polynomial is an expression of the form $p(\lambda)=\sum_{k=0}^{n}a_k\lambda^k$ where as usual λ^0 is defined to equal 1. Two polynomials are said to be equal if their corresponding coefficients are the same. Thus, in particular, $p(\lambda)=0$ means each of the $a_k=0$. An element of the field λ is said to be a root of the polynomial if $p(\lambda)=0$ in the sense that when you plug in λ into the formula and do the indicated operations, you get 0. The degree of a nonzero polynomial is the highest exponent appearing on λ. The degree of the zero polynomial $p(\lambda)=0$ is not defined.*

Example 16.4.2 *Consider the polynomial $p(\lambda)=\lambda^2+\lambda$ where the coefficients are in $\mathbb{Z}_2$. Is this polynomial equal to 0? Not according to the above definition, because its coefficients are not all equal to 0. However, $p(1)=p(0)=0$ so it sends every element of $\mathbb{Z}_2$ to 0. Note the distinction between saying it sends everything in the field to 0 with having the polynomial be the zero polynomial.*

The fundamental result is the division theorem for polynomials.

Lemma 16.4.3 *Let $f(\lambda)$ and $g(\lambda) \neq 0$ be polynomials. Then there exists a polynomial, $q(\lambda)$ such that*

$$f(\lambda) = q(\lambda) g(\lambda) + r(\lambda)$$

where the degree of $r(\lambda)$ is less than the degree of $g(\lambda)$ or $r(\lambda) = 0$. These polynomials $q(\lambda)$ and $r(\lambda)$ are unique.

Proof: Suppose that $f(\lambda) - q(\lambda) g(\lambda)$ is never equal to 0 for any $q(\lambda)$. If it is, then the conclusion follows. Now suppose

$$r(\lambda) = f(\lambda) - q(\lambda) g(\lambda)$$

and the degree of $r(\lambda)$ is $m \geq n$ where n is the degree of $g(\lambda)$. Say the leading term of $r(\lambda)$ is $b\lambda^m$ while the leading term of $g(\lambda)$ is $\hat{b}\lambda^n$. Then letting $a = b/\hat{b}$, $a\lambda^{m-n} g(\lambda)$ has the same leading term as $r(\lambda)$. Thus the degree of $r_1(\lambda) \equiv r(\lambda) - a\lambda^{m-n} g(\lambda)$ is no more than $m-1$. Then

$$r_1(\lambda) = f(\lambda) - \left(q(\lambda) g(\lambda) + a\lambda^{m-n} g(\lambda)\right) = f(\lambda) - \left(\overbrace{q(\lambda) + a\lambda^{m-n}}^{q_1(\lambda)}\right) g(\lambda)$$

Denote by S the set of polynomials $f(\lambda) - g(\lambda) l(\lambda)$. Out of all these polynomials, there exists one which has smallest degree $r(\lambda)$. Let this take place when $l(\lambda) = q(\lambda)$. Then by the above argument, the degree of $r(\lambda)$ is less than the degree of $g(\lambda)$. Otherwise, there is one which has smaller degree. Thus $f(\lambda) = g(\lambda) q(\lambda) + r(\lambda)$.

As to uniqueness, if you have $r(\lambda), \hat{r}(\lambda), q(\lambda), \hat{q}(\lambda)$ which work, then you would have

$$(\hat{q}(\lambda) - q(\lambda)) g(\lambda) = r(\lambda) - \hat{r}(\lambda)$$

Now if the polynomial on the right is not zero, then neither is the one on the left. Hence this would involve two polynomials which are equal although their degrees are different. This is impossible. Hence $r(\lambda) = \hat{r}(\lambda)$ and so, matching coefficients implies that $\hat{q}(\lambda) = q(\lambda)$. ■

Now with this lemma, here is another one which is very fundamental. First here is a definition. A polynomial is **monic** means it is of the form

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

That is, the leading coefficient is 1. In what follows, the coefficients of polynomials are in $\mathbb{F}$, a field of scalars which is completely arbitrary. Think $\mathbb{R}$ if you need an example.

Definition 16.4.4 *A polynomial f is said to divide a polynomial g if $g(\lambda) = f(\lambda) r(\lambda)$ for some polynomial $r(\lambda)$. Let $\{\phi_i(\lambda)\}$ be a finite set of polynomials. The greatest common divisor will be the* ***monic*** *polynomial $q(\lambda)$ such that $q(\lambda)$ divides each $\phi_i(\lambda)$ and if $p(\lambda)$ divides each $\phi_i(\lambda)$, then $p(\lambda)$ divides $q(\lambda)$. The finite set of polynomials $\{\phi_i\}$ is said to be relatively prime if their greatest common divisor is 1. A polynomial $f(\lambda)$ is irreducible if there is no polynomial with coefficients in $\mathbb{F}$ which divides it except nonzero scalar multiples of $f(\lambda)$ and constants.*

Proposition 16.4.5 *The greatest common divisor is unique.*

Proof: Suppose both $q(\lambda)$ and $q'(\lambda)$ work. Then $q(\lambda)$ divides $q'(\lambda)$ and the other way around and so

$$q'(\lambda) = q(\lambda) l(\lambda), \ q(\lambda) = l'(\lambda) q'(\lambda)$$

Therefore, the two must have the same degree. Hence $l'(\lambda), l(\lambda)$ are both constants. However, this constant must be 1 because both $q(\lambda)$ and $q'(\lambda)$ are monic. ■

Theorem 16.4.6 *Let $\psi(\lambda)$ be the greatest commondivisor of $\{\phi_i(\lambda)\}$, not all of which are zero polynomials. Then there exist polynomials $r_i(\lambda)$ such that*

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

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

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

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

where $r_i(\lambda)$ is a polynomial. Then $S \neq \emptyset$ because some $\phi_i(\lambda) \neq 0$. Then let the r_i be chosen such that the degree of the expression $\sum_{i=1}^{p} r_i(\lambda) \phi_i(\lambda)$ is as small as possible. Letting $\psi(\lambda)$ equal this sum, it remains to verify it is the greatest common divisor. First, does it divide each $\phi_i(\lambda)$? Suppose it fails to divide $\phi_1(\lambda)$. Then by Lemma 16.4.3

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

where degree of $r(\lambda)$ is less than that of $\psi(\lambda)$. Then dividing $r(\lambda)$ by the leading coefficient if necessary and denoting the result by $\psi_1(\lambda)$, it follows the degree of $\psi_1(\lambda)$ is less than the degree of $\psi(\lambda)$ and $\psi_1(\lambda)$ equals

$$\psi_1(\lambda) = (\phi_1(\lambda) - \psi(\lambda) l(\lambda)) a = \left(\phi_1(\lambda) - \sum_{i=1}^{p} r_i(\lambda) \phi_i(\lambda) l(\lambda)\right) a$$

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

for a suitable $a \in \mathbb{F}$. This is one of the polynomials in S. Therefore, $\psi(\lambda)$ does not have the smallest degree after all because the degree of $\psi_1(\lambda)$ is smaller. This is a contradiction. Therefore, $\psi(\lambda)$ divides $\phi_1(\lambda)$. Similarly it divides all the other $\phi_i(\lambda)$.

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

Lemma 16.4.7 *Suppose $\phi(\lambda)$ and $\psi(\lambda)$ are monic polynomials which are irreducible and not equal. Then they are relatively prime.*

Proof: Suppose $\eta(\lambda)$ is a nonconstant polynomial. If $\eta(\lambda)$ divides $\phi(\lambda)$, then since $\phi(\lambda)$ is irreducible, $\eta(\lambda)$ equals $a\phi(\lambda)$ for some $a \in \mathbb{F}$. If $\eta(\lambda)$ divides $\psi(\lambda)$ then it must be of the form $b\psi(\lambda)$ for some $b \in \mathbb{F}$ and so it follows

$$\psi(\lambda) = \frac{a}{b} \phi(\lambda)$$

but both $\psi(\lambda)$ and $\phi(\lambda)$ are monic polynomials which implies $a = b$ and so $\psi(\lambda) = \phi(\lambda)$. This is assumed not to happen. It follows the only polynomials which divide both $\psi(\lambda)$ and $\phi(\lambda)$ are constants and so the two polynomials are relatively prime. Thus a polynomial which divides them both must be a constant, and if it is monic, then it must be 1. Thus 1 is the greatest common divisor. ■

Lemma 16.4.8 *Let $\psi(\lambda)$ be an irreducible monic polynomial not equal to 1 which divides*

$$\prod_{i=1}^{p} \phi_i(\lambda)^{k_i},\ k_i \text{ a positive integer,}$$

where each $\phi_i(\lambda)$ is an irreducible monic polynomial not equal to 1. Then $\psi(\lambda)$ equals some $\phi_i(\lambda)$.

Proof : Say $\psi(\lambda) l(\lambda) = \prod_{i=1}^{p} \phi_i(\lambda)^{k_i}$. Suppose $\psi(\lambda) \neq \phi_i(\lambda)$ for all i. Then by Lemma 16.4.7, there exist polynomials $m_i(\lambda), n_i(\lambda)$ such that

$$\begin{aligned} 1 &= \psi(\lambda) m_i(\lambda) + \phi_i(\lambda) n_i(\lambda) \\ \phi_i(\lambda) n_i(\lambda) &= 1 - \psi(\lambda) m_i(\lambda) \end{aligned}$$

Hence,

$$\begin{aligned} \psi(\lambda) n(\lambda) &\equiv \psi(\lambda) l(\lambda) \prod_{i=1}^{p} n_i(\lambda)^{k_i} = \prod_{i=1}^{p} (n_i(\lambda) \phi_i(\lambda))^{k_i} \\ &= \prod_{i=1}^{p} (1 - \psi(\lambda) m_i(\lambda))^{k_i} = 1 + g(\lambda) \psi(\lambda) \end{aligned}$$

for a polynomial $g(\lambda)$. Thus

$$1 = \psi(\lambda)(n(\lambda) - g(\lambda))$$

which is impossible because $\psi(\lambda)$ is not equal to 1. ■

Now here is a simple lemma about canceling monic polynomials.

Lemma 16.4.9 *Suppose $p(\lambda)$ is a monic polynomial and $q(\lambda)$ is a polynomial such that*

$$p(\lambda) q(\lambda) = 0.$$

Then $q(\lambda) = 0$. Also if

$$p(\lambda) q_1(\lambda) = p(\lambda) q_2(\lambda)$$

then $q_1(\lambda) = q_2(\lambda)$.

Proof: Let

$$p(\lambda) = \sum_{j=1}^{k} p_j \lambda^j,\ q(\lambda) = \sum_{i=1}^{n} q_i \lambda^i,\ p_k = 1.$$

Then the product equals

$$\sum_{j=1}^{k} \sum_{i=1}^{n} p_j q_i \lambda^{i+j}.$$

Then look at those terms involving λ^{k+n}. This is $p_k q_n \lambda^{k+n}$ and is given to be 0. Since $p_k = 1$, it follows $q_n = 0$. Thus

$$\sum_{j=1}^{k} \sum_{i=1}^{n-1} p_j q_i \lambda^{i+j} = 0.$$

Then consider the term involving λ^{n-1+k} and conclude that since $p_k = 1$, it follows $q_{n-1} = 0$. Continuing this way, each $q_i = 0$. This proves the first part. The second follows from

$$p(\lambda)(q_1(\lambda) - q_2(\lambda)) = 0. \blacksquare$$

The following is the analog of the fundamental theorem of arithmetic for polynomials.

Theorem 16.4.10 *Let $f(\lambda)$ be a nonconstant polynomial with coefficients in $\mathbb{F}$. Then there is some $a \in \mathbb{F}$ such that $f(\lambda) = a\prod_{i=1}^{n} \phi_i(\lambda)$ where $\phi_i(\lambda)$ is an irreducible nonconstant monic polynomial and repeats are allowed. Furthermore, this factorization is unique in the sense that any two of these factorizations have the same nonconstant factors in the product, possibly in different order and the same constant a.*

Proof: That such a factorization exists is obvious. If $f(\lambda)$ is irreducible, you are done. Factor out the leading coefficient. If not, then $f(\lambda) = a\phi_1(\lambda)\phi_2(\lambda)$ where these are monic polynomials. Continue doing this with the ϕ_i and eventually arrive at a factorization of the desired form.

It remains to argue the factorization is unique except for order of the factors. Suppose

$$a\prod_{i=1}^{n} \phi_i(\lambda) = b\prod_{i=1}^{m} \psi_i(\lambda)$$

where the $\phi_i(\lambda)$ and the $\psi_i(\lambda)$ are all irreducible monic nonconstant polynomials and $a, b \in \mathbb{F}$. If $n > m$, then by Lemma 16.4.8, each $\psi_i(\lambda)$ equals one of the $\phi_j(\lambda)$. By the above cancellation lemma, Lemma 16.4.9, you can cancel all these $\psi_i(\lambda)$ with appropriate $\phi_j(\lambda)$ and obtain a contradiction because the resulting polynomials on either side would have different degrees. Similarly, it cannot happen that $n < m$. It follows $n = m$ and the two products consist of the same polynomials. Then it follows $a = b$. ■

The following corollary will be well used. This corollary seems rather believable but does require a proof.

Corollary 16.4.11 *Let $q(\lambda) = \prod_{i=1}^{p} \phi_i(\lambda)^{k_i}$ where the k_i are positive integers and the $\phi_i(\lambda)$ are irreducible monic polynomials. Suppose also that $p(\lambda)$ is a monic polynomial which divides $q(\lambda)$. Then*

$$p(\lambda) = \prod_{i=1}^{p} \phi_i(\lambda)^{r_i}$$

where r_i is a nonnegative integer no larger than k_i.

Proof: Using Theorem 16.4.10, let $p(\lambda) = b\prod_{i=1}^{s} \psi_i(\lambda)^{r_i}$ where the $\psi_i(\lambda)$ are each irreducible and monic and $b \in \mathbb{F}$. Since $p(\lambda)$ is monic, $b = 1$. Then there exists a polynomial $g(\lambda)$ such that

$$p(\lambda)g(\lambda) = g(\lambda)\prod_{i=1}^{s} \psi_i(\lambda)^{r_i} = \prod_{i=1}^{p} \phi_i(\lambda)^{k_i}$$

Hence $g(\lambda)$ must be monic. Therefore,

$$p(\lambda)g(\lambda) = \overbrace{\prod_{i=1}^{s}\psi_i(\lambda)^{r_i}}^{p(\lambda)}\prod_{j=1}^{l}\eta_j(\lambda) = \prod_{i=1}^{p}\phi_i(\lambda)^{k_i}$$

for η_j monic and irreducible. By uniqueness, each ψ_i equals one of the $\phi_j(\lambda)$ and the same holding true of the $\eta_i(\lambda)$. Therefore, $p(\lambda)$ is of the desired form. ■

16.4.2 Polynomials And Fields

When you have a polynomial like $x^2 - 3$ which has no rational roots, it turns out you can enlarge the field of rational numbers to obtain a larger field such that this polynomial does have roots in this larger field. I am going to discuss a systematic way to do this. It will turn out that for any polynomial with coefficients in any field, there always exists a possibly larger field such that the polynomial has roots in this larger field. This book has mainly featured the field of real or complex numbers but this procedure will show how to obtain many other fields which could be used in most of what was presented earlier in the book. Here is an important idea concerning equivalence relations which I hope is familiar.

Definition 16.4.12 *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 16.4.13 *$[x]$ 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.*

Also recall the notion of equivalence classes.

Theorem 16.4.14 *Let $\sim$ be an equivalence relation defined on a set, S and let $\mathscr{H}$ 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 $[x] \cap [y] = \emptyset$.*

Definition 16.4.15 *Let $\mathbb{F}$ be a field, for example the rational numbers, and denote by $\mathbb{F}[x]$ the polynomials having coefficients in $\mathbb{F}$. Suppose $p(x)$ is a polynomial. Let $a(x) \sim b(x)$ ($a(x)$ is similar to $b(x)$) when*

$$a(x) - b(x) = k(x)p(x)$$

for some polynomial $k(x)$.

Proposition 16.4.16 *In the above definition, $\sim$ is an equivalence relation.*

Proof: First of all, note that $a(x) \sim a(x)$ because their difference equals $0p(x)$. If $a(x) \sim b(x)$, then $a(x) - b(x) = k(x) p(x)$ for some $k(x)$. But then

$$b(x) - a(x) = -k(x) p(x)$$

and so $b(x) \sim a(x)$. Next suppose $a(x) \sim b(x)$ and $b(x) \sim c(x)$. Then $a(x) - b(x) = k(x) p(x)$ for some polynomial $k(x)$ and also $b(x) - c(x) = l(x) p(x)$ for some polynomial $l(x)$. Then

$$\begin{aligned} a(x) - c(x) &= a(x) - b(x) + b(x) - c(x) \\ &= k(x) p(x) + l(x) p(x) = (l(x) + k(x)) p(x) \end{aligned}$$

and so $a(x) \sim c(x)$ and this shows the transitive law. ■

With this proposition, here is another definition which essentially describes the elements of the new field. It will eventually be necessary to assume the polynomial $p(x)$ in the above definition is irreducible so I will begin assuming this.

Definition 16.4.17 *Let $\mathbb{F}$ be a field and let $p(x) \in \mathbb{F}[x]$ be a monic irreducible polynomial of degree greater than 0. Thus there is no polynomial having coefficients in $\mathbb{F}$ which divides $p(x)$ except for itself and constants. For the similarity relation defined in Definition 16.4.15, define the following operations on the equivalence classes. $[a(x)]$ is an equivalence class means that it is the set of all polynomials which are similar to $a(x)$.*

$$[a(x)] + [b(x)] = [a(x) + b(x)]$$

$$[a(x)][b(x)] = [a(x) b(x)]$$

This collection of equivalence classes is sometimes denoted by $\mathbb{F}[x]/(p(x))$.

Proposition 16.4.18 *In the situation of Definition 16.4.17 where $p(x)$ is a monic irreducible polynomial, the following are valid.*

1. *$p(x)$ and $q(x)$ are relatively prime for any $q(x) \in \mathbb{F}[x]$ which is not a multiple of $p(x)$.*
2. *The definitions of addition and multiplication are well defined.*
3. *If $a, b \in \mathbb{F}$ and $[a] = [b]$, then $a = b$. Thus $\mathbb{F}$ is a subset of $\mathbb{F}[x]/(p(x))$.*
4. *$\mathbb{F}[x]/(p(x))$ is a field in which the polynomial $p(x)$ has a root.*
5. *$\mathbb{F}[x]/(p(x))$ is a vector space with field of scalars $\mathbb{F}$ and its dimension is m where m is the degree of the irreducible polynomial $p(x)$.*

Proof: First consider the claim about $p(x), q(x)$ being relatively prime. If $\psi(x)$ is the greatest common divisor, it follows $\psi(x)$ is either equal to $p(x)$ or 1. If it is $p(x)$, then $q(x)$ is a multiple of $p(x)$ which does not happen. If it is 1, then by definition, the two polynomials are relatively prime.

To show the operations are well defined, suppose

$$[a(x)] = \left[a'(x)\right], [b(x)] = \left[b'(x)\right]$$

It is necessary to show

$$[a(x) + b(x)] = \left[a'(x) + b'(x)\right]$$

$$[a(x)b(x)] = [a'(x)b'(x)]$$

Consider the second of the two.

$$\begin{aligned} & a'(x)b'(x) - a(x)b(x) \\ = \;& a'(x)b'(x) - a(x)b'(x) + a(x)b'(x) - a(x)b(x) \\ = \;& b'(x)\left(a'(x) - a(x)\right) + a(x)\left(b'(x) - b(x)\right) \end{aligned}$$

Now by assumption $(a'(x) - a(x))$ is a multiple of $p(x)$ as is $(b'(x) - b(x))$, so the above is a multiple of $p(x)$ and by definition this shows $[a(x)b(x)] = [a'(x)b'(x)]$. The case for addition is similar.

Now suppose $[a] = [b]$. This means $a - b = k(x)p(x)$ for some polynomial $k(x)$. Then $k(x)$ must equal 0 since otherwise the two polynomials $a - b$ and $k(x)p(x)$ could not be equal because they would have different degree.

It is clear that the axioms of a field are satisfied except for the one which says that non zero elements of the field have a multiplicative inverse. Let $[q(x)] \in \mathbb{F}[x]/(p(x))$ where $[q(x)] \neq [0]$. Then $q(x)$ is not a multiple of $p(x)$ and so by the first part, $q(x), p(x)$ are relatively prime. Thus there exist $n(x), m(x)$ such that

$$1 = n(x)q(x) + m(x)p(x)$$

Hence

$$[1] = [1 - n(x)p(x)] = [n(x)q(x)] = [n(x)][q(x)]$$

which shows that $[q(x)]^{-1} = [n(x)]$. Thus this is a field. The polynomial has a root in this field because if

$$p(x) = x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0,$$

$$[0] = [p(x)] = [x]^m + [a_{m-1}][x]^{m-1} + \cdots + [a_1][x] + [a_0]$$

Thus $[x]$ is a root of this polynomial in the field $\mathbb{F}[x]/(p(x))$.

Consider the last claim. Let $f(x) \in \mathbb{F}[x]/(p(x))$. Thus $[f(x)]$ is a typical thing in $\mathbb{F}[x]/(p(x))$. Then from the division algorithm,

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

where $r(x)$ is either 0 or has degree less than the degree of $p(x)$. Thus

$$[r(x)] = [f(x) - p(x)q(x)] = [f(x)]$$

but clearly $[r(x)] \in \text{span}\left([1], \cdots, [x]^{m-1}\right)$. Thus

$$\text{span}\left([1], \cdots, [x]^{m-1}\right) = \mathbb{F}[x]/(p(x)).$$

Then $\left\{[1], \cdots, [x]^{m-1}\right\}$ is a basis if these vectors are linearly independent. Suppose then that

$$\sum_{i=0}^{m-1} c_i [x]^i = \left[\sum_{i=0}^{m-1} c_i x^i\right] = 0$$

Then you would need to have $p(x) / \sum_{i=0}^{m-1} c_i x^i$ which is impossible unless each $c_i = 0$ because $p(x)$ has degree m. ■

The last assertion in the proof follows from the definition of addition and multiplication in $\mathbb{F}[x]/(p(x))$ and math induction. If each $a_i \in \mathbb{F}$,

$$\left[a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0\right] = [a_n][x]^n + [a_{n-1}][x]^{n-1} + \cdots [a_1][x] + [a_0] \tag{16.7}$$

Remark 16.4.19 *The polynomials consisting of all polynomial multiples of $p(x)$, denoted by $(p(x))$ is called an ideal. An ideal I is a subset of the commutative ring (Here the ring is $\mathbb{F}[x]$.) with unity consisting of all polynomials which is itself a ring and which has the property that whenever $f(x) \in \mathbb{F}[x]$, and $g(x) \in I$, $f(x)g(x) \in I$. In this case, you could argue that $(p(x))$ is an ideal and that the only ideal containing it is itself or the entire ring $\mathbb{F}[x]$. This is called a maximal ideal.*

Example 16.4.20 *The polynomial $x^2 - 2$ is irreducible in $\mathbb{Q}[x]$. This is because if $x^2 - 2 = p(x)q(x)$ where $p(x), q(x)$ both have degree less than 2, then they both have degree 1. Hence you would have $x^2 - 2 = (x+a)(x+b)$ which requires that $a + b = 0$ so this factorization is of the form $(x-a)(x+a)$ and now you need to have $a = \sqrt{2} \notin \mathbb{Q}$. Now $\mathbb{Q}[x]/\left(x^2 - 2\right)$ is of the form $a + b[x]$ where $a, b \in \mathbb{Q}$ and $[x]^2 - 2 = 0$. Thus one can regard $[x]$ as $\sqrt{2}$. $\mathbb{Q}[x]/\left(x^2 - 2\right)$ is of the form $a + b\sqrt{2}$.*

Thus the above is an illustration of something general described in the following definition.

Definition 16.4.21 *Let $F \subseteq K$ be two fields. Then clearly K is also a vector space over F. Then also, K is called a finite field extension of F if the dimension of this vector space, denoted by $[K : F]$ is finite.*

There are some easy things to observe about this.

Proposition 16.4.22 *Let $F \subseteq K \subseteq L$ be fields. Then $[L : F] = [L : K][K : F]$.*

Proof: Let $\{l_i\}_{i=1}^n$ be a basis for L over K and let $\left\{k_j\right\}_{j=1}^m$ be a basis of K over F. Then if $l \in L$, there exist unique scalars x_i in K such that $l = \sum_{i=1}^n x_i l_i$. Now $x_i \in K$ so there exist f_{ji} such that $x_i = \sum_{j=1}^m f_{ji} k_j$. Then it follows that

$$l = \sum_{i=1}^n \sum_{j=1}^m f_{ji} k_j l_i$$

It follows that $\left\{k_j l_i\right\}$ is a spanning set. Is it linearly independent? Suppose

$$\sum_{i=1}^n \sum_{j=1}^m f_{ji} k_j l_i = 0.$$

Then, since the l_i are independent, $\sum_{j=1}^m f_{ji} k_j = 0$ and since $\left\{k_j\right\}_j$ is independent, each $f_{ji} = 0$ for each j for a given arbitrary i. Therefore, $\left\{k_j l_i\right\}$ is a basis. ■

Note that if $p(x)$ were not irreducible, then you could find a field extension $\mathbb{G}$ such that $[\mathbb{G} : \mathbb{F}] \leq n$. You could do this by working with an irreducible factor of $p(x)$.

Usually, people simply write b rather than $[b]$ if $b \in \mathbb{F}$. Then with this convention,

$$[b\phi(x)] = [b][\phi(x)] = b[\phi(x)].$$

This shows how to enlarge a field to get a new one in which the polynomial has a root. By using a succession of such enlargements, called field extensions, there will exist a field in which the given polynomial can be factored into a product of polynomials having degree one. The field you obtain in this process of enlarging in which the given polynomial factors in terms of linear factors is called a splitting field.

Theorem 16.4.23 *Let $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a polynomial with coefficients in a field of scalars $\mathbb{F}$. There exists a larger field $\mathbb{G}$ such that there exist $\{z_1, \cdots, z_n\}$ listed according to multiplicity such that*

$$p(x) = \prod_{i=1}^{n} (x - z_i)$$

This larger field is called a splitting field. Furthermore,

$$[\mathbb{G} : \mathbb{F}] \leq n!$$

Proof: From Proposition 16.4.18, there exists a field $\mathbb{F}_1$ such that $p(x)$ has a root, z_1 ($= [x]$ if p is irreducible.) Then by the Euclidean algorithm

$$p(x) = (x - z_1) q_1(x) + r$$

where $r \in \mathbb{F}_1$. Since $p(z_1) = 0$, this requires $r = 0$. Now do the same for $q_1(x)$ that was done for $p(x)$, enlarging the field to $\mathbb{F}_2$ if necessary, such that in this new field

$$q_1(x) = (x - z_2) q_2(x).$$

and so

$$p(x) = (x - z_1)(x - z_2) q_2(x)$$

After n such extensions, you will have obtained the necessary field $\mathbb{G}$.

Finally consider the claim about dimension. By Proposition 16.4.18, there is a larger field $\mathbb{G}_1$ such that $p(x)$ has a root a_1 in $\mathbb{G}_1$ and $[\mathbb{G} : \mathbb{F}] \leq n$. Then

$$p(x) = (x - a_1) q(x)$$

Continue this way until the polynomial equals the product of linear factors. Then by Proposition 16.4.22 applied multiple times, $[\mathbb{G} : \mathbb{F}] \leq n!$. ■

Example 16.4.24 *The polynomial $x^2 + 1$ is irreducible in $\mathbb{R}(x)$, polynomials having real coefficients. To see this is the case, suppose $\psi(x)$ divides $x^2 + 1$. Then*

$$x^2 + 1 = \psi(x) q(x)$$

If the degree of $\psi(x)$ is less than 2, then it must be either a constant or of the form $ax + b$. In the latter case, $-b/a$ must be a zero of the right side, hence of the left but $x^2 + 1$ has no real zeros. Therefore, the degree of $\psi(x)$ must be two and $q(x)$ must be a constant. Thus the only polynomial which divides $x^2 + 1$ are constants and multiples of $x^2 + 1$. Therefore, this shows $x^2 + 1$ is irreducible. Find the inverse of $\left[x^2 + x + 1\right]$ in the space of equivalence classes, $\mathbb{R}/\left(x^2 + 1\right)$.

You can solve this with partial fractions.

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

and so

$$1 = (-x)\left(x^2+x+1\right) + (x+1)\left(x^2+1\right)$$

which implies

$$1 \sim (-x)\left(x^2+x+1\right)$$

and so the inverse is $[-x]$.

The following proposition is interesting. It was essentially proved above but to emphasize it, here it is again.

Proposition 16.4.25 *Suppose $p(x) \in \mathbb{F}[x]$ is irreducible and has degree n. Then every element of $\mathbb{G} = \mathbb{F}[x]/(p(x))$ is of the form $[0]$ or $[r(x)]$ where the degree of $r(x)$ is less than n.*

Proof: This follows right away from the Euclidean algorithm for polynomials. If $k(x)$ has degree larger than $n-1$, then

$$k(x) = q(x)\,p(x) + r(x)$$

where $r(x)$ is either equal to 0 or has degree less than n. Hence

$$[k(x)] = [r(x)]. \blacksquare$$

Example 16.4.26 *In the situation of the above example, find $[ax+b]^{-1}$ assuming*

$$a^2+b^2 \neq 0.$$

Note this includes all cases of interest thanks to the above proposition.

You can do it with partial fractions as above.

$$\frac{1}{(x^2+1)(ax+b)} = \frac{b-ax}{(a^2+b^2)(x^2+1)} + \frac{a^2}{(a^2+b^2)(ax+b)}$$

and so

$$1 = \frac{1}{a^2+b^2}(b-ax)(ax+b) + \frac{a^2}{(a^2+b^2)}\left(x^2+1\right)$$

Thus

$$\frac{1}{a^2+b^2}(b-ax)(ax+b) \sim 1$$

and so

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

You might find it interesting to recall that $(ai+b)^{-1} = \frac{b-ai}{a^2+b^2}$.

16.4.3 The Algebraic Numbers

Each polynomial having coefficients in a field $\mathbb{F}$ has a splitting field. Consider the case of all polynomials $p(x)$ having coefficients in a field $\mathbb{F} \subseteq \mathbb{G}$ and consider all roots which are also in $\mathbb{G}$. The theory of vector spaces is very useful in the study of these algebraic numbers. Here is a definition.

Definition 16.4.27 *The algebraic numbers $\mathbb{A}$ are those numbers which are in $\mathbb{G}$ and also roots of some polynomial $p(x)$ having coefficients in $\mathbb{F}$. The minimal polynomial of $a \in \mathbb{A}$ is defined to be the monic polynomial $p(x)$ having smallest degree such that $p(a) = 0$.*

Theorem 16.4.28 *Let $a \in \mathbb{A}$. Then there exists a unique monic irreducible polynomial $p(x)$ having coefficients in $\mathbb{F}$ such that $p(a) = 0$. This polynomial is the minimal polynomial.*

Proof: Let $p(x)$ be the monic polynomial having smallest degree such that $p(a) = 0$. Then $p(x)$ is irreducible because if not, there would exist a polynomial having smaller degree which has a as a root. Now suppose $q(x)$ is monic and irreducible such that $q(a) = 0$.

$$q(x) = p(x)l(x) + r(x)$$

where if $r(x) \neq 0$, then it has smaller degree than $p(x)$. But in this case, the equation implies $r(a) = 0$ which contradicts the choice of $p(x)$. Hence $r(x) = 0$ and so, since $q(x)$ is irreducible, $l(x) = 1$ showing that $p(x) = q(x)$. ■

Definition 16.4.29 *For a an algebraic number, let* $\deg(a)$ *denote the degree of the minimal polynomial of a.*

Also, here is another definition.

Definition 16.4.30 *Let $a_1, \cdots, a_m$ be in $\mathbb{A}$. A polynomial in $\{a_1, \cdots, a_m\}$ will be an expression of the form*

$$\sum_{k_1 \cdots k_n} a_{k_1 \cdots k_n} a_1^{k_1} \cdots a_n^{k_n}$$

where the $a_{k_1 \cdots k_n}$ are in $\mathbb{F}$, each k_j is a nonnegative integer, and all but finitely many of the $a_{k_1 \cdots k_n}$ equal zero. The collection of such polynomials will be denoted by

$$\mathbb{F}[a_1, \cdots, a_m].$$

Now notice that for a an algebraic number, $\mathbb{F}[a]$ is a vector space with field of scalars $\mathbb{F}$. Similarly, for $\{a_1, \cdots, a_m\}$ algebraic numbers, $\mathbb{F}[a_1, \cdots, a_m]$ is a vector space with field of scalars $\mathbb{F}$. The following fundamental proposition is important.

Proposition 16.4.31 *Let $\{a_1, \cdots, a_m\}$ be algebraic numbers. Then*

$$\dim \mathbb{F}[a_1, \cdots, a_m] \leq \prod_{j=1}^{m} \deg(a_j)$$

and for an algebraic number a,

$$\dim \mathbb{F}[a] = \deg(a)$$

Every element of $\mathbb{F}[a_1, \cdots, a_m]$ is in $\mathbb{A}$ and $\mathbb{F}[a_1, \cdots, a_m]$ is a field.

Proof: Let the minimal polynomial be

$$p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

If $q(a) \in \mathbb{F}[a]$, then

$$q(x) = p(x) l(x) + r(x)$$

where $r(x)$ has degree less than the degree of $p(x)$ if it is not zero. Thus $\mathbb{F}[a]$ is spanned by

$$\left\{1, a, a^2, \cdots, a^{n-1}\right\}$$

Since $p(x)$ has smallest degree of all polynomial which have a as a root, the above set is also linearly independent. This proves the second claim.

Now consider the first claim. By definition, $\mathbb{F}[a_1, \cdots, a_m]$ is obtained from all linear combinations of products of $\left\{a_1^{k_1}, a_2^{k_2}, \cdots, a_n^{k_n}\right\}$ where the k_i are nonnegative integers. From the first part, it suffices to consider only $k_j \leq \deg(a_j)$. Therefore, there exists a spanning set for $\mathbb{F}[a_1, \cdots, a_m]$ which has

$$\prod_{i=1}^{m} \deg(a_i)$$

entries. By Theorem 16.3.5 this proves the first claim.

Finally consider the last claim. Let $g(a_1, \cdots, a_m)$ be a polynomial in

$$\mathbb{F}[a_1, \cdots, a_m]$$

Since

$$\dim \mathbb{F}[a_1, \cdots, a_m] \equiv p \leq \prod_{j=1}^{m} \deg(a_j) < \infty,$$

it follows

$$1, g(a_1, \cdots, a_m), g(a_1, \cdots, a_m)^2, \cdots, g(a_1, \cdots, a_m)^p$$

are dependent. It follows $g(a_1, \cdots, a_m)$ is the root of some polynomial having coefficients in $\mathbb{F}$. Thus everything in $\mathbb{F}[a_1, \cdots, a_m]$ is algebraic. Why is $\mathbb{F}[a_1, \cdots, a_m]$ a field? Let $g(a_1, \cdots, a_m)$ be as just mentioned. Then it has a minimal polynomial,

$$p(x) = x^q + a_{q-1}x^{q-1} + \cdots + a_1 x + a_0$$

where the $a_i \in \mathbb{F}$. Then $a_0 \neq 0$ or else the polynomial would not be minimal. Therefore,

$$g(a_1, \cdots, a_m)\left(g(a_1, \cdots, a_m)^{q-1} + a_{q-1} g(a_1, \cdots, a_m)^{q-2} + \cdots + a_1\right) = -a_0$$

and so the multiplicative inverse for $g(a_1, \cdots, a_m)$ is

$$\frac{g(a_1, \cdots, a_m)^{q-1} + a_{q-1} g(a_1, \cdots, a_m)^{q-2} + \cdots + a_1}{-a_0} \in \mathbb{F}[a_1, \cdots, a_m].$$

The other axioms of a field are obvious. ■

Now from this proposition, it is easy to obtain the following interesting result about the algebraic numbers.

Theorem 16.4.32 *The algebraic numbers $\mathbb{A}$, those roots of polynomials in $\mathbb{F}[x]$ which are in $\mathbb{G}$, are a field.*

Proof: By definition, each $a \in \mathbb{A}$ has a minimal polynomial. Let $a \neq 0$ be an algebraic number and let $p(x)$ be its minimal polynomial. Then $p(x)$ is of the form

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

where $a_0 \neq 0$. Otherwise $p(x)$ would not have minimal degree. Then plugging in a yields

$$a\frac{\left(a^{n-1} + a_{n-1}a^{n-2} + \cdots + a_1\right)(-1)}{a_0} = 1.$$

and so $a^{-1} = \frac{\left(a^{n-1} + a_{n-1}a^{n-2} + \cdots + a_1\right)(-1)}{a_0} \in \mathbb{F}[a]$. By the proposition, every element of $\mathbb{F}[a]$ is in $\mathbb{A}$ and this shows that for every nonzero element of $\mathbb{A}$, its inverse is also in $\mathbb{A}$. What about products and sums of things in $\mathbb{A}$? Are they still in $\mathbb{A}$? Yes. If $a, b \in \mathbb{A}$, then both $a+b$ and $ab \in \mathbb{F}[a,b]$ and from the proposition, each element of $\mathbb{F}[a,b]$ is in $\mathbb{A}$. ■

A typical example of what is of interest here is when the field $\mathbb{F}$ of scalars is $\mathbb{Q}$, the rational numbers and the field $\mathbb{G}$ is $\mathbb{R}$. However, you can certainly conceive of many other examples by considering the integers mod a prime, for example (See Problem 4 on Page 379 for example.) or any of the fields which occur as field extensions in the above.

There is a very interesting thing about $\mathbb{F}[a_1 \cdots a_n]$ in the case where $\mathbb{F}$ is infinite which says that there exists a single algebraic γ such that $\mathbb{F}[a_1 \cdots a_n] = \mathbb{F}[\gamma]$. In other words, every field extension of this sort is a simple field extension. I found this fact in an early version of [5].

Proposition 16.4.33 *There exists γ such that $\mathbb{F}[a_1 \cdots a_n] = \mathbb{F}[\gamma]$.*

Proof: To begin with, consider $\mathbb{F}[\alpha, \beta]$. Let $\gamma = \alpha + \lambda\beta$. Then by Proposition 16.4.31 γ is an algebraic number and it is also clear

$$\mathbb{F}[\gamma] \subseteq \mathbb{F}[\alpha, \beta]$$

I need to show the other inclusion. This will be done for a suitable choice of λ. To do this, it suffices to verify that both α and β are in $\mathbb{F}[\gamma]$.

Let the minimal polynomials of α and β be $f(x)$ and $g(x)$ respectively. Let the distinct roots of $f(x)$ and $g(x)$ be $\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$ and $\{\beta_1, \beta_2, \cdots, \beta_m\}$ respectively. These roots are in a field which contains splitting fields of both $f(x)$ and $g(x)$. Let $\alpha = \alpha_1$ and $\beta = \beta_1$. Now define

$$h(x) \equiv f(\alpha + \lambda\beta - \lambda x) \equiv f(\gamma - \lambda x)$$

so that $h(\beta) = f(\alpha) = 0$. It follows $(x - \beta)$ divides both $h(x)$ and $g(x)$. If $(x - \eta)$ is a different linear factor of both $g(x)$ and $h(x)$ then it must be $\left(x - \beta_j\right)$ for some β_j for some $j > 1$ because these are the only factors of $g(x)$. Therefore, this would require

$$0 = h\left(\beta_j\right) = f\left(\alpha_1 + \lambda\beta_1 - \lambda\beta_j\right)$$

and so it would be the case that $\alpha_1 + \lambda\beta_1 - \lambda\beta_j = \alpha_k$ for some k. Hence

$$\lambda = \frac{\alpha_k - \alpha_1}{\beta_1 - \beta_j}$$

Now there are finitely many quotients of the above form and if λ is chosen to not be any of them, then the above cannot happen and so in this case, the only linear factor of both $g(x)$ and $h(x)$ will be $(x-\beta)$. Choose such a λ.

Let $\phi(x)$ be the minimal polynomial of β with respect to the field $\mathbb{F}[\gamma]$. Then this minimal polynomial must divide both $h(x)$ and $g(x)$ because $h(\beta)=g(\beta)=0$. However, the only factor these two have in common is $x-\beta$ and so $\phi(x)=x-\beta$ which requires $\beta\in\mathbb{F}[\gamma]$. Now also $\alpha=\gamma-\lambda\beta$ and so $\alpha\in\mathbb{F}[\gamma]$ also. Therefore, both $\alpha,\beta\in\mathbb{F}[\gamma]$ which forces $\mathbb{F}[\alpha,\beta]\subseteq\mathbb{F}[\gamma]$. This proves the proposition in the case that $n=2$. The general result follows right away by observing that

$$\mathbb{F}[a_1\cdots a_n]=\mathbb{F}[a_1\cdots a_{n-1}][a_n]$$

and using induction. ■

When you have a field $\mathbb{F}$, $\mathbb{F}(a)$ denotes the smallest field which contains both $\mathbb{F}$ and a. When a is algebraic over $\mathbb{F}$, it follows that $\mathbb{F}(a)=\mathbb{F}[a]$. The latter is easier to think about because it just involves polynomials.

16.4.4 The Lindemannn Weierstrass Theorem And Vector Spaces

As another application of the abstract concept of vector spaces, there is an amazing theorem due to Weierstrass and Lindemannn.

Theorem 16.4.34 *Suppose $a_1,\cdots,a_n$ are algebraic numbers, roots of a polynomial with rational coefficients, and suppose $\alpha_1,\cdots,\alpha_n$ are distinct algebraic numbers. Then*

$$\sum_{i=1}^{n}a_ie^{\alpha_i}\neq 0$$

In other words, the $\{e^{\alpha_1},\cdots,e^{\alpha_n}\}$ are independent as vectors with field of scalars equal to the algebraic numbers.

A number is transcendental, as opposed to algebraic, if it is not a root of a polynomial which has integer (rational) coefficients. Most numbers are this way but it is hard to verify that specific numbers are transcendental. That π is transcendental follows from

$$e^0+e^{i\pi}=0.$$

By the above theorem, this could not happen if π were algebraic because then $i\pi$ would also be algebraic. Recall these algebraic numbers form a field and i is clearly algebraic, being a root of x^2+1. This fact about π was first proved by Lindemannn in 1882 and then the general theorem above was proved by Weierstrass in 1885. This fact that π is transcendental solved an old problem called squaring the circle which was to construct a square with the same area as a circle using a straight edge and compass. It can be shown that the fact π is transcendental implies this problem is impossible.[1]

[1] Gilbert, the librettist of the Savoy operas, may have heard about this great achievement. In Princess Ida which opened in 1884 he has the following lines. “As for fashion they forswear it, so the say - so they say; and the circle - they will square it some fine day some fine day.” Of course it had been proved impossible to do this a couple of years before.

16.5 Exercises

1. 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.

2. 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.

3. If you have 5 vectors in $\mathbb{F}^5$ and the vectors are linearly independent, can it always be concluded they span $\mathbb{F}^5$? Here $\mathbb{F}$ is an arbitrary field. Explain.

4. If you have 6 vectors in $\mathbb{F}^5$, is it possible they are linearly independent? Here $\mathbb{F}$ is an arbitrary field. Explain.

5. Show in any vector space, $\mathbf{0}$ is unique. This is done in the book. You do it yourself.

6. $\uparrow$In any vector space, show that if $\mathbf{x} + \mathbf{y} = \mathbf{0}$, then $\mathbf{y} = -\mathbf{x}$.This is done in the book. You do it yourself.

7. $\uparrow$Show that in any vector space, $0\mathbf{x} = \mathbf{0}$. That is, the scalar 0 times the vector $\mathbf{x}$ gives the vector $\mathbf{0}$. This is done in the book. You do it yourself.

8. $\uparrow$Show that in any vector space, $(-1)\,\mathbf{x} = -\mathbf{x}$.This is done in the book. You do it yourself.

9. Let X be a vector space and suppose $\{\mathbf{x}_1, \cdots, \mathbf{x}_k\}$ is a set of vectors from X. Show that $\mathbf{0}$ is in $\operatorname{span}(\mathbf{x}_1, \cdots, \mathbf{x}_k)$.This is done in the book. You do it yourself.

10. Let X consist of the real valued functions which are defined on an interval $[a, b]$. For $f, g \in X$, $f + g$ is the name of the function which satisfies $(f + g)(x) = f(x) + g(x)$ and for α a real number, $(\alpha f)(x) \equiv \alpha(f(x))$. Show this is a vector space with field of scalars equal to $\mathbb{R}$. Also explain why it cannot possibly be finite dimensional.

11. Let S be a nonempty set and let V denote the set of all functions which are defined on S and have values in W a vector space having field of scalars $\mathbb{F}$. Also define vector addition according to the usual rule, $(f + g)(s) \equiv f(s) + g(s)$ and scalar multiplication by $(\alpha f)(s) \equiv \alpha f(s)$. Show that V is a vector space with field of scalars $\mathbb{F}$.

12. Verify that any field $\mathbb{F}$ is a vector space with field of scalars $\mathbb{F}$. However, show that $\mathbb{R}$ is a vector space with field of scalars $\mathbb{Q}$.

13. Let $\mathbb{F}$ be a field and consider functions defined on $\{1, 2, \cdots, n\}$ having values in $\mathbb{F}$. Explain how, if V is the set of all such functions, V can be considered as $\mathbb{F}^n$.

14. Let V be the set of all functions defined on $\mathbb{N} \equiv \{1, 2, \cdots\}$ having values in a field $\mathbb{F}$ such that vector addition and scalar multiplication are defined by $(\mathbf{f} + \mathbf{g})(s) \equiv \mathbf{f}(s) + \mathbf{g}(s)$ and $(\alpha\mathbf{f})(s) \equiv \alpha\mathbf{f}(s)$ respectively, for $\mathbf{f}, \mathbf{g} \in V$ and $\alpha \in \mathbb{F}$. Explain how this is a vector space and show that for $\mathbf{e}_i$ given by

$$\mathbf{e}_i(k) \equiv \begin{cases} 1 \text{ if } i = k \\ 0 \text{ if } i \neq k \end{cases},$$

the vectors $\{\mathbf{e}_k\}_{k=1}^{\infty}$ are linearly independent.

15. Suppose, you have smooth functions $\{y_1, y_2, \cdots, y_n\}$ (all derivatives exist) defined on an interval $[a,b]$. Then the Wronskian of these functions is the determinant

$$W(y_1, \cdots, y_n)(x) = \det \begin{pmatrix} y_1(x) & \cdots & y_n(x) \\ y_1'(x) & \cdots & y_n'(x) \\ \vdots & & \vdots \\ y_1^{(n-1)}(x) & \cdots & y_n^{(n-1)}(x) \end{pmatrix}$$

Show that if $W(y_1, \cdots, y_n)(x) \neq 0$ for some x, then the functions are linearly independent.

16. Give an example of two functions, y_1, y_2 defined on $[-1,1]$ such that

$$W(y_1, y_2)(x) = 0$$

for all $x \in [-1,1]$ and yet $\{y_1, y_2\}$ is linearly independent.

17. Let the vectors be polynomials of degree no more than 3. Show that with the usual definitions of scalar multiplication and addition wherein, for $p(x)$ a polynomial, $(\alpha p)(x) = \alpha p(x)$ and for p, q polynomials $(p+q)(x) \equiv p(x) + q(x)$, this is a vector space.

18. In the previous problem show that a basis for the vector space is $\{1, x, x^2, x^3\}$.

19. Let V be the polynomials of degree no more than 3. Determine which of the following are bases for this vector space.

    (a) $\{x+1, x^3+x^2+2x, x^2+x, x^3+x^2+x\}$

    (b) $\{x^3+1, x^2+x, 2x^3+x^2, 2x^3-x^2-3x+1\}$

20. In the context of the above problem, consider polynomials

$$\{a_i x^3 + b_i x^2 + c_i x + d_i,\ i = 1,2,3,4\}$$

Show that this collection of polynomials is linearly independent on an interval $[a,b]$ if and only if

$$\begin{pmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{pmatrix}$$

is an invertible matrix.

21. Let the field of scalars be $\mathbb{Q}$, the rational numbers and let the vectors be of the form $a + b\sqrt{2}$ where a, b are rational numbers. Show that this collection of vectors is a vector space with field of scalars $\mathbb{Q}$ and give a basis for this vector space.

22. Suppose V is a finite dimensional vector space. Based on the exchange theorem above, it was shown that any two bases have the same number of vectors in them.

Give a different proof of this fact using the earlier material in the book. **Hint:** Suppose $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ and $\{\mathbf{y}_1, \cdots, \mathbf{y}_m\}$ are two bases with $m < n$. Then define $\phi : \mathbb{F}^n \mapsto V$ and $\psi : \mathbb{F}^m \mapsto V$ by $\phi(\mathbf{a}) \equiv \sum_{k=1}^n a_k \mathbf{x}_k$, and $\psi(\mathbf{b}) \equiv \sum_{j=1}^m b_j \mathbf{y}_j$. Consider the linear transformation, $\psi^{-1} \circ \phi$. Argue it is a one to one and onto mapping from $\mathbb{F}^n$ to $\mathbb{F}^m$. Now consider a matrix of this linear transformation and its row reduced echelon form.

23. This and the following problems will present most of a differential equations course. To begin with, consider the scalar initial value problem

$$y' = ay,\ y(t_0) = y_0$$

When a is real, show the unique solution to this problem is $y = y_0 e^{a(t-t_0)}$. Next suppose

$$y' = (a + ib) y,\ y(t_0) = y_0 \tag{16.8}$$

where $y(t) = u(t) + iv(t)$. Show there exists a unique solution and it is

$$y(t) = y_0 e^{a(t-t_0)} \left(\cos b(t-t_0) + i \sin b(t-t_0)\right) \equiv e^{(a+ib)(t-t_0)} y_0. \tag{16.9}$$

Next show that for a real or complex there exists a unique solution to the initial value problem

$$y' = ay + f,\ y(t_0) = y_0$$

and it is given by

$$y(t) = e^{a(t-t_0)} y_0 + e^{at} \int_{t_0}^{t} e^{-as} f(s)\, ds.$$

Hint: For the first part write as $y' - ay = 0$ and multiply both sides by e^{-at}. Then explain why you get

$$\frac{d}{dt}\left(e^{-at} y(t)\right) = 0,\ y(t_0) = 0.$$

Now you finish the argument. To show uniqueness in the second part, suppose

$$y' = (a + ib) y,\ y(0) = 0$$

and verify this requires $y(t) = 0$. To do this, note

$$\bar{y}' = (a - ib)\bar{y},\ \bar{y}(0) = 0$$

and that

$$\begin{aligned} \frac{d}{dt}|y(t)|^2 &= y'(t)\bar{y}(t) + \bar{y}'(t) y(t) = (a+ib) y(t)\bar{y}(t) + (a-ib)\bar{y}(t) y(t) \\ &= 2a|y(t)|^2,\ |y|^2(t_0) = 0 \end{aligned}$$

Thus from the first part $|y(t)|^2 = 0e^{-2at} = 0$. Finally observe by a simple computation that 16.8 is solved by 16.9. For the last part, write the equation as

$$y' - ay = f$$

and multiply both sides by e^{-at} and then integrate from t_0 to t using the initial condition.

24. ↑Now consider A an $n\times n$ matrix. By Schur's theorem there exists unitary Q such that
$$Q^{-1}AQ=T$$
where T is upper triangular. Now consider the first order initial value problem
$$\mathbf{x}'=A\mathbf{x},\ \mathbf{x}(t_0)=\mathbf{x}_0.$$
Show there exists a unique solution to this first order system. **Hint:** Let $\mathbf{y}=Q^{-1}\mathbf{x}$ and so the system becomes
$$\mathbf{y}'=T\mathbf{y},\ \mathbf{y}(t_0)=Q^{-1}\mathbf{x}_0 \tag{16.10}$$
Now letting $\mathbf{y}=(y_1,\cdots,y_n)^T$, the bottom equation becomes
$$y_n'=t_{nn}y_n,\ y_n(t_0)=\left(Q^{-1}\mathbf{x}_0\right)_n.$$
Then use the solution you get in this to get the solution to the initial value problem which occurs one level up, namely
$$y_{n-1}'\ =t_{(n-1)(n-1)}y_{n-1}+t_{(n-1)n}y_n,\ y_{n-1}(t_0)=\left(Q^{-1}\mathbf{x}_0\right)_{n-1}$$
Continue doing this to obtain a unique solution to 16.10.

25. ↑Now suppose $\Phi(t)$ is an $n\times n$ matrix of the form
$$\Phi(t)=\begin{pmatrix}\mathbf{x}_1(t) & \cdots & \mathbf{x}_n(t)\end{pmatrix} \tag{16.11}$$
where $\mathbf{x}_k'(t)=A\mathbf{x}_k(t)$. Explain why
$$\Phi'(t)=A\Phi(t)$$
if and only if $\Phi(t)$ is given in the form of 16.11. Also explain why if $\mathbf{c}\in\mathbb{F}^n,\mathbf{y}(t)\equiv\Phi(t)\mathbf{c}$ solves the equation
$$\mathbf{y}'(t)=A\mathbf{y}(t).$$

26. ↑In the above problem, consider the question whether all solutions to
$$\mathbf{x}'=A\mathbf{x} \tag{16.12}$$
are obtained in the form $\Phi(t)\mathbf{c}$ for some choice of $\mathbf{c}\in\mathbb{F}^n$. In other words, is the general solution to this equation $\Phi(t)\mathbf{c}$ for $\mathbf{c}\in\mathbb{F}^n$? Prove the following theorem using linear algebra.

Theorem 16.5.1 *Suppose $\Phi(t)$ is an $n\times n$ matrix which satisfies*
$$\Phi'(t)=A\Phi(t).$$
Then the general solution to 16.12 is $\Phi(t)\mathbf{c}$ if and only if $\Phi(t)^{-1}$ exists for some t. Furthermore, if $\Phi'(t)=A\Phi(t)$, then either $\Phi(t)^{-1}$ exists for all t or $\Phi(t)^{-1}$ never exists for any t.

($\det(\Phi(t))$ is called the Wronskian and this theorem is sometimes called the Wronskian alternative.)

Hint: Suppose first the general solution is of the form $\Phi(t)\mathbf{c}$ where $\mathbf{c}$ is an arbitrary constant vector in $\mathbb{F}^n$. You need to verify $\Phi(t)^{-1}$ exists for some t. In fact, show $\Phi(t)^{-1}$ exists for every t. Suppose then that $\Phi(t_0)^{-1}$ does not exist. Explain why there exists $\mathbf{c} \in \mathbb{F}^n$ such that there is no solution $\mathbf{x}$ to $\mathbf{c} = \Phi(t_0)\mathbf{x}$. By the existence part of Problem 24 there exists a solution to $\mathbf{x}' = A\mathbf{x}$, $\mathbf{x}(t_0) = \mathbf{c}$, but this cannot be in the form $\Phi(t)\mathbf{c}$. Thus for every t, $\Phi(t)^{-1}$ exists. Next suppose for some $t_0, \Phi(t_0)^{-1}$ exists. Let $\mathbf{z}' = A\mathbf{z}$ and choose $\mathbf{c}$ such that $\mathbf{z}(t_0) = \Phi(t_0)\mathbf{c}$. Then both $\mathbf{z}(t), \Phi(t)\mathbf{c}$ solve

$$\mathbf{x}' = A\mathbf{x},\ \mathbf{x}(t_0) = \mathbf{z}(t_0)$$

Apply uniqueness to conclude $\mathbf{z} = \Phi(t)\mathbf{c}$. Finally, consider that $\Phi(t)\mathbf{c}$ for $\mathbf{c} \in \mathbb{F}^n$ either is the general solution or it is not the general solution. If it is, then $\Phi(t)^{-1}$ exists for all t. If it is not, then $\Phi(t)^{-1}$ cannot exist for any t from what was just shown.

27. ↑Let $\Phi'(t) = A\Phi(t)$. Then $\Phi(t)$ is called a fundamental matrix if $\Phi(t)^{-1}$ exists for all t. Show there exists a unique solution to the equation

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f},\ \mathbf{x}(t_0) = \mathbf{x}_0 \tag{16.13}$$

and it is given by the formula

$$\mathbf{x}(t) = \Phi(t)\Phi(t_0)^{-1}\mathbf{x}_0 + \Phi(t)\int_{t_0}^{t}\Phi(s)^{-1}\mathbf{f}(s)\,ds$$

Now these few problems have done virtually everything of significance in an entire undergraduate differential equations course, illustrating the superiority of linear algebra. The above formula is called the variation of constants formula.

Hint: Uniqueness is easy. If $\mathbf{x}_1, \mathbf{x}_2$ are two solutions then let $\mathbf{u}(t) = \mathbf{x}_1(t) - \mathbf{x}_2(t)$ and argue $\mathbf{u}' = A\mathbf{u}, \mathbf{u}(t_0) = \mathbf{0}$. Then use Problem 24. To verify there exists a solution, you could just differentiate the above formula using the fundamental theorem of calculus and verify it works. Another way is to assume the solution in the form

$$\mathbf{x}(t) = \Phi(t)\mathbf{c}(t)$$

and find $\mathbf{c}(t)$ to make it all work out. This is called the method of variation of parameters.

28. ↑Show there exists a special Φ such that $\Phi'(t) = A\Phi(t)$, $\Phi(0) = I$, and $\Phi(t)^{-1}$ exists for all t. Show using uniqueness that

$$\Phi(-t) = \Phi(t)^{-1}$$

and that for all $t, s \in \mathbb{R}$

$$\Phi(t+s) = \Phi(t)\Phi(s)$$

Explain why with this special Φ, the solution to 16.13 can be written as

$$\mathbf{x}(t) = \Phi(t - t_0)\mathbf{x}_0 + \int_{t_0}^{t}\Phi(t - s)\mathbf{f}(s)\,ds.$$

Hint: Let $\Phi(t)$ be such that the j^{th} column is $\mathbf{x}_j(t)$ where

$$\mathbf{x}_j' = A\mathbf{x}_j,\ \mathbf{x}_j(0) = \mathbf{e}_j.$$

Use uniqueness as required.

29. *Using the Lindemann Weierstrass theorem show that if σ is an algebraic number $\sin\sigma, \cos\sigma, \ln\sigma$, and e are all transcendental. **Hint:** Observe, that

$$ee^{-1} + (-1)e^0 = 0,\ 1e^{\ln(\sigma)} + (-1)\sigma e^0 = 0,$$

$$\frac{1}{2i}e^{i\sigma} - \frac{1}{2i}e^{-i\sigma} + (-1)\sin(\sigma)e^0 = 0.$$

Chapter 17

Inner Product Spaces

17.1 Basic Definitions And Examples

An inner product space V is a vector space which also has an inner product. It is usually assumed, when considering inner product spaces that the field of scalars is either $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. This terminology has already been considered in the context of $\mathbb{F}^n$. In this section, it will be assumed that the field of scalars is $\mathbb{C}$, the complex numbers, unless specified to be something else. An inner product is a mapping $\langle \cdot, \cdot \rangle : V \times V \mapsto \mathbb{C}$ which satisfies the following axioms.

Axioms For Inner Product

1. $\langle \mathbf{u}, \mathbf{v} \rangle \in \mathbb{C}$, $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$.
2. If a, b are numbers and $\mathbf{u}, \mathbf{v}, \mathbf{z}$ are vectors then $\langle (a\mathbf{u} + b\mathbf{v}), \mathbf{z} \rangle = a \langle \mathbf{u}, \mathbf{z} \rangle + b \langle \mathbf{v}, \mathbf{z} \rangle$.
3. $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$ and it equals 0 if and only if $\mathbf{u} = \mathbf{0}$.

Note this implies $\langle \mathbf{x}, \alpha \mathbf{y} \rangle = \overline{\alpha} \langle \mathbf{x}, \mathbf{y} \rangle$ because

$$\langle \mathbf{x}, \alpha \mathbf{y} \rangle = \overline{\langle \alpha \mathbf{y}, \mathbf{x} \rangle} = \overline{\alpha \langle \mathbf{y}, \mathbf{x} \rangle} = \overline{\alpha} \langle \mathbf{x}, \mathbf{y} \rangle$$

Example 17.1.1 *Let V be the continuous complex valued functions defined on a finite closed interval I. Define an inner product as follows.*

$$\langle f, g \rangle \equiv \int_I f(x) \overline{g(x)} p(x) \, dx$$

where $p(x)$ some function which is strictly positive on the closed interval I. It is understood in writing this that

$$\int_I f(x) + ig(x) \, dx \equiv \int_I f(x) \, dx + i \int_I g(x) \, dx$$

Then with this convention, the usual calculus theorems hold about evaluating integrals using the fundamental theorem of calculus and so forth. You simply apply these theorems to the real and imaginary parts of a complex valued function.

Example 17.1.2 *Let V be the polynomials of degree at most n which are defined on a closed interval I and let $\{x_0, x_1, \cdots, x_n\}$ be $n+1$ distinct points in I. Then define*

$$\langle f, g\rangle \equiv \sum_{k=0}^{n} f(x_k)\overline{g(x_k)}$$

This last example clearly satisfies all the axioms for an inner product except for the one which says that $\langle \mathbf{u}, \mathbf{u}\rangle = 0$ if and only if $\mathbf{u} = \mathbf{0}$. Suppose then that $\langle f, f\rangle = 0$. Then f must vanish at $n+1$ distinct points but f is a polynomial of degree n. Therefore, it has at most n zeros unless it is identically equal to 0. Hence the second case holds and so f equals 0.

Example 17.1.3 *Let V be any complex vector space and let $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ be a basis.* ***Decree*** *that*

$$\langle \mathbf{v}_i, \mathbf{v}_j\rangle = \delta_{ij}.$$

Then define

$$\left\langle \sum_{j=1}^{n} c_j \mathbf{v}_j, \sum_{k=1}^{n} d_k \mathbf{v}_k \right\rangle \equiv \sum_{j,k} c_j \overline{d_k} \langle \mathbf{v}_j, \mathbf{v}_k\rangle = \sum_{k=1}^{n} c_k \overline{d_k}$$

This makes the complex vector space into an inner product space.

Example 17.1.4 *Let V consist of sequences $\mathbf{a} = \{a_k\}_{k=1}^{\infty}$, $a_k \in \mathbb{C}$, with the property that*

$$\sum_{k=1}^{\infty} |a_k|^2 < \infty$$

and the inner product is then defined as

$$\langle \mathbf{a}, \mathbf{b}\rangle \equiv \sum_{k=1}^{\infty} a_k \overline{b_k}$$

All of the axioms of the inner product are obvious for this example except the most basic one which says that the inner product has values in $\mathbb{C}$. Why does the above sum even converge? It converges from a comparison test.

$$\left|a_k \overline{b_k}\right| \leq \frac{|a_k|^2}{2} + \frac{|b_k|^2}{2}$$

and by assumption,

$$\sum_{k=1}^{\infty} \left(\frac{|a_k|^2}{2} + \frac{|b_k|^2}{2} \right) < \infty$$

and therefore, the given sum which defines the inner product is absolutely convergent. Therefore, thanks to completeness of $\mathbb{C}$ this sum also converges. This fact should be familiar to anyone who has had a calculus class in the context that the sequences are real valued. The case where they are complex valued follows right away from a consideration of real and imaginary parts.

By far the most important example of an inner product space is $L^2(\Omega)$, the space of Lebesgue measurable square integrable functions defined on Ω. However, this is a book on algebra, not analysis, so this example will be ignored.

17.1.1 The Cauchy Schwarz Inequality And Norms

The most fundamental theorem relative to inner products is the Cauchy Schwarz inequality.

Theorem 17.1.5 *(Cauchy Schwarz)The following inequality holds for* $\mathbf{x}$ *and* $\mathbf{y} \in V$, *an inner product space.*

$$|\langle \mathbf{x},\mathbf{y}\rangle| \leq \langle \mathbf{x},\mathbf{x}\rangle^{1/2} \langle \mathbf{y},\mathbf{y}\rangle^{1/2} \tag{17.1}$$

Equality holds in this inequality if and only if one vector is a multiple of the other.

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

$$\theta \langle \mathbf{x},\mathbf{y}\rangle = |\langle \mathbf{x},\mathbf{y}\rangle|$$

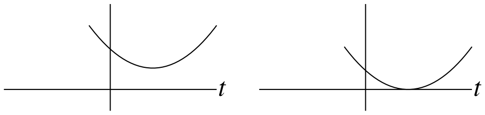
Consider $p(t) \equiv \left\langle \mathbf{x} + \overline{\theta} t \mathbf{y}, \mathbf{x} + t\overline{\theta}\mathbf{y} \right\rangle$ where $t \in \mathbb{R}$. Then from the above list of properties of the dot product,

$$\begin{aligned} 0 &\leq p(t) = \langle \mathbf{x},\mathbf{x}\rangle + t\theta \langle \mathbf{x},\mathbf{y}\rangle + t\overline{\theta} \langle \mathbf{y},\mathbf{x}\rangle + t^2 \langle \mathbf{y},\mathbf{y}\rangle \\ &= \langle \mathbf{x},\mathbf{x}\rangle + t\theta \langle \mathbf{x},\mathbf{y}\rangle + t\overline{\theta \langle \mathbf{x},\mathbf{y}\rangle} + t^2 (\mathbf{y},\mathbf{y}) \\ &= \langle \mathbf{x},\mathbf{x}\rangle + 2t \operatorname{Re} \theta \langle \mathbf{x},\mathbf{y}\rangle + t^2 \langle \mathbf{y},\mathbf{y}\rangle \\ &= \langle \mathbf{x},\mathbf{x}\rangle + 2t |\langle \mathbf{x},\mathbf{y}\rangle| + t^2 \langle \mathbf{y},\mathbf{y}\rangle \end{aligned} \tag{17.2}$$

and this must hold for all $t \in \mathbb{R}$. Therefore, if $\langle \mathbf{y},\mathbf{y}\rangle = 0$ it must be the case that $|\langle \mathbf{x},\mathbf{y}\rangle| = 0$ also since otherwise the above inequality would be violated for large negative t. Therefore, in this case,

$$|\langle \mathbf{x},\mathbf{y}\rangle| \leq \langle \mathbf{x},\mathbf{x}\rangle^{1/2} \langle \mathbf{y},\mathbf{y}\rangle^{1/2}.$$

In the other case, if $\langle \mathbf{y},\mathbf{y}\rangle \neq 0$, then $p(t) \geq 0$ for all t means the graph of $y = p(t)$ is a parabola which opens up and it either has exactly one real zero in the case its vertex touches the t axis or it has no real zeros.



From the quadratic formula this happens exactly when

$$4|\langle \mathbf{x},\mathbf{y}\rangle|^2 - 4\langle \mathbf{x},\mathbf{x}\rangle \langle \mathbf{y},\mathbf{y}\rangle \leq 0$$

which is equivalent to 17.1.

It is clear from a computation that if one vector is a scalar multiple of the other that equality holds in 17.1. Conversely, suppose equality does hold. Then this is equivalent to saying $4|\langle \mathbf{x},\mathbf{y}\rangle|^2 - 4\langle \mathbf{x},\mathbf{x}\rangle \langle \mathbf{y},\mathbf{y}\rangle = 0$ and so from the quadratic formula, there exists one real zero to $p(t) = 0$. Call it t_0. Then

$$p(t_0) \equiv \left\langle \mathbf{x} + \overline{\theta} t_0 \mathbf{y}, \mathbf{x} + t_0 \overline{\theta} \mathbf{y} \right\rangle = \left|\mathbf{x} + \overline{\theta} t_0 \mathbf{y}\right|^2 = 0$$

and so $\mathbf{x} = -\overline{\theta} t_0 \mathbf{y}$. ■

So what does the Cauchy Schwarz inequality say in the above examples? In Example 17.1.1 it says that

$$\left|\int_I f(x) \overline{g(x)} p(x)\, dx\right| \leq \left(\int_I |f(x)|^2 p(x)\, dx\right)^{1/2} \left(\int_I |g(x)|^2 p(x)\, dx\right)^{1/2}$$

With the Cauchy Schwarz inequality, it is possible to obtain the triangle inequality. This is the inequality in the next theorem. First it is necessary to define the norm or length of a vector. This is what is in the next definition.

Definition 17.1.6 *Let V be an inner product space and let $\mathbf{z} \in V$. Then $|\mathbf{z}| \equiv \langle \mathbf{z}, \mathbf{z} \rangle^{1/2}$. $|\mathbf{z}|$ is called the norm of $\mathbf{z}$ and also the length of $\mathbf{z}$.*

With the definition of length of a vector, here are the main properties of length.

Theorem 17.1.7 *For length defined in Definition 17.1.6, the following hold.*

$$|\mathbf{z}| \geq 0 \text{ and } |\mathbf{z}| = 0 \text{ if and only if } \mathbf{z} = \mathbf{0} \tag{17.3}$$

$$\text{If } \alpha \text{ is a scalar, } |\alpha \mathbf{z}| = |\alpha| \, |\mathbf{z}| \tag{17.4}$$

$$|\mathbf{z} + \mathbf{w}| \leq |\mathbf{z}| + |\mathbf{w}| \, . \tag{17.5}$$

Proof: The first two claims are left as exercises. To establish the third,

$$\begin{aligned} |\mathbf{z}+\mathbf{w}|^2 &= \langle \mathbf{z}+\mathbf{w}, \mathbf{z}+\mathbf{w} \rangle \\ &= \langle \mathbf{z}, \mathbf{z} \rangle + \langle \mathbf{w}, \mathbf{w} \rangle + \langle \mathbf{w}, \mathbf{z} \rangle + \langle \mathbf{z}, \mathbf{w} \rangle \\ &= |\mathbf{z}|^2 + |\mathbf{w}|^2 + 2 \operatorname{Re} \langle \mathbf{w}, \mathbf{z} \rangle \\ &\leq |\mathbf{z}|^2 + |\mathbf{w}|^2 + 2 \left| \langle \mathbf{w}, \mathbf{z} \rangle \right| \\ &\leq |\mathbf{z}|^2 + |\mathbf{w}|^2 + 2 |\mathbf{w}| \, |\mathbf{z}| = (|\mathbf{z}| + |\mathbf{w}|)^2 . \blacksquare \end{aligned}$$

The properties 17.3 - 17.5 are the axioms for a norm. A vector space which has a norm is called a normed linear space or a normed vector space.

17.2 The Gram Schmidt Process

The Gram Schmidt process is also valid in an inner product space. If you have a linearly independent set of vectors, there is an orthonormal set of vectors which has the same span. Recall the definition of an orthonormal set. It is the same as before.

Definition 17.2.1 *Let V be an inner product space and let $\{\mathbf{u}_i\}$ be a collection of vectors. It is an orthonormal set if*

$$\langle \mathbf{u}_k, \mathbf{u}_j \rangle = \delta_{jk}.$$

As before, every orthonormal set of vectors is linearly independent. If

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

where $\{\mathbf{u}_k\}_{k=1}^n$ is an orthonormal set of vectors, why is each $c_k = 0$?

This is true because you can take the inner product of both sides with $\mathbf{u}_j$. Then

$$\left\langle \sum_{k=1}^{n} c_k \mathbf{u}_k, \mathbf{u}_j \right\rangle = \langle \mathbf{0}, \mathbf{u}_j \rangle \, .$$

The right side equals 0 because

$$\langle \mathbf{0},\mathbf{u}\rangle = \langle \mathbf{0}+\mathbf{0},\mathbf{u}\rangle = \langle \mathbf{0},\mathbf{u}\rangle + \langle \mathbf{0},\mathbf{u}\rangle$$

Subtracting $\langle \mathbf{0},\mathbf{u}\rangle$ from both sides shows that $\langle \mathbf{0},\mathbf{u}\rangle = 0$. Therefore, from the properties of the inner product,

$$0 = \langle \mathbf{0},\mathbf{u}_j\rangle = \left\langle \sum_{k=1}^{n} c_k \mathbf{u}_k, \mathbf{u}_j \right\rangle = \sum_{k=1}^{n} c_k \langle \mathbf{u}_k, \mathbf{u}_j\rangle = \sum_{k=1}^{n} c_k \delta_{kj} = c_j.$$

Since c_j was arbitrary, this verifies that an orthonormal set of vectors is linearly independent.

Now consider the Gram Schmidt process.

Lemma 17.2.2 *Let $\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ be a linearly independent subset of an inner product space V. Then there exists an orthonormal set of vectors $\{\mathbf{u}_1,\cdots,\mathbf{u}_n\}$ which has the property that for each $k \leq n$, span$(\mathbf{x}_1,\cdots,\mathbf{x}_k) = \text{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$.*

Proof: Let $\mathbf{u}_1 \equiv \mathbf{x}_1/|\mathbf{x}_1|$. Thus for $k=1$, $\text{span}(\mathbf{u}_1) = \text{span}(\mathbf{x}_1)$ and $\{\mathbf{u}_1\}$ is an orthonormal set. Now suppose for some $k<n$, $\mathbf{u}_1$, $\cdots$, $\mathbf{u}_k$ have been chosen such that $(\mathbf{u}_j,\mathbf{u}_l) = \delta_{jl}$ and $\text{span}(\mathbf{x}_1,\cdots,\mathbf{x}_k) = \text{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$. Then define

$$\mathbf{u}_{k+1} \equiv \frac{\mathbf{x}_{k+1} - \sum_{j=1}^{k} \langle \mathbf{x}_{k+1},\mathbf{u}_j\rangle \mathbf{u}_j}{\left|\mathbf{x}_{k+1} - \sum_{j=1}^{k} \langle \mathbf{x}_{k+1},\mathbf{u}_j\rangle \mathbf{u}_j\right|}, \tag{17.6}$$

where the denominator is not equal to zero because the $\mathbf{x}_j$ form a basis, and so

$$\mathbf{x}_{k+1} \notin \text{span}(\mathbf{x}_1,\cdots,\mathbf{x}_k) = \text{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)$$

Thus by induction,

$$\mathbf{u}_{k+1} \in \text{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k,\mathbf{x}_{k+1}) = \text{span}(\mathbf{x}_1,\cdots,\mathbf{x}_k,\mathbf{x}_{k+1}).$$

Also, $\mathbf{x}_{k+1} \in \text{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k,\mathbf{u}_{k+1})$ which is seen easily by solving 17.6 for $\mathbf{x}_{k+1}$ and it follows

$$\text{span}(\mathbf{x}_1,\cdots,\mathbf{x}_k,\mathbf{x}_{k+1}) = \text{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k,\mathbf{u}_{k+1}).$$

If $l \leq k$,

$$\begin{aligned}
\langle \mathbf{u}_{k+1},\mathbf{u}_l\rangle &= C\left(\langle \mathbf{x}_{k+1},\mathbf{u}_l\rangle - \sum_{j=1}^{k} \langle \mathbf{x}_{k+1},\mathbf{u}_j\rangle \langle \mathbf{u}_j,\mathbf{u}_l\rangle\right) \\
&= C\left(\langle \mathbf{x}_{k+1},\mathbf{u}_l\rangle - \sum_{j=1}^{k} \langle \mathbf{x}_{k+1},\mathbf{u}_j\rangle \delta_{lj}\right) \\
&= C\left(\langle \mathbf{x}_{k+1},\mathbf{u}_l\rangle - \langle \mathbf{x}_{k+1},\mathbf{u}_l\rangle\right) = 0.
\end{aligned}$$

The vectors, $\{\mathbf{u}_j\}_{j=1}^{n}$, generated in this way are therefore orthonormal because each vector has unit length. ■

As in the case of $\mathbb{F}^n$, if you have a finite dimensional subspace of an inner product space, you can begin with a basis and then apply the Gram Schmidt process above to obtain an orthonormal basis.

There is nothing wrong with the above algorithm, but when you use it, it tends to get pretty intricate and it is easy to get lost in the details. There is a way to simplify it to produce fewer steps using matrices. I will illustrate in the case of three vectors. Say $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\}$ is linearly independent and you wish to find an orthonormal set $\{\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3\}$ which has the same span such that $\operatorname{span}(\mathbf{u}_1,\cdots,\mathbf{u}_k)=\operatorname{span}(\mathbf{v}_1,\cdots,\mathbf{v}_k)$ for each $k=1,2,3$. Then you would have

$$\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix} R$$

where R is an upper triangular matrix. Then

$$\begin{aligned}\delta_{jk} = \langle \mathbf{v}_j,\mathbf{v}_k\rangle &= \left\langle \sum_{r=1}^{n} \mathbf{u}_r R_{rj}, \sum_{s=1}^{n} \mathbf{u}_s R_{sk} \right\rangle \\ &= \sum_{r,s} R_{rj} \langle \mathbf{u}_r,\mathbf{u}_s\rangle R_{sk}\end{aligned}$$

Let G be the matrix whose rs entry is $\langle \mathbf{u}_r,\mathbf{u}_s\rangle$. This is called the Grammian matrix. Then the above reduces to the following matrix equation.

$$I = R^T G R$$

Taking inverses of both sides yields

$$I = R^{-1} G^{-1} \left(R^T\right)^{-1}$$

Then it follows that

$$RR^T = G^{-1}. \tag{17.7}$$

Example 17.2.3 *Let the real inner product space V consist of the continuous real functions defined on $[0,1]$ with the inner product given by*

$$\langle f,g\rangle \equiv \int_0^1 f(x)\, g(x)\, dx$$

and consider the functions (vectors) $\{1,x,x^2,x^3\}$. Show this is a linearly independent set of vectors and obtain an orthonormal set of vectors having the same span.

First, why is this a linearly independent set of vectors? This follows easily from Problem 15 on Page 403. You consider the Wronskian of these functions.

$$\det \begin{pmatrix} 1 & x & x^2 & x^3 \\ 0 & 1 & 2x & 3x^2 \\ 0 & 0 & 2 & 6x \\ 0 & 0 & 0 & 6 \end{pmatrix} \neq 0.$$

Therefore, the vectors are linearly independent. Now following the above procedure with matrices, let

$$R = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ 0 & a_5 & a_6 & a_7 \\ 0 & 0 & a_8 & a_9 \\ 0 & 0 & 0 & a_{10} \end{pmatrix}$$

Also it is necessary to compute the Grammian.

$$\begin{pmatrix} \int_0^1 dx & \int_0^1 xdx & \int_0^1 x^2dx & \int_0^1 x^3dx \\ \int_0^1 xdx & \int_0^1 x^2dx & \int_0^1 x^3dx & \int_0^1 x^4dx \\ \int_0^1 x^2dx & \int_0^1 x^3dx & \int_0^1 x^4dx & \int_0^1 x^5dx \\ \int_0^1 x^3dx & \int_0^1 x^4dx & \int_0^1 x^5dx & \int_0^1 x^6dx \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{pmatrix}$$

You also need the inverse of this Grammian. However, a computer algebra system can provide this right away.

$$G^{-1} = \begin{pmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{pmatrix}$$

Now it just remains to find the a_i.

$$\begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ 0 & a_5 & a_6 & a_7 \\ 0 & 0 & a_8 & a_9 \\ 0 & 0 & 0 & a_{10} \end{pmatrix} \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ 0 & a_5 & a_6 & a_7 \\ 0 & 0 & a_8 & a_9 \\ 0 & 0 & 0 & a_{10} \end{pmatrix}^T =$$

$$\begin{pmatrix} a_1^2+a_2^2+a_3^2+a_4^2 & a_2a_5+a_3a_6+a_4a_7 & a_3a_8+a_4a_9 & a_4a_{10} \\ a_2a_5+a_3a_6+a_4a_7 & a_5^2+a_6^2+a_7^2 & a_6a_8+a_7a_9 & a_7a_{10} \\ a_3a_8+a_4a_9 & a_6a_8+a_7a_9 & a_8^2+a_9^2 & a_9a_{10} \\ a_4a_{10} & a_7a_{10} & a_9a_{10} & a_{10}^2 \end{pmatrix} =$$

$$\begin{pmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{pmatrix}$$

Thus you can take (There is more than one solution.)

$$a_{10} = \sqrt{2800} = 20\sqrt{7},\ a_9 = -30\sqrt{7}, a_8 = 6\sqrt{5},\ a_7 = 12\sqrt{7},\ a_6 = -6\sqrt{5}$$

$$a_5 = 2\sqrt{3},\ a_4 = -\sqrt{7},\ a_3 = \sqrt{5},\ a_2 = -\sqrt{3},\ a_1 = 1$$

Thus the desired orthonormal basis is given by

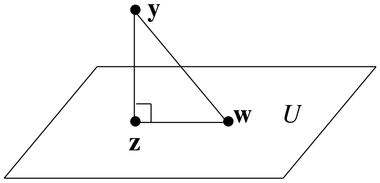
$$\begin{pmatrix} 1 & x & x^2 & x^3 \end{pmatrix} \begin{pmatrix} 1 & -\sqrt{3} & \sqrt{5} & -\sqrt{7} \\ 0 & 2\sqrt{3} & -6\sqrt{5} & 12\sqrt{7} \\ 0 & 0 & 6\sqrt{5} & -30\sqrt{7} \\ 0 & 0 & 0 & 20\sqrt{7} \end{pmatrix}$$

which yields

$$1,\ 2\sqrt{3}x - \sqrt{3},\ 6\sqrt{5}x^2 - 6\sqrt{5}x + \sqrt{5},\ 20\sqrt{7}x^3 - 30\sqrt{7}x^2 + 12\sqrt{7}x - \sqrt{7}$$

17.3 Approximation And Least Squares

Let V be an inner product space and let U be a finite dimensional subspace. Given $\mathbf{y} \in V$, how can you find the vector of U which is closest to $\mathbf{y}$ out of all such vectors in U? Does there even exist such a closest vector? The following picture is suggestive of the conclusion of the following lemma. It turns out that pictures like this do not mislead when you are dealing with inner product spaces in any number of dimensions.



Note that in the picture, $\mathbf{z}$ is a point in U and also $\mathbf{w}$ is a point of U. The following lemma states that for $\mathbf{z}$ to be closest to $\mathbf{y}$ out of all vectors in U, the vector from $\mathbf{z}$ to $\mathbf{y}$ should be perpendicular to any vector $\mathbf{w} \in U$. Since U is a subspace, this is the same as saying that the vector $\mathbf{y}-\mathbf{z}$ is perpendicular to the vector $\mathbf{w}-\mathbf{z}$ which is the situation illustrated by the above picture.

Lemma 17.3.1 *Suppose $\mathbf{y} \in V$, an inner product space and U is a subspace of V. Then*

$$|\mathbf{y}-\mathbf{z}| \leq |\mathbf{y}-\mathbf{w}|$$

for all $\mathbf{w} \in U$ if and only if, for all $\mathbf{w} \in U$,

$$\langle \mathbf{y}-\mathbf{z}, \mathbf{w} \rangle = 0. \tag{17.8}$$

Furthermore, there is at most one $\mathbf{z}$ which minimizes $|\mathbf{y}-\mathbf{w}|$ for $\mathbf{w} \in U$.

Proof: First suppose condition 17.8. Letting $\mathbf{w} \in U$, and using the properties of the inner product and the definition of the norm,

$$\begin{aligned} |\mathbf{y}-\mathbf{w}|^2 &= |\mathbf{y}-\mathbf{z}+\mathbf{z}-\mathbf{w}|^2 = |\mathbf{y}-\mathbf{z}|^2 + |\mathbf{z}-\mathbf{w}|^2 + 2\operatorname{Re}\langle \mathbf{y}-\mathbf{z}, \mathbf{z}-\mathbf{w} \rangle \\ &= |\mathbf{y}-\mathbf{z}|^2 + |\mathbf{z}-\mathbf{w}|^2 \end{aligned}$$

It follows then that $|\mathbf{y}-\mathbf{w}|$ is minimized when $\mathbf{w}=\mathbf{z}$.Next suppose $\mathbf{z}$ is a minimizer. Then pick $\mathbf{w} \in U$ and let $t \in \mathbb{R}$. Let $\theta \in \mathbb{C}$ be such that $|\theta| = 1$ and $\theta \langle \mathbf{y}-\mathbf{z}, \mathbf{w} \rangle = |\langle \mathbf{y}-\mathbf{z}, \mathbf{w} \rangle|$. Then

$$t \mapsto \left|(\mathbf{y}-\mathbf{z}) + t\overline{\theta}\mathbf{w}\right|^2$$

has a minimum when $t = 0$. But from the axioms of the inner product and definition of the norm, this function of t equals

$$\begin{aligned} & |\mathbf{y}-\mathbf{z}|^2 + |t\theta\mathbf{w}|^2 + 2t\operatorname{Re}\left\langle \mathbf{y}-\mathbf{z}, \overline{\theta}\mathbf{w} \right\rangle \\ = {} & |\mathbf{y}-\mathbf{z}|^2 + |t\theta\mathbf{w}|^2 + 2t\operatorname{Re}\theta\langle \mathbf{y}-\mathbf{z}, \mathbf{w} \rangle \\ = {} & |\mathbf{y}-\mathbf{z}|^2 + t^2|\mathbf{w}|^2 + 2t|\langle \mathbf{y}-\mathbf{z}, \mathbf{w} \rangle| \end{aligned}$$

Hence its derivative when $t = 0$ which is $2|\langle \mathbf{y}-\mathbf{z}, \mathbf{w} \rangle|$ equals 0.

Suppose now that $\mathbf{z}_i, i = 1,2$ both are minimizers. Then, as above,

$$|\mathbf{y}-\mathbf{z}_1|^2 = |\mathbf{y}-\mathbf{z}_2|^2 + |\mathbf{z}_1-\mathbf{z}_2|^2$$

and this is a contradiction unless $\mathbf{z}_1 = \mathbf{z}_2$ because $|\mathbf{y}-\mathbf{z}_1|^2 = |\mathbf{y}-\mathbf{z}_2|^2$. ■

This $\mathbf{z}$ described above is called the orthogonal projection of $\mathbf{y}$ onto U. The picture suggests why it is called this. The vector $\mathbf{y}-\mathbf{z}$ is perpendicular to the vectors in U.

With the above lemma, here is a theorem about existence, uniqueness and properties of a minimizer. The following theorem shows that the orthogonal projection is obtained by the linear transformation given by the formula

$$T\mathbf{y} \equiv \sum_{k=1}^{n} \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k$$

Note that the formula as well as the geometric interpretation suggested in the above picture shows that $T^2 = T$.

Theorem 17.3.2 *Let V be an inner product space and let U be an n dimensional subspace of V. Then if $\mathbf{y} \in V$ is given, there exists a unique $\mathbf{x} \in U$ such that*

$$|\mathbf{y}-\mathbf{x}| \leq |\mathbf{y}-\mathbf{w}|$$

for all $\mathbf{w} \in U$ and in addition, there is a formula for $\mathbf{x}$ in terms of any orthonormal basis for $U, \{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$,

$$\mathbf{x} = \sum_{k=1}^{n} \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k$$

Proof: By Lemma 17.3.1 there is at most one minimizer and it is characterized by the condition

$$\langle \mathbf{y}-\mathbf{x}, \mathbf{w} \rangle = 0$$

for all $\mathbf{w} \in U$. Let $\{\mathbf{u}_k\}_{k=1}^n$ be an orthonormal basis for U. By the Gram Schmidt process, Lemma 17.2.2, there exists such an orthonormal basis. Now it only remains to verify that

$$\left\langle \mathbf{y} - \sum_{k=1}^{n} \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k, \mathbf{w} \right\rangle = 0$$

for all $\mathbf{w}$. Since $\{\mathbf{u}_k\}_{k=1}^n$ is a basis, it suffices to verify that

$$\left\langle \mathbf{y} - \sum_{k=1}^{n} \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k, \mathbf{u}_l \right\rangle = 0, \text{ all } l = 1,2,\cdots,n$$

However, from the properties of the inner product,

$$\left\langle \mathbf{y} - \sum_{k=1}^{n} \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k, \mathbf{u}_l \right\rangle = \langle \mathbf{y}, \mathbf{u}_l \rangle - \sum_{k=1}^{n} \langle \mathbf{y}, \mathbf{u}_k \rangle \langle \mathbf{u}_k, \mathbf{u}_l \rangle$$

$$= \langle \mathbf{y}, \mathbf{u}_l \rangle - \sum_{k=1}^{n} \langle \mathbf{y}, \mathbf{u}_k \rangle \delta_{kl} = \langle \mathbf{y}, \mathbf{u}_l \rangle - \langle \mathbf{y}, \mathbf{u}_l \rangle = 0. \blacksquare$$

Note it follows that for any orthonormal basis $\{\mathbf{u}_k\}_{k=1}^n$, the same unique vector $\mathbf{x}$ is obtained as

$$\sum_{k=1}^{n} \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k$$

and it is the unique minimizer of $\mathbf{w} \mapsto |\mathbf{y} - \mathbf{w}|$. This is stated in the following corollary for the sake of emphasis.

Corollary 17.3.3 *Let V be an inner product space and let U be an n dimensional subspace of V. Then for $\mathbf{y}$ given in V, and $\{\mathbf{u}_k\}_{k=1}^n$, $\{\mathbf{v}_k\}_{k=1}^n$ two orthonormal bases for U,*

$$\sum_{k=1}^{n} \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k = \sum_{k=1}^{n} \langle \mathbf{y}, \mathbf{v}_k \rangle \mathbf{v}_k$$

The scalars $\langle \mathbf{y}, \mathbf{u}_k \rangle$ are called the Fourier coefficients.

Example 17.3.4 *Let V denote the real inner product space consisting of continuous functions defined on $[0,1]$ with the inner product*

$$\langle f, g \rangle \equiv \int_0^1 f(x) g(x) \, dx.$$

Let $U = \text{span}\left(1, x, x^2\right)$. It is desired to find the vector (function) in U which is closest to sin *in the norm determined by this inner product. Thus it is desired to minimize*

$$\left(\int_0^1 |\sin(x) - p(x)|^2 \, dx \right)^{1/2}$$

out of all functions p contained in U.

By Example 17.2.3, an orthonormal basis for U is

$$\left\{ 1,\ 2\sqrt{3}x - \sqrt{3},\ 6\sqrt{5}x^2 - 6\sqrt{5}x + \sqrt{5} \right\}$$

Then by Theorem 17.3.2, the closest vector (function) in U to sin can be computed as follows. First determine the Fourier coefficients.

$$\int_0^1 1 \sin(x) \, dx = 1 - \cos(1)$$

$$\int_0^1 \left(2\sqrt{3}x - \sqrt{3}\right) \sin(x) \, dx = \sqrt{3}\left(-\cos 1 + 2\sin 1 - 1\right)$$

$$\int_0^1 \left(6\sqrt{5}x^2 - 6\sqrt{5}x + \sqrt{5}\right) \sin(x) \, dx = \sqrt{5}\left(11\cos 1 + 6\sin 1 - 11\right)$$

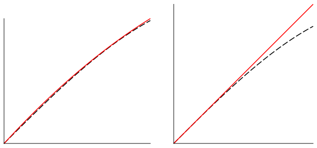
Next, from Theorem 17.3.2, the closest point to sin is

$$\begin{gathered}(1 - \cos(1)) + \left(\sqrt{3}(-\cos 1 + 2\sin 1 - 1)\right)\left(2\sqrt{3}x - \sqrt{3}\right) \\ + \left(\sqrt{5}(11\cos 1 + 6\sin 1 - 11)\right)\left(6\sqrt{5}x^2 - 6\sqrt{5}x + \sqrt{5}\right)\end{gathered}$$

Simplifying and approximating things like $\sin 1$, this yields the following for the approximation to $\sin x$.

$$-0.23546x^2 + 1.0913x - 7.4649\times 10^{-3}$$

If this is graphed along with $\sin x$ for $x\in[0,1]$ the result is as follows. One of the functions is represented by the solid line and the other by the dashed line.



There are two graphs. The left is the least squares approximation of the given function and the right is the result of using the Taylor series, both up to degree 2. You see the difference. The approximation using the inner product norm, called mean square approximation, attempts to approximate the given function on the whole interval while the Taylor series approximation is only good for small x.

17.4 Orthogonal Complement

Theorem 17.4.1 *Let V be a finite dimensional inner product space and let H be a subspace. Then $H^{\perp}$ defined by $H^{\perp}=\{y\in V:\langle y,h\rangle=0 \text{ for all } h\in H\}$ is also a subspace and*

$$V = H\oplus H^{\perp}$$

where the symbol means that every vector in V can be obtained as a sum of two vectors, one in H and the other in $H^{\perp}$ in exactly one way.

Proof: Let $\{\mathbf{u}_1,\cdots,\mathbf{u}_r\}$ be an orthonormal basis for H. Define the projection map

$$P\mathbf{v}\equiv\sum_{k=1}^{r}\langle\mathbf{v},\mathbf{u}_k\rangle\mathbf{u}_k$$

As shown in Theorem 17.3.2, $P\mathbf{v}$ is the unique point of H closest to $\mathbf{v}$ and $\langle\mathbf{v}-P\mathbf{v},\mathbf{h}\rangle=0$ for all $\mathbf{h}\in H$. Thus $\mathbf{v}=\mathbf{v}-P\mathbf{v}+P\mathbf{v}$ which shows that $V=H+H^{\perp}$. It remains to verify that there is a unique way to represent $\mathbf{v}$ as such a sum. Suppose then that

$$\mathbf{v}=\mathbf{h}+\mathbf{y}=\hat{\mathbf{h}}+\hat{\mathbf{y}}$$

where the $\mathbf{y}$ vectors are in $H^{\perp}$ and the $\mathbf{h}$ vectors in H. Then

$$\mathbf{h}-\hat{\mathbf{h}}=\hat{\mathbf{y}}-\mathbf{y}$$

Now $H^{\perp}$ is a subspace and so, taking the inner product of both sides with $\mathbf{h}-\hat{\mathbf{h}}$, you get

$$\left|\mathbf{h}-\hat{\mathbf{h}}\right|^2=0$$

and so $\mathbf{h}=\hat{\mathbf{h}}$ which then requires that $\mathbf{y}=\hat{\mathbf{y}}$. ■

Note in the above that it is routine from the formula to see that P is linear and also $P^2=P$. This is why it is called a projection.

17.5 Fourier Series

One of the most important applications of these ideas about approximation is to Fourier series. Much more can be said about these than will be presented here. However, Theorem 17.3.2 is a very useful framework for discussing these series.

For $x \in \mathbb{R}$, define e^{ix} by the following formula

$$e^{ix} \equiv \cos x + i \sin x$$

The reason for defining it this way is that $e^{i0} = 1$, and $\left(e^{ix}\right)' = ie^{ix}$ if you use this definition. Also it follows from the trigonometry identities that $e^{i(x+y)} = e^{ix}e^{iy}$. This is because

$$e^{ix}e^{iy} = (\cos x + i \sin x)(\cos y + i \sin y)$$

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

$$= \cos (x+y) + i \sin (x+y) = e^{i(x+y)}$$

In addition, $\overline{e^{ix}} = e^{-ix}$ because

$$\overline{e^{ix}} = \cos x - i \sin x = \cos (-x) + i \sin (-x) = e^{-ix}$$

It follows that the functions $\frac{1}{\sqrt{2\pi}}e^{ikx}$ for

$$k \in \{-n, -(n-1), \cdots, -1, 0, 1, \cdots, (n-1), n\} \equiv I_n$$

form an orthonormal set in V, the inner product space of continuous functions defined on $[-\pi, \pi]$ with the inner product given by

$$\int_{-\pi}^{\pi} f(x) \overline{g(x)} dx \equiv \langle f, g \rangle .$$

I will verify this now. Let k, l be two integers in $I_n, k \neq l$

$$\begin{aligned} \int_{-\pi}^{\pi} e^{ikx}\overline{e^{ilx}} dx &= \int_{-\pi}^{\pi} e^{i(k-l)x} dx = \frac{e^{i(k-l)x}}{i(k-l)} |_{-\pi}^{\pi} \\ &= \cos (k-l) \pi - \cos (k-l)(-\pi) = 0 \end{aligned}$$

Also

$$\int_{-\pi}^{\pi} \frac{1}{\sqrt{2\pi}} e^{ikx} \frac{1}{\sqrt{2\pi}} \overline{e^{ikx}} dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{0x} dx = 1.$$

Example 17.5.1 *Let V be the inner product space of piecewise continuous functions defined on $[-\pi, \pi]$ and let U be the span of the (vectors) $\left\{e^{ikx}\right\}_{k=-n}^{n}$. Let*

$$f(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ -1 \text{ if } x < 0 \end{cases}$$

Find the vector of U which is closest to f in the mean square sense (In the norm defined by this inner product).

First of all, you need to find the Fourier coefficients. Since $x \mapsto \cos(x)$ is even and $x \mapsto \sin(x)$ is odd,

$$\int_{-\pi}^{\pi} f(x) \frac{1}{\sqrt{2\pi}} e^{-ikx} dx = \frac{-i2}{\sqrt{2\pi}} \int_{0}^{\pi} \sin(-kx)\, dx$$

$$= \frac{-i\sqrt{2}}{\sqrt{\pi}} \frac{1-\cos(k\pi)}{k}, \quad \int_{-\pi}^{\pi} f(x)\, dx = 0.$$

Therefore, the best approximation is

$$\sum_{k=-n}^{n} \left(\frac{-i}{\sqrt{\pi}} \frac{1-\cos(k\pi)}{k} \right) \frac{\sqrt{2}}{\sqrt{2\pi}} e^{ikx}$$

The term for k can be combined with the term for $-k$ to yield

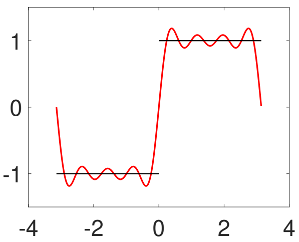
$$\left(\frac{-i}{\sqrt{\pi}} \frac{1-\cos(k\pi)}{k} \right) \frac{\sqrt{2}}{\sqrt{2\pi}} \left(e^{ikx} - e^{-ikx} \right) = \left(\frac{-i}{\sqrt{\pi}} \frac{1-\cos(k\pi)}{k} \right) \frac{\sqrt{2}}{\sqrt{2\pi}} (2i \sin kx)$$

$$= \left(\frac{2}{\pi} \frac{1-\cos(k\pi)}{k} \right) \sin kx$$

The terms when k is even are all 0. Therefore, the above reduces to

$$\frac{4}{\pi} \sum_{k=1}^{n} \left(\frac{1}{2k-1} \right) \sin(2k-1)x$$

In the case where $n = 4$, the graph of the function f being approximated along with the above function which is approximating it are as shown in the following picture on $[-\pi, \pi]$. This sum which



delivers the closest point in U will be denoted by $S_n f$. Note how the approximate function, closest in the mean square norm, is not equal to the given function at very many points but is trying to be close to it across the entire interval $[-\pi, \pi]$, except for a small interval centered at 0. You might try doing similar graphs on a calculator or computer in which you take larger and larger values of n. What will happen is that there will be a little bump near the point of discontinuity which won't go away, but this little bump will get thinner and thinner. The reason this must happen is roughly because the functions in the sum are continuous and the function being approximated is not. Therefore, convergence cannot take place uniformly. This is all I will say about these considerations because this is not an analysis book. See Problem 20 below for a discussion of where the Fourier series does converge at jumps.

17.6 The Discreet Fourier Transform

Everything done above for the Fourier series on $[-\pi,\pi]$ could have been done just as easily on $[0,2\pi]$ because all the functions are periodic of period 2π. Thus, for f a function defined on $[0,2\pi]$, you could consider the partial sums of the Fourier series on $[0,2\pi]$

$$\sum_{k=-n}^{n} a_k e^{ikx}$$

where

$$a_k = \frac{1}{2\pi}\int_0^{2\pi} f(y)\,e^{-iky}dy$$

and all the results of the last section continue to apply except now it is on the new interval. This is done to make the presentation of what follows easier to write.

The idea is that maybe you don't know what the function f is at all points, only at certain points

$$x_j = j\frac{2\pi}{N},\ j = 0,1,\cdots,N-1$$

Then instead of the integral given above, you could write a Riemann sum which approximates it. I will simply write the left Riemann sum. This yields the approximation b_k for a_k. Assuming f is continuous, the approximation would improve as $N\to\infty$.

$$b_k = \frac{1}{2\pi}\sum_{j=0}^{N-1} f\left(j\frac{2\pi}{N}\right) e^{-ik\left(j\frac{2\pi}{N}\right)}\frac{2\pi}{N} = \frac{1}{N}\sum_{j=0}^{N-1}\left(e^{-i\frac{2\pi}{N}}\right)^{kj} y_j$$

where y_j is defined to be the value of the function at x_j. This is called the discreet Fourier transform. In terms of matrix multiplication, let $\omega_N = e^{-i\frac{2\pi}{N}}$. Then

$$\begin{pmatrix} b_{-(N-1)} \\ \vdots \\ b_{-1} \\ b_0 \\ b_1 \\ \vdots \\ b_{N-1} \end{pmatrix} = \frac{1}{N}\begin{pmatrix} 1 & (\overline{\omega_N})^{N-1} & \cdots & (\overline{\omega_N})^{(N-1)(N-1)} \\ \vdots & \vdots & & \vdots \\ 1 & \overline{\omega_N} & \cdots & (\overline{\omega_N})^{(N-1)} \\ 1 & 1 & \cdots & 1 \\ 1 & \omega_N & \cdots & \omega_N^{(N-1)} \\ \vdots & \vdots & & \vdots \\ 1 & \omega_N^{N-1} & \cdots & \omega_N^{(N-1)(N-1)} \end{pmatrix}\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{pmatrix}$$

Thus you can find these approximate values by matrix multiplication.

Example 17.6.1 *Suppose you have the following table of values for the function f.*

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

Note that the above only uses the first four values.

In this case, $N = 4$ and so $\omega_N = e^{-i(\pi/2)} = -i$. Then the approximate Fourier coefficients are given by

$$\begin{pmatrix} b_{-3} \\ b_{-2} \\ b_{-1} \\ b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 1 & i^3 & i^6 & i^9 \\ 1 & i^2 & \left(i^2\right)^2 & \left(i^2\right)^3 \\ 1 & i & i^2 & i^3 \\ 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & (-i)^2 & (-i)^4 & (-i)^6 \\ 1 & (-i)^3 & (-i)^6 & (-i)^9 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ -1 \\ 1 \end{pmatrix}$$

$$= \frac{1}{4} \begin{pmatrix} 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \\ 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ -1 \\ 1 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 2-i \\ -3 \\ 2+i \\ 3 \\ 2-i \\ -3 \\ 2+i \end{pmatrix}$$

It follows that the approximate Fourier series for the given function is

$$\frac{1}{4} \left(\begin{array}{c} (2-i) e^{-3ix} + (-3) e^{-2ix} + (2+i) e^{-ix} + 3 + (2-i) e^{ix} \\ + (-3) e^{2ix} + (2+i) e^{3ix} \end{array} \right)$$

This simplifies to

$$\frac{3}{4} + 2\left(\frac{1}{2}\cos x - \frac{1}{4}\sin x\right) - \frac{3}{2}\cos(2x) + 2\left(\frac{1}{2}\cos 3x - \frac{1}{4}\sin 3x\right)$$

If you graph this, it will not do all that well in approximating some functions which have the given values at the given points. This is not surprising since only four points were considered. This is why in practice, people like to use a large number of points and when you do, the computations become sufficiently long that a special algorithm was developed for doing them. It is called the fast Fourier transform. So when you see this mentioned, this is what it is about, efficiently computing the discreet Fourier transform which can be thought of as a way to approximate the Fourier coefficients based on incomplete information for a given function.

17.7 Exercises

1. Verify that Examples 17.1.1 - 17.1.4 are each inner product spaces.

2. In each of the examples 17.1.1 - 17.1.4 write the Cauchy Schwarz inequality.

3. Verify 17.3 and 17.4.

4. Consider the Cauchy Schwarz inequality. Show that it still holds under the assumptions $\langle \mathbf{u},\mathbf{v}\rangle = \overline{\langle \mathbf{v},\mathbf{u}\rangle}, \langle (a\mathbf{u}+b\mathbf{v}),\mathbf{z}\rangle = a\langle \mathbf{u},\mathbf{z}\rangle + b\langle \mathbf{v},\mathbf{z}\rangle$, and $\langle \mathbf{u},\mathbf{u}\rangle \geq 0$. Thus it is not necessary to say that $\langle \mathbf{u},\mathbf{u}\rangle = 0$ only if $\mathbf{u}=\mathbf{0}$. It is enough to simply state that $\langle \mathbf{u},\mathbf{u}\rangle \geq 0$.

5. Consider the integers modulo a prime, $\mathbb{Z}_p$. This is a field of scalars. Now let the vector space be $(\mathbb{Z}_p)^n$ where $n \geq p$. Define now

$$\langle \mathbf{z},\mathbf{w}\rangle \equiv \sum_{i=1}^{n} z_i w_i$$

Does this satisfy the axioms of an inner product? Does the Cauchy Schwarz inequality hold for this $\langle\rangle$? Does the Cauchy Schwarz inequality even make any sense?

6. If you only know that $\langle \mathbf{u},\mathbf{u}\rangle \geq 0$ along with the other axioms of the inner product and if you define $|\mathbf{z}|$ the same way, how do the conclusions of Theorem 17.1.7 change?

7. In an inner product space, an open ball is the set

$$B(\mathbf{x},r) \equiv \{\mathbf{y} : |\mathbf{y}-\mathbf{x}| < r\}.$$

If $\mathbf{z} \in B(\mathbf{x},r)$, show there exists $\delta > 0$ such that $B(\mathbf{z},\delta) \subseteq B(\mathbf{x},r)$. In words, this says that an open ball is open. **Hint:** This depends on the triangle inequality.

8. Let V be the real inner product space consisting of continuous functions defined on $[-1,1]$ with the inner product given by

$$\int_{-1}^{1} f(x)\,g(x)\,dx$$

Show that $\{1,x,x^2\}$ are linearly independent and find an orthonormal basis for the span of these vectors.

9. A **regular Sturm Liouville problem** involves the differential equation for an unknown function of x which is denoted here by y,

$$\left(p(x)\,y'\right)' + (\lambda q(x) + r(x))\,y = 0,\ x \in [a,b]$$

and it is assumed that $p(t), q(t) > 0$ for any t along with boundary conditions,

$$\begin{aligned} C_1 y(a) + C_2 y'(a) &= 0 \\ C_3 y(b) + C_4 y'(b) &= 0 \end{aligned}$$

where

$$C_1^2 + C_2^2 > 0, \text{ and } C_3^2 + C_4^2 > 0.$$

There is an immense theory connected to these important problems. The constant λ is called an eigenvalue. Show that if y is a solution to the above problem corresponding to $\lambda = \lambda_1$ and if z is a solution corresponding to $\lambda = \lambda_2 \neq \lambda_1$, then

$$\int_a^b q(x)\,y(x)\,z(x)\,dx = 0. \tag{17.9}$$

Hint: Do something like this:

$$\left(p(x)y'\right)'z+(\lambda_1 q(x)+r(x))yz=0,$$

$$\left(p(x)z'\right)'y+(\lambda_2 q(x)+r(x))zy=0.$$

Now subtract and either use integration by parts or show

$$\left(p(x)y'\right)'z-\left(p(x)z'\right)'y=\left(\left(p(x)y'\right)z-\left(p(x)z'\right)y\right)'$$

and then integrate. Boundary conditions to show that $y'(a)z(a)-z'(a)y(a)=0$ and $y'(b)z(b)-z'(b)y(b)=0$.

10. Using the above problem or standard techniques of calculus, show that

$$\left\{\frac{\sqrt{2}}{\sqrt{\pi}}\sin(nx)\right\}_{n=1}^{\infty}$$

are orthonormal with respect to the inner product

$$\langle f,g\rangle=\int_0^{\pi}f(x)g(x)\,dx$$

Hint: If you want to use the above problem, show that $\sin(nx)$ is a solution to the boundary value problem

$$y''+n^2y=0,\ y(0)=y(\pi)=0$$

11. Find $S_5f(x)$ where $f(x)=x$ on $[-\pi,\pi]$. Then graph both $S_5f(x)$ and $f(x)$ if you have access to a system which will do a good job of it.

12. Find $S_5f(x)$ where $f(x)=|x|$ on $[-\pi,\pi]$. Then graph both $S_5f(x)$ and $f(x)$ if you have access to a system which will do a good job of it.

13. Find $S_5f(x)$ where $f(x)=x^2$ on $[-\pi,\pi]$. Then graph both $S_5f(x)$ and $f(x)$ if you have access to a system which will do a good job of it.

14. Let V be the set of real polynomials defined on $[0,1]$ which have degree at most 2. Make this into a real inner product space by defining

$$\langle f,g\rangle\equiv f(0)g(0)+f(1/2)g(1/2)+f(1)g(1)$$

Find an orthonormal basis and explain why this is an inner product.

15. Consider $\mathbb{R}^n$ with the following definition.

$$\langle\mathbf{x},\mathbf{y}\rangle\equiv\sum_{k=1}^{n}x_ky_kk$$

Does this define an inner product? If so, explain why and state the Cauchy Schwarz inequality in terms of sums.

16. From the above, for f a piecewise continuous function,

$$S_n f(x) = \frac{1}{2\pi} \sum_{k=-n}^{n} e^{ikx} \left(\int_{-\pi}^{\pi} f(y) e^{-iky} dy \right).$$

Show this can be written in the form

$$S_n f(x) = \int_{-\pi}^{\pi} f(y) D_n(x-y) \, dy$$

where

$$D_n(t) = \frac{1}{2\pi} \sum_{k=-n}^{n} e^{ikt}$$

This is called the Dirichlet kernel. Show that

$$D_n(t) = \frac{1}{2\pi} \frac{\sin(n+(1/2))t}{\sin(t/2)}$$

For V the vector space of piecewise continuous functions, define $S_n : V \mapsto V$ by

$$S_n f(x) = \int_{-\pi}^{\pi} f(y) D_n(x-y) \, dy.$$

Show that S_n is a linear transformation. (In fact, $S_n f$ is not just piecewise continuous but infinitely differentiable. Why?) Explain why $\int_{-\pi}^{\pi} D_n(t) \, dt = 1$. **Hint:** To obtain the formula, do the following.

$$\begin{aligned} e^{i(t/2)} D_n(t) &= \frac{1}{2\pi} \sum_{k=-n}^{n} e^{i(k+(1/2))t} \\ e^{i(-t/2)} D_n(t) &= \frac{1}{2\pi} \sum_{k=-n}^{n} e^{i(k-(1/2))t} \end{aligned}$$

Change the variable of summation in the bottom sum and then subtract and solve for $D_n(t)$.

17. ↑Let V be an inner product space and let U be a finite dimensional subspace with an orthonormal basis $\{\mathbf{u}_i\}_{i=1}^n$. If $\mathbf{y} \in V$, show

$$|\mathbf{y}|^2 \geq \sum_{k=1}^{n} |\langle \mathbf{y}, \mathbf{u}_k \rangle|^2$$

Now suppose that $\{\mathbf{u}_k\}_{k=1}^{\infty}$ is an orthonormal set of vectors of V. Explain why

$$\lim_{k \to \infty} \langle \mathbf{y}, \mathbf{u}_k \rangle = 0.$$

When applied to functions, this is a special case of the Riemann Lebesgue lemma.

18. ↑Let f be any piecewise continuous function which is bounded on $[-\pi, \pi]$. Show, using the above problem, that

$$\lim_{n \to \infty} \int_{-\pi}^{\pi} f(t) \sin(nt) \, dt = \lim_{n \to \infty} \int_{-\pi}^{\pi} f(t) \cos(nt) \, dt = 0$$

19. ↑*Let f be a function which is defined on $(-\pi,\pi]$. The 2π periodic extension is given by the formula $f(x+2\pi)=f(x)$. In the rest of this problem, f will refer to this 2π periodic extension. Assume that f is piecewise continuous, bounded, and also that the following limits exist

$$\lim_{y\to 0+}\frac{f(x+y)-f(x+)}{y},\ \lim_{y\to 0+}\frac{f(x-y)-f(x+)}{y}$$

Here it is assumed that

$$f(x+)\equiv\lim_{h\to 0+}f(x+h),\ \ f(x-)\equiv\lim_{h\to 0+}f(x-h)$$

both exist at every point. The above conditions rule out functions where the slope taken from either side becomes infinite. Justify the following assertions and eventually conclude that under these very reasonable conditions

$$\lim_{n\to\infty}S_nf(x)=(f(x+)+f(x-))/2$$

the mid point of the jump. In words, the Fourier series converges to the midpoint of the jump of the function.

$$S_nf(x)=\int_{-\pi}^{\pi}f(x-y)D_n(y)\,dy$$

$$\left|S_nf(x)-\frac{f(x+)+f(x-)}{2}\right|$$

$$=\left|\int_{-\pi}^{\pi}\left(f(x-y)-\frac{f(x+)+f(x-)}{2}\right)D_n(y)\,dy\right|$$

$$=\left|\begin{array}{c}\int_0^{\pi}f(x-y)D_n(y)\,dy+\int_0^{\pi}f(x+y)D_n(y)\,dy\\ -\int_0^{\pi}(f(x+)+f(x-))D_n(y)\,dy\end{array}\right|$$

$$\leq\left|\int_0^{\pi}(f(x-y)-f(x-))D_n(y)\,dy\right|+\left|\int_0^{\pi}(f(x+y)-f(x+))D_n(y)\,dy\right|$$

Now apply some trig. identities and use the result of Problem 18 to conclude that both of these terms must converge to 0.

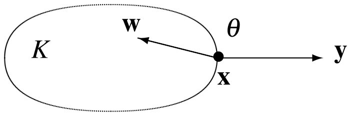
20. ↑Using the Fourier series obtained in Problem 11 and the result of Problem 19 above, find an interesting formula by examining where the Fourier series converges when $x=\pi/2$. Of course you can get many other interesting formulas in the same way. **Hint:** You should get

$$S_nf(x)=\sum_{k=1}^{n}\frac{2(-1)^{k+1}}{k}\sin(kx)$$

21. Let V be an inner product space and let K be a convex subset of V. This means that if $\mathbf{x},\mathbf{z}\in K$, then the line segment $\mathbf{x}+t(\mathbf{z}-\mathbf{x})=(1-t)\mathbf{x}+t\mathbf{z}$ is contained in K for all $t\in[0,1]$. Note that every subspace is a convex set. Let $\mathbf{y}\in V$ and let $\mathbf{x}\in K$. Show that $\mathbf{x}$ is the closest point to $\mathbf{y}$ out of all points in K if and only if for all $\mathbf{w}\in K$,

$$\operatorname{Re}\langle\mathbf{y}-\mathbf{x},\mathbf{w}-\mathbf{x}\rangle\leq 0.$$

In $\mathbb{R}^n$, a picture of the above situation where $\mathbf{x}$ is the closest point to $\mathbf{y}$ is as follows.



The condition of the above **variational inequality** is that the angle θ shown in the picture is larger than 90 degrees. Recall the geometric description of the dot product presented earlier. See Page 36.

22. Show that in any inner product space the parallelogram identity holds.

$$|\mathbf{x}+\mathbf{y}|^2 + |\mathbf{x}-\mathbf{y}|^2 = 2\,|\mathbf{x}|^2 + 2\,|\mathbf{y}|^2$$

Next show that in a real inner product space, the polarization identity holds.

$$\langle \mathbf{x},\mathbf{y}\rangle = \frac{1}{4}\left(|\mathbf{x}+\mathbf{y}|^2 - |\mathbf{x}-\mathbf{y}|^2\right).$$

23. *This problem is for those who know about Cauchy sequences and completeness of $\mathbb{F}^p$ and about closed sets. Suppose K is a closed nonempty convex subset of a finite dimensional subspace U of an inner product space V. Let $\mathbf{y} \in V$. Then show there exists a unique point $\mathbf{x} \in K$ which is closest to $\mathbf{y}$. **Hint:** Let

$$\lambda = \inf\{|\mathbf{y}-\mathbf{z}| : \mathbf{z} \in K\}$$

Let $\{\mathbf{x}_n\}$ be a minimizing sequence,

$$|\mathbf{y}-\mathbf{x}_n| \to \lambda.$$

Use the parallelogram identity in the above problem to show that $\{\mathbf{x}_n\}$ is a Cauchy sequence. Now let $\{\mathbf{u}_k\}_{k=1}^p$ be an orthonormal basis for U. Say

$$\mathbf{x}_n = \sum_{k=1}^p c_k^n \mathbf{u}_k$$

Verify that for $\mathbf{c}^n \equiv \left(c_1^n, \cdots, c_p^n\right) \in \mathbb{F}^p$

$$|\mathbf{x}_n - \mathbf{x}_m| = |\mathbf{c}^n - \mathbf{c}^m|_{\mathbb{F}^p}.$$

Now use completeness of $\mathbb{F}^p$ and the assumption that K is closed to get the existence of $\mathbf{x} \in K$ such that $|\mathbf{x}-\mathbf{y}| = \lambda$.

24. *Let K be a closed nonempty convex subset of a finite dimensional subspace U of a real inner product space V. (It is true for complex ones also.) For $\mathbf{x} \in V$, denote by $P\mathbf{x}$ the unique closest point to $\mathbf{x}$ in K. Verify that P is Lipschitz continuous with Lipschitz constant 1,

$$|P\mathbf{x} - P\mathbf{y}| \leq |\mathbf{x}-\mathbf{y}|.$$

Hint: Use Problem 21.

25. * This problem is for people who know about compactness. It is an analysis problem. If you have only had the usual undergraduate calculus course, don't waste your time with this problem. Suppose V is a finite dimensional normed linear space. Recall this means that there exists a norm $\|\cdot\|$ defined on V as described above,

$$\|\mathbf{v}\| \geq 0 \text{ equals } 0 \text{ if and only if } \mathbf{v} = \mathbf{0}$$

$$\|\mathbf{v}+\mathbf{u}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|, \ \|\alpha \mathbf{v}\| = |\alpha| \|\mathbf{v}\|.$$

Let $|\cdot|$ denote the norm which comes from Example 17.1.3, the inner product by decree. Show $|\cdot|$ and $\|\cdot\|$ are equivalent. That is, there exist constants $\delta, \Delta > 0$ such that for all $\mathbf{x} \in V$,

$$\delta |\mathbf{x}| \leq \|\mathbf{x}\| \leq \Delta |\mathbf{x}|.$$

Explain why every two norms on a finite dimensional vector space must be equivalent in the above sense.

26. Let A be an $n \times n$ matrix such that $A = A^*$. Verify that $\langle A\mathbf{x}, \mathbf{x} \rangle$ is always real. Then A is said to be nonnegative if $\langle A\mathbf{x}, \mathbf{x} \rangle \geq 0$ for every $\mathbf{x}$. Verify that $[\cdot, \cdot]$ given by $[\mathbf{x}, \mathbf{y}] \equiv \langle A\mathbf{x}, \mathbf{y} \rangle$ satisfies all the axioms of an inner product except for the one which says that $[\mathbf{x}, \mathbf{x}] = 0$ if and only if $\mathbf{x} = \mathbf{0}$. Also verify that the Cauchy Schwarz inequality holds for $[\cdot, \cdot]$.

27. Verify that for V equal to the space of continuous complex valued functions defined on an interval $[a, b]$, an inner product is

$$\langle f, g \rangle = \int_a^b f(x) \bar{g}(x) \, dx$$

If the functions are only assumed to be Riemann integrable, why is this no longer an inner product? In this last case, does the Cauchy Schwarz inequality still hold?

Chapter 18

Linear Transformations

18.1 Matrix Multiplication As A Linear Transformation

Definition 18.1.1 *Let V and W be two finite dimensional vector spaces. A function, L which maps V to W is called a linear transformation and $L \in \mathscr{L}(V,W)$ if for all scalars α and β, and vectors $\mathbf{v},\mathbf{w}$,*

$$L(\alpha\mathbf{v}+\beta\mathbf{w}) = \alpha L(\mathbf{v}) + \beta L(\mathbf{w}).$$

These linear transformations are also called homomorphisms. If one of them is one to one, it is called injective and if it is onto, it is called surjective. When a linear transformation is both one to one and onto, it is called bijective. ,

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 $L : \mathbb{F}^n \mapsto \mathbb{F}^m$ is given by

$$(L\mathbf{v})_i \equiv \sum_{j=1}^{n} a_{ij} v_j.$$

Here

$$\mathbf{v} \equiv \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{F}^n.$$

18.2 The Linear Maps as a Vector Space

In what follows I will often continue the practice of denoting vectors in bold face to distinguish them from scalars. However, this does not mean they are in $\mathbb{F}^n$.

Definition 18.2.1 *Given $L,M \in \mathscr{L}(V,W)$ define a new element of $\mathscr{L}(V,W)$, denoted by $L+M$ according to the rule*

$$(L+M)\mathbf{v} \equiv L\mathbf{v} + M\mathbf{v}.$$

For α a scalar and $L \in \mathscr{L}(V,W)$, define $\alpha L \in \mathscr{L}(V,W)$ by

$$\alpha L(\mathbf{v}) \equiv \alpha(L\mathbf{v}).$$

You should verify that all the axioms of a vector space hold for $\mathscr{L}(V,W)$ with the above definitions of vector addition and scalar multiplication. What about the dimension of $\mathscr{L}(V,W)$?

Lemma 18.2.2 *Let V and W be vector spaces and suppose $\{v_1,\cdots,v_n\}$ is a basis for V. Then if $L:V\to W$ is given by $Lv_k = w_k \in W$ and*

$$L\left(\sum_{k=1}^n a_k v_k\right) \equiv \sum_{k=1}^n a_k L v_k = \sum_{k=1}^n a_k w_k$$

then L is well defined and is in $\mathscr{L}(V,W)$. Also, if L,M are two linear transformations such that $Lv_k = Mv_k$ for all k, then $M = L$.

Proof: L is well defined on V because, since $\{v_1,\cdots,v_n\}$ is a basis, there is exactly one way to write a given vector of V as a linear combination. Next, observe that L is obviously linear from the definition. If L,M are equal on the basis, then an arbitrary vector in V is of the form $\sum_{k=1}^n a_k v_k$. Therefore,

$$L\left(\sum_{k=1}^n a_k v_k\right) = \sum_{k=1}^n a_k L v_k = \sum_{k=1}^n a_k M v_k = M\left(\sum_{k=1}^n a_k v_k\right)$$

and so $L = M$ because they give the same result for every vector in V. ■

The message is that when you define a linear transformation, it suffices to tell what it does to a basis.

Lemma 18.2.3 *Suppose $\theta \in \mathscr{L}(X,Y)$ where X,Y are vector spaces and θ is one to one and onto. Then if $\{\mathbf{y}_1,\cdots,\mathbf{y}_n\}$ is a basis for Y, it follows that*

$$\left\{\theta^{-1}\mathbf{y}_1,\cdots,\theta^{-1}\mathbf{y}_n\right\}$$

is a basis for X.

Proof: Let $\mathbf{x}_k = \theta^{-1}\mathbf{y}_k$. If $\sum_k c_k \mathbf{x}_k = 0$, then

$$\theta\left(\sum_k c_k \mathbf{x}_k\right) = \sum_k c_k \theta \mathbf{x}_k = \sum_k c_k \mathbf{y}_k = \mathbf{0}$$

and so, each $c_k = 0$. Hence $\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ is independent. If $\mathbf{x} \in X$, then $\theta\mathbf{x} \in Y$ so it equals an expression of the form

$$\sum_{k=1}^n c_k \mathbf{y}_k = \sum_k c_k \theta \mathbf{x}_k.$$

Hence

$$\theta\left(\mathbf{x} - \sum_k c_k \mathbf{x}_k\right) = \mathbf{0}$$

and so, since θ is one to one, $\mathbf{x} - \sum_k c_k \mathbf{x}_k = \mathbf{0}$ which shows that $\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ also spans and is therefore, a basis. ■

Theorem 18.2.4 *Let V and W be finite dimensional linear spaces of dimension n and m respectively Then* $\dim(\mathscr{L}(V,W)) = mn$.

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 L be a linear transformation. Then there are unique (since the $\mathbf{w}_j$ form a basis) scalars c_{ij} such that

$$L\mathbf{v}_k = \sum_{i=1}^{m} c_{ik}\mathbf{w}_i$$

Let C denote the $m\times n$ matrix whose ij^{th} entry is c_{ij}. Let θ be the mapping which takes $L\in\mathscr{L}(V,W)$ to this matrix which is just defined. Then θ is one to one because if $\theta L=\theta M$, then both L and M are equal on the basis for V. Therefore, $L=M$. Given such an $m\times n$ matrix $C=(c_{ij})$, use the above formula to define L. Thus θ is a one to one and onto map from $\mathscr{L}(V,W)$ to the space of $m\times n$ matrices $\mathscr{M}_{mn}$. It is also clear that θ is a linear map because if $\theta L=C$ and $\theta M=D$ and a,b scalars,

$$(aL+bM)\mathbf{v}_k = \sum_{i=1}^{m} ac_{ik}\mathbf{v}_i + \sum_{i=1}^{m} bd_{ik}\mathbf{v}_i = \sum_{i=1}^{m}(ac_{ik}+bd_{ik})\mathbf{v}_i$$

and so $\theta(aL+bM)=aC+bD$.

Obviously this space of matrices is of dimension mn, a basis consisting of E_{ij} the matrix which has a 1 in the ij^{th} position and zeros elsewhere. Therefore, $\left\{\theta^{-1}E_{ij}\right\}_{i.j}$ is a basis for $\mathscr{L}(V,W)$ by the above lemma. ■

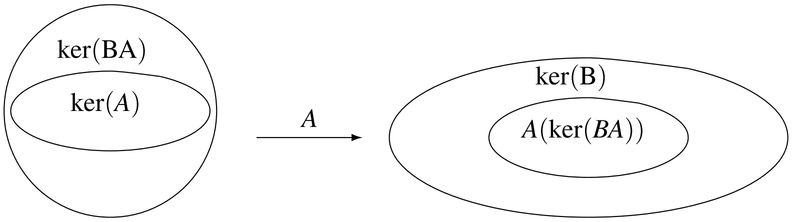
18.3 Eigenvalues And Eigenvectors Of Linear Transformations

Here is a very useful theorem due to Sylvester.

Theorem 18.3.1 *Let $A\in\mathscr{L}(V,W)$ and $B\in\mathscr{L}(W,U)$ where V,W,U are all vector spaces over a field $\mathbb{F}$. Suppose also that* $\ker(A)$ *and* $A(\ker(BA))$ *are finite dimensional subspaces. Then*

$$\dim(\ker(BA)) \leq \dim(\ker(B)) + \dim(\ker(A)).$$

Proof: If $\mathbf{x}\in\ker(BA)$, then $A\mathbf{x}\in\ker(B)$ and so $A(\ker(BA))\subseteq\ker(B)$. The following picture may help.



Now let $\{x_1,\cdots,x_n\}$ be a basis of $\ker(A)$ and let $\{Ay_1,\cdots,Ay_m\}$ be a basis for

$$A(\ker(BA)).$$

Take any $z \in \ker(BA)$. Then $Az = \sum_{i=1}^{m} a_i A y_i$ and so

$$A\left(z - \sum_{i=1}^{m} a_i y_i\right) = \mathbf{0}$$

which means $z - \sum_{i=1}^{m} a_i y_i \in \ker(A)$ and so there are scalars b_i such that

$$z - \sum_{i=1}^{m} a_i y_i = \sum_{j=1}^{n} b_i x_i.$$

It follows span $(x_1, \cdots, x_n, y_1, \cdots, y_m) \supseteq \ker(BA)$ and so by the first part, (See the picture.)

$$\dim(\ker(BA)) \leq n + m \leq \dim(\ker(A)) + \dim(\ker(B)) \ \blacksquare$$

Of course this result holds for any finite product of linear transformations by induction. One way this is quite useful is in the case where you have a finite product of linear transformations $\prod_{i=1}^{l} L_i$ all in $\mathscr{L}(V,V)$. Then

$$\dim\left(\ker \prod_{i=1}^{l} L_i\right) \leq \sum_{i=1}^{l} \dim(\ker L_i)$$

and so if you can find a linearly independent set of vectors in $\ker\left(\prod_{i=1}^{l} L_i\right)$ of size

$$\sum_{i=1}^{l} \dim(\ker L_i),$$

then it must be a basis for $\ker\left(\prod_{i=1}^{l} L_i\right)$.

Definition 18.3.2 *Let $\{V_i\}_{i=1}^{r}$ be subspaces of V. Then*

$$\sum_{i=1}^{r} V_i$$

denotes all sums of the form $\sum_{i=1}^{r} \mathbf{v}_i$ where $\mathbf{v}_i \in V_i$. If whenever

$$\sum_{i=1}^{r} \mathbf{v}_i = 0, \ \mathbf{v}_i \in V_i, \tag{18.1}$$

it follows that $\mathbf{v}_i = 0$ for each i, then 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.

Lemma 18.3.3 *If $V = V_1 \oplus \cdots \oplus V_r$ and if $\beta_i = \left\{\mathbf{v}_1^i, \cdots, \mathbf{v}_{m_i}^i\right\}$ is a basis for V_i, then a basis for V is $\{\beta_1, \cdots, \beta_r\}$.*

Proof: Suppose $\sum_{i=1}^{r} \sum_{j=1}^{m_i} c_{ij} \mathbf{v}_j^i = 0$. then since it is a direct sum, it follows for each i,

$$\sum_{j=1}^{m_i} c_{ij} \mathbf{v}_j^i = 0$$

and now since $\left\{\mathbf{v}_1^i, \cdots, \mathbf{v}_{m_i}^i\right\}$ is a basis, each $c_{ij} = 0$. $\blacksquare$

Here is a useful lemma.

Lemma 18.3.4 *Let L_i be in $\mathscr{L}(V,V)$ and suppose for $i \neq j, L_iL_j = L_jL_i$ and also L_i is one to one on* $\ker(L_j)$ *whenever $i \neq j$. Then*

$$\ker\left(\prod_{i=1}^{p} L_i\right) = \ker(L_1) \oplus + \cdots + \oplus \ker(L_p)$$

Here $\prod_{i=1}^{p} L_i$ is the product of all the linear transformations. A symbol like $\prod_{j\neq i} L_j$ is the product of all of them but L_i.

Proof: Note that since the operators commute, $L_j : \ker(L_i) \mapsto \ker(L_i)$. Here is why. If $L_i\mathbf{y} = \mathbf{0}$ so that $\mathbf{y} \in \ker(L_i)$, then

$$L_iL_j\mathbf{y} = L_jL_i\mathbf{y} = L_j\mathbf{0} = \mathbf{0}$$

and so $L_j : \ker(L_i) \mapsto \ker(L_i)$. Next observe that it is obvious that, since the operators commute,

$$\sum_{i=1}^{p} \ker(L_p) \subseteq \ker\left(\prod_{i=1}^{p} L_i\right)$$

Suppose

$$\sum_{i=1}^{p} \mathbf{v}_i = \mathbf{0},\ \mathbf{v}_i \in \ker(L_i),$$

but some $\mathbf{v}_i \neq \mathbf{0}$. Then do $\prod_{j\neq i} L_j$ to both sides. Since the linear transformations commute, this results in

$$\prod_{j\neq i} L_j\mathbf{v}_i = \mathbf{0}$$

which contradicts the assumption that these L_j are one to one and the observation that they map $\ker(L_i)$ to $\ker(L_i)$. Thus if

$$\sum_{i} \mathbf{v}_i = \mathbf{0},\ \mathbf{v}_i \in \ker(L_i)$$

then each $\mathbf{v}_i = \mathbf{0}$. It follows that

$$\ker(L_1) \oplus + \cdots + \oplus \ker(L_p) \subseteq \ker\left(\prod_{i=1}^{p} L_i\right) \tag{*}$$

From Sylvester's theorem and the observation about direct sums in Lemma 18.3.3,

$$\begin{aligned} \sum_{i=1}^{p} \dim(\ker(L_i)) &= \dim(\ker(L_1) \oplus + \cdots + \oplus \ker(L_p)) \\ &\leq \dim\left(\ker\left(\prod_{i=1}^{p} L_i\right)\right) \leq \sum_{i=1}^{p} \dim(\ker(L_i)) \end{aligned}$$

which implies all these are equal. Now in general, if W is a subspace of V, a finite dimensional vector space and the two have the same dimension, then $W = V$. This is because W has a basis and if $\mathbf{v}$ is not in the span of this basis, then $\mathbf{v}$ adjoined to the basis of W would be a linearly independent set so the dimension of V would then be strictly larger than the dimension of W. It follows from * that

$$\ker(L_1) \oplus + \cdots + \oplus \ker(L_p) = \ker\left(\prod_{i=1}^{p} L_i\right) \blacksquare$$

Here is a situation in which the above holds. $\ker(A-\lambda_i I)^r$ is sometimes called a generalized eigenspace. The following is an important result on generalized eigenspaces.

Theorem 18.3.5 *Let V be a vector space of dimension n and A a linear transformation and suppose $\{\lambda_1, \cdots, \lambda_k\}$ are distinct scalars. Define for $r_i \in \mathbb{N}$*

$$V_i = \ker(A-\lambda_i I)^{r_i} \tag{18.2}$$

Then

$$\ker\left(\prod_{i=1}^{p}(A-\lambda_i I)^{r_i}\right) = V_i \oplus \cdots \oplus V_p. \tag{18.3}$$

Proof: It is obvious the linear transformations $(A-\lambda_i I)^{r_i}$ commute. Now here is a claim.

Claim : Let $\mu \neq \lambda_i$, Then $(A-\mu I)^m : V_i \mapsto V_i$ and is one to one and onto for every $m \in \mathbb{N}$.

Proof: It is clear $(A-\mu I)^m$ maps V_i to V_i because if $\mathbf{v} \in V_i$ then $(A-\lambda_i I)^{r_i}\mathbf{v} = \mathbf{0}$. Consequently,

$$(A-\lambda_i I)^{r_i}(A-\mu I)^m \mathbf{v} = (A-\mu I)^m (A-\lambda_i I)^{r_i}\mathbf{v} = (A-\mu I)^m \mathbf{0} = \mathbf{0}$$

which shows that $(A-\mu I)^m \mathbf{v} \in V_i$.

It remains to verify $(A-\mu I)^m$ is one to one. This will be done by showing that $(A-\mu I)$ is one to one. Let $\mathbf{w} \in V_i$ and suppose $(A-\mu I)\mathbf{w} = 0$ so that $A\mathbf{w} = \mu\mathbf{w}$. Then for $m \equiv r_i$, $(A-\lambda_i I)^m \mathbf{w} = 0$ and so by the binomial theorem,

$$(\mu-\lambda_i)^m \mathbf{w} = \sum_{l=0}^{m}\binom{m}{l}(-\lambda_i)^{m-l}\mu^l \mathbf{w}$$

$$\sum_{l=0}^{m}\binom{m}{l}(-\lambda_i)^{m-l}A^l \mathbf{w} = (A-\lambda_i I)^m \mathbf{w} = \mathbf{0}.$$

Therefore, since $\mu \neq \lambda_i$, it follows $\mathbf{w} = 0$ and this verifies $(A-\mu I)$ is one to one. Thus $(A-\mu I)^m$ is also one to one on V_i. Letting $\{\mathbf{u}_1^i, \cdots, \mathbf{u}_{r_k}^i\}$ be a basis for V_i, it follows

$$\left\{(A-\mu I)^m \mathbf{u}_1^i, \cdots, (A-\mu I)^m \mathbf{u}_{r_k}^i\right\}$$

is also a basis and so $(A-\mu I)^m$ is also onto. The desired result now follows from Lemma 18.3.4. ■

Let V be a finite dimensional vector space with field of scalars $\mathbb{C}$. For example, it could be a subspace of $\mathbb{C}^n$. Also suppose $A \in \mathscr{L}(V,V)$. Does A have eigenvalues and eigenvectors just like the case where A is a $n \times n$ matrix?

Theorem 18.3.6 *Let V be a nonzero finite dimensional vector space of dimension n. Suppose also the field of scalars equals $\mathbb{C}$.*[1] *Suppose $A \in \mathscr{L}(V,V)$. Then there exists $\mathbf{v} \neq \mathbf{0}$ and $\lambda \in \mathbb{C}$ such that*

$$A\mathbf{v} = \lambda \mathbf{v}.$$

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 18.2.4 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 minimal polynomial and will be denoted here by $p(\lambda)$. By the fundamental theorem of algebra, $p(\lambda)$ is of the form

$$p(\lambda) = \prod_{k=1}^{m} (\lambda - \lambda_k).$$

where some of the λ_k might be repeated. Thus, since p has minimal degree,

$$\prod_{k=1}^{m} (A - \lambda_k I) = 0, \text{ but } \prod_{k=1}^{m-1} (A - \lambda_k I) \neq 0.$$

Therefore, there exists $\mathbf{u} \neq 0$ such that

$$\mathbf{v} \equiv \left(\prod_{k=1}^{m-1} (A - \lambda_k I) \right) (\mathbf{u}) \neq 0.$$

But then

$$(A - \lambda_m I)\mathbf{v} = (A - \lambda_m I) \left(\prod_{k=1}^{m-1} (A - \lambda_k I) \right) (\mathbf{u}) = \mathbf{0}. \blacksquare$$

As a corollary, it is good to mention that the minimal polynomial just discussed is unique.

Corollary 18.3.7 *Let $A \in \mathscr{L}(V,V)$ where V is an n dimensional vector space, the field of scalars being $\mathbb{F}$. Then there exists a polynomial $q(\lambda)$ having coefficients in $\mathbb{F}$ such that $q(A) = 0$. Letting $p(\lambda)$ be the monic polynomial having smallest degree such that $p(A) = 0$, it follows that $p(\lambda)$ is unique.*

Proof: The existence of $p(\lambda)$ follows from the above theorem. Suppose then that $p_1(\lambda)$ is another one. That is, it has minimal degree of all polynomials $q(\lambda)$ satisfying $q(A) = 0$ and is monic. Then by Lemma 16.4.3 there exists $r(\lambda)$ which is either equal to 0 or has degree smaller than that of $p(\lambda)$ and a polynomial $l(\lambda)$ such that

$$p_1(\lambda) = p(\lambda) l(\lambda) + r(\lambda)$$

[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.

By assumption, $r(A)=0$. Therefore, $r(\lambda)=0$. Also by assumption, $p_1(\lambda)$ and $p(\lambda)$ have the same degree and so $l(\lambda)$ is a scalar. Since $p_1(\lambda)$ and $p(\lambda)$ are both monic, it follows this scalar must equal 1. This shows uniqueness. ■

Corollary 18.3.8 *In the above theorem, each of the scalars λ_k has the property that there exists a nonzero $\mathbf{v}$ such that $(A-\lambda_i I)\mathbf{v}=0$. 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 $(A-\lambda_m I)$. Next suppose

$$(A-\mu I)\mathbf{v}=\mathbf{0}$$

for some μ and $\mathbf{v}\neq\mathbf{0}$. Then

$$\begin{aligned}
0 &= \prod_{k=1}^{m}(A-\lambda_k I)\mathbf{v}=\prod_{k=1}^{m-1}(A-\lambda_k I)\left(\overbrace{A\mathbf{v}}^{=\mu\mathbf{v}}-\lambda_m\mathbf{v}\right)\\
&= (\mu-\lambda_m)\left(\prod_{k=1}^{m-1}(A-\lambda_k I)\right)\mathbf{v}\\
&= (\mu-\lambda_m)\left(\prod_{k=1}^{m-2}(A-\lambda_k I)\right)(A\mathbf{v}-\lambda_{m-1}\mathbf{v})\\
&= (\mu-\lambda_m)(\mu-\lambda_{m-1})\left(\prod_{k=1}^{m-2}(A-\lambda_k I)\right)\mathbf{v}
\end{aligned}$$

continuing this way yields $=\prod_{k=1}^{m}(\mu-\lambda_k)\mathbf{v}$, a contradiction unless $\mu=\lambda_k$ for some k. ■

Therefore, these are eigenvectors and eigenvalues with the usual meaning. This leads to the following definition.

Definition 18.3.9 *For $A\in\mathscr{L}(V,V)$ where $\dim(V)=n$, the scalars, λ_k in the minimal polynomial,*

$$p(\lambda)=\prod_{k=1}^{m}(\lambda-\lambda_k)\equiv\prod_{k=1}^{p}(\lambda-\lambda_k)^{r_k}$$

are called the eigenvalues of A. In the last expression, λ_k is a repeated root which occurs r_k times. The collection of eigenvalues of A is denoted by $\sigma(A)$. The generalized eigenspaces are

$$\ker(A-\lambda_k I)^{r_k}\equiv V_k.$$

Theorem 18.3.10 *In the situation of the above definition,*

$$V=V_1\oplus\cdots\oplus V_p$$

That is, the vector space equals the direct sum of its generalized eigenspaces.

Proof: Since $V=\ker\left(\prod_{k=1}^{p}(A-\lambda_k I)^{r_k}\right)$, the conclusion follows from Theorem 18.3.5. ■

18.4 Block Diagonal Matrices

In this section the vector space will be $\mathbb{C}^n$ and the linear transformations will be those which result by multiplication by $n \times n$ matrices.

Definition 18.4.1 *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 18.4.2 *Let A be an $n \times n$ matrix. Letting $\lambda_1, \lambda_2, \cdots, \lambda_r$ be the distinct eigenvalues of A,arranged in some 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. 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, $\left\{\mathbf{u}_1^k, \cdots, \mathbf{u}_{r_k}^k\right\} = D_k$ constitute a basis for V_{λ_k}.

Proof: By Theorem 18.3.9 and Lemma 18.3.3,

$$\mathbb{C}^n = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_k}$$

and a basis for $\mathbb{C}^n$ is $\{D_1, \cdots, D_r\}$ where D_k is a basis for V_{λ_k}, $\ker(A - \lambda_k I)^{r_k}$.

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 $C_iB_i = I_{r_i \times r_i}$. Also, if $i \neq j$, then $C_iAB_j = 0$ the last claim holding because $A : V_{\lambda_j} \mapsto V_{\lambda_j}$ 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 since $C_iB_j = 0$ if $i \neq j$. 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 $C_{r_k}AB_{r_k}$ is an $r_k \times r_k$ matrix.

What about the eigenvalues of $C_{r_k}AB_{r_k}$? 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

$$(A - \lambda_k I)^{r_k} \mathbf{x} = (A - \mu I + (\mu - \lambda_k) I)^{r_k} \mathbf{x}$$

$$= \sum_{j=0}^{r_k} \begin{pmatrix} r_k \\ j \end{pmatrix} (\mu - \lambda_k)^{r_k - j} (A - \mu I)^j \mathbf{x} = (\mu - \lambda_k)^{r_k} \mathbf{x} \neq \mathbf{0}$$

contrary to the assumption that $\mathbf{x} \in V_{\lambda_k}$. Suppose then that $C_{r_k}AB_{r_k}\mathbf{x} = \lambda\mathbf{x}$ 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

$$S^{-1}A\mathbf{y} = S^{-1}AS \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{x} \\ \vdots \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \vdots \\ C_{r_k}AB_{r_k}\mathbf{x} \\ \vdots \\ \mathbf{0} \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{x} \\ \vdots \\ \mathbf{0} \end{pmatrix}$$

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 let $P_k = C_{r_k}AB_{r_k}$. ■

The above theorem contains a result which is of sufficient importance to state as a corollary.

Corollary 18.4.3 *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 13.2.11 on Page 309. 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 18.4.2 there exists S such that

$$S^{-1}AS = \begin{pmatrix} P_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_r \end{pmatrix}.$$

Therefore,

$$\begin{aligned} U^* SASU &= \begin{pmatrix} U_1^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_r^* \end{pmatrix} \begin{pmatrix} P_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_r \end{pmatrix} \begin{pmatrix} U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_r \end{pmatrix} \\ &= \begin{pmatrix} U_1^* P_1 U_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_r^* P_r U_r \end{pmatrix} = \begin{pmatrix} T_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T_r \end{pmatrix}. \end{aligned}$$

This proves most of the following corollary of Theorem 18.4.2.

Corollary 18.4.4 *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

$$m_k = \dim\left(\ker\left(A - \lambda_k I\right)^{r_k}\right)$$

where the minimal polynomial of A is

$$\prod_{k=1}^{p} \left(\lambda - \lambda_k\right)^{r_k}$$

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 observation 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

$$\prod_{k=1}^{r}(\lambda_k-\lambda)^{m_k}. \blacksquare$$

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 18.4.4 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

$$T_s=\begin{pmatrix}\alpha & \cdots & *\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \alpha\end{pmatrix} \tag{18.4}$$

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

[2] The S here is written as S^{-1} in the corollary.

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 $k \geq m_s$.

$$\begin{aligned} T_s^k &= (D+N)^k = \sum_{j=0}^{k} \begin{pmatrix} k \\ j \end{pmatrix} D^{k-j} N^j \\ &= \sum_{j=0}^{m_s} \begin{pmatrix} k \\ j \end{pmatrix} D^{k-j} N^j, \end{aligned} \tag{18.5}$$

the third equation holding because $N^{m_s} = 0$. Thus T_s^k is of the form

$$T_s^k = \begin{pmatrix} \alpha^k & \cdots & * \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha^k \end{pmatrix}.$$

Lemma 18.4.5 *Suppose T is of the form T_s described above in 18.4 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 18.5, it follows that for large k, and $j \leq m_s$,

$$\begin{pmatrix} k \\ j \end{pmatrix} \leq \frac{k(k-1)\cdots(k-m_s+1)}{m_s!}.$$

Therefore, letting C be the largest value of $\left| \left(N^j \right)_{pq} \right|$ for $0 \leq j \leq m_s$,

$$\left| \left(T^k \right)_{pq} \right| \leq m_s C \left(\frac{k(k-1)\cdots(k-m_s+1)}{m_s!} \right) |\alpha|^{k-m_s}$$

which converges to zero as $k \to \infty$. This is most easily seen by applying the ratio test to the series

$$\sum_{k=m_s}^{\infty} \left(\frac{k(k-1)\cdots(k-m_s+1)}{m_s!} \right) |\alpha|^{k-m_s}$$

and then noting that if a series converges, then the k^{th} term converges to zero. ■

18.5 The Matrix Of A Linear Transformation

To begin with, here is an easy lemma which relates the vectors in two vector spaces having the same dimension.

Lemma 18.5.1 *Let $q : V \to W$ be one to one, onto and linear. Then q maps any basis of V to a basis for W. Conversely, if q is linear and maps a basis to a basis, then it is one to one onto.*

Proof: Let $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ be a basis for V, why is $\{q\mathbf{v}_1,\cdots,q\mathbf{v}_n\}$ a basis for W? First consider why it is linearly independent. Suppose $\sum_{k=1}^n c_k q\mathbf{v}_k = \mathbf{0}$. Then $q\left(\sum_{k=1}^n c_k\mathbf{v}_k\right) = \mathbf{0}$ and since q is one to one, it follows that $\sum_{k=1}^n c_k\mathbf{v}_k = \mathbf{0}$ which requires each $c_k = 0$ because $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is independent. Next take $\mathbf{w}\in W$. Since q is onto, there exists $\mathbf{v}\in V$ such that $q\mathbf{v} = \mathbf{w}$. Since $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ is a basis, there are scalars c_k such that $q\left(\sum_{k=1}^n c_k\mathbf{v}_k\right) = q(\mathbf{v}) = \mathbf{w}$ and so $\mathbf{w} = \sum_{k=1}^n c_k q\mathbf{v}_k$ which is in the span$(q\mathbf{v}_1,\cdots,q\mathbf{v}_n)$. Therefore, $\{q\mathbf{v}_1,\cdots,q\mathbf{v}_n\}$ is a basis as claimed.

Suppose now that $q\mathbf{v}_i = \mathbf{w}_i$ where $\{\mathbf{v}_1,\cdots,\mathbf{v}_n\}$ and $\{\mathbf{w}_1,\cdots,\mathbf{w}_n\}$ are bases for V and W. If $q\left(\sum_{k=1}^n c_k\mathbf{v}_k\right) = \mathbf{0}$, then $\sum_{k=1}^n c_k q\mathbf{v}_k = \sum_{k=1}^n c_k\mathbf{w}_k = \mathbf{0}$ and so each c_k is 0 because it is given that $\{\mathbf{w}_1,\cdots,\mathbf{w}_n\}$ is linearly independent. For $\sum_{k=1}^n c_k\mathbf{w}_k$ an arbitrary vector of W, this vector equals $\sum_{k=1}^n c_k q\mathbf{v}_k = q\left(\sum_{k=1}^n c_k\mathbf{v}_k\right)$. Therefore, q is also onto. ■

Such a mapping is called an isomorphism.

Definition 18.5.2 *Let V be a vector space of dimension n and W a vector space of dimension m with respect to the same field of scalars $\mathbb{F}$. Let*

$$\begin{aligned} q_\alpha &: \mathbb{F}^n \to V \\ q_\beta &: \mathbb{F}^m \to W \end{aligned}$$

be two isomorphisms as in the above lemma. Here α denotes the basis

$$\{q_\alpha\mathbf{e}_1,\cdots,q_\alpha\mathbf{e}_n\}$$

and β denotes the basis $\{q_\beta\mathbf{e}_1,\cdots,q_\beta\mathbf{e}_m\}$ for V and W respectively. For $L\in\mathscr{L}(V,W)$, the matrix of L with respect to these two bases,

$$\{q_\alpha\mathbf{e}_1,\cdots,q_\alpha\mathbf{e}_n\},\{q_\beta\mathbf{e}_1,\cdots,q_\beta\mathbf{e}_m\}$$

denoted by $[L]_{\beta\alpha}$ or $[L]$ for short satisfies

$$Lq_\alpha = q_\beta [L]_{\beta\alpha} \tag{18.6}$$

In terms of a diagram,

$$\begin{tikzcd} V \arrow[r, "L"] & W \\ \mathbb{F}^n \arrow[u, "q_\alpha"] \arrow[r, "{[L]_{\beta\alpha}}"'] \arrow[ur, phantom, "\circ"] & \mathbb{F}^m \arrow[u, "q_\beta"'] \end{tikzcd} \tag{18.7}$$

Starting at $\mathbb{F}^n$, you go up and then to the right using L and the result is the same if you go to the right by matrix multiplication by $[L]_{\beta\alpha}$ and then up using q_β.

So how can we find this matrix? Let

$$\begin{aligned} \alpha &\equiv \{q_\alpha\mathbf{e}_1,\cdots,q_\alpha\mathbf{e}_n\} \equiv \{\mathbf{v}_1,\cdots,\mathbf{v}_n\} \\ \beta &\equiv \{q_\beta\mathbf{e}_1,\cdots,q_\beta\mathbf{e}_m\} \equiv \{\mathbf{w}_1,\cdots,\mathbf{w}_m\} \end{aligned}$$

and suppose the ij^{th} entry of the desired matrix is a_{ij}. Letting $\mathbf{b}\in\mathbb{F}^n$, the requirement 18.6 is equivalent to

$$\sum_i \overbrace{\left(\sum_j a_{ij}b_j\right)}^{=([L]\mathbf{b})_i} \mathbf{w}_i = L\sum_j b_j\mathbf{v}_j = \sum_j b_j L\mathbf{v}_j.$$

Thus interchanging the order in the sum on the left,

$$\sum_j \left(\sum_i a_{ij} \mathbf{w}_i \right) b_j = \sum_j (L\mathbf{v}_j) b_j,$$

and this must hold for all $\mathbf{b}$ which happens if and only if

$$L\mathbf{v}_j = \sum_i a_{ij} \mathbf{w}_i \tag{18.8}$$

It may help to write 18.8 in the form

$$\begin{pmatrix} L\mathbf{v}_1 & \cdots & L\mathbf{v}_n \end{pmatrix} = \begin{pmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_m \end{pmatrix} A \tag{18.9}$$

where $[L] = A = (a_{ij})$.

A little less precisely, you need for $\mathbf{b} \in \mathbb{F}^n$,

$$\begin{pmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_m \end{pmatrix} A\mathbf{b} = L \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{pmatrix} \mathbf{b} = \begin{pmatrix} L\mathbf{v}_1 & \cdots & L\mathbf{v}_n \end{pmatrix} \mathbf{b} \tag{18.10}$$

where we use the usual conventions for notation like the above. Then since 18.10 is to hold for all $\mathbf{b}$, 18.9 follows.

Example 18.5.3 *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 18.9,

$$\begin{pmatrix} 0 & 1 & 2x & 3x^2 \end{pmatrix} = \begin{pmatrix} 1 & x & x^2 \end{pmatrix} [D].$$

It follows from this that

$$[D] = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

Now consider the important case where $V = \mathbb{F}^n$, $W = \mathbb{F}^m$, and the basis chosen is the standard basis of vectors $\mathbf{e}_i \equiv \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}^T$ the 1 in the i^{th} position. Let L be a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$ and let $[L]$ be the matrix of the transformation with respect to these bases. Thus

$$\alpha = \{\mathbf{e}_1, \cdots, \mathbf{e}_n\}, \beta = \{\mathbf{e}_1, \cdots, \mathbf{e}_m\}$$

and so, in this case the coordinate maps q_α and q_β are simply the identity map and the requirement that A is the matrix of the transformation amounts to

$$(L\mathbf{b})_i = ([L]\mathbf{b})_i$$

where this above indicates the i^{th} entry of the indicated vectors taken 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 if the entries of $[L]$ are a_{ij}, you would need to have

$$(L\mathbf{b})_i = \sum_j a_{ij} b_j$$

In terms of matrix notation, you would need

$$\begin{pmatrix} L\mathbf{e}_i & \cdots & L\mathbf{e}_n \end{pmatrix} = I[L]$$

The following example illustrates what happens when you consider the matrix of a linear transformation with respect to two different bases.

Example 18.5.4 *You have a linear transformation $L : \mathbb{R}^3 \to \mathbb{R}^3$ and it is given on a basis by*

$$L\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, L\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, L\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

Find the matrix of this linear transformation relative to the basis

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

Then find the matrix of this transformation with respect to the usual basis on $\mathbb{R}^3$.

The matrix with columns equal to $L\mathbf{v}_i$ for the $\mathbf{v}_i$ the basis vectors is on the left in what is below. Thus if A is the matrix of the transformation with respect to this basis,

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

Then multiplying by the inverse of the first matrix on the right, you need

$$A = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 & 0 \\ -1 & 2 & 1 \\ 1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -5 & 1 \\ -1 & 4 & 0 \\ 0 & -2 & 1 \end{pmatrix}$$

Note how it works. Start with a column vector $(x, y, z)^T$. Then do q_α to it to get

$$x\begin{pmatrix} 1 & 0 & 1 \end{pmatrix}^T + y\begin{pmatrix} 1 & 1 & 1 \end{pmatrix}^T + z\begin{pmatrix} -1 & 1 & 0 \end{pmatrix}^T.$$

Then you do L to this and get

$$x\begin{pmatrix}1\\-1\\1\end{pmatrix}+y\begin{pmatrix}1\\2\\-1\end{pmatrix}+z\begin{pmatrix}0\\1\\1\end{pmatrix}=\begin{pmatrix}x+y\\2y-x+z\\x-y+z\end{pmatrix}$$

Now take the matrix of the transformation times the given column vector.

$$\begin{pmatrix}2&-5&1\\-1&4&0\\0&-2&1\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}2x-5y+z\\4y-x\\z-2y\end{pmatrix}$$

Is this the coordinate vector of the above relative to the given basis?

$$(2x-5y+z)\begin{pmatrix}1\\0\\1\end{pmatrix}+(4y-x)\begin{pmatrix}1\\1\\1\end{pmatrix}+(z-2y)\begin{pmatrix}-1\\1\\0\end{pmatrix}=\begin{pmatrix}x+y\\2y-x+z\\x-y+z\end{pmatrix}$$

You see it is the same thing.

Now lets find the matrix of L with respect to the usual basis. Let B be this matrix. That is, multiplication by B is the same as doing L. Thus

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

Hence

$$B=\begin{pmatrix}1&1&0\\-1&2&1\\1&-1&1\end{pmatrix}\begin{pmatrix}1&1&-1\\0&1&1\\1&1&0\end{pmatrix}^{-1}=\begin{pmatrix}0&0&1\\2&3&-3\\-3&-2&4\end{pmatrix}$$

Of course this is a very different matrix than the matrix of the linear transformation with respect to the funny basis.

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.

$$\begin{tikzcd}
\mathbb{F}^n \arrow[r, "A_2"] \arrow[d, "q_2"'] \arrow[dr, phantom, "\circ"] & \mathbb{F}^m \arrow[d, "p_2"'] \\
V \arrow[r, "L"] \arrow[dr, phantom, "\circ"] & W \\
\mathbb{F}^n \arrow[r, "A_1"] \arrow[u, "q_1"] & \mathbb{F}^m \arrow[u, "p_1"]
\end{tikzcd}$$

In this diagram q_i and p_i are coordinate maps as described above. From the diagram,

$$p_1^{-1}p_2A_2q_2^{-1}q_1=A_1,$$

where $q_2^{-1}q_1$ and $p_1^{-1}p_2$ are one to one, onto, and linear maps. Thus the effect of these maps is identical to multiplication by a suitable matrix.

Definition 18.5.5 *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 $\mathbb{F}^n$,

$$S^{-1} A_2 S = A_1. \tag{18.11}$$

When this occurs, A_1 is said to be similar to A_2 and $A \mapsto S^{-1}AS$ is called a similarity transformation.

Here is some terminology.

Definition 18.5.6 *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 18.5.7 *$[x]$ 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 18.5.8 *Let $\sim$ be an equivalence class defined on a set, S and let $\mathscr{H}$ 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 18.5.9 *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 $L \in \mathscr{L}(V,V)$ 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.

It was pointed out in the above definition that if A, B are matrices which come from a single linear transformation, then they are similar. Suppose now that A, B are similar.

$A=S^{-1}BS$. Pick a basis α for V and let q_α be as described above. Then in the following diagram, define $L\equiv q_\alpha A q_\alpha^{-1}$.

$$\begin{tikzcd} V \arrow[r, "L"] & V \\ \mathbb{F}^n \arrow[u, "q_\alpha"] \arrow[r, "A"'] \arrow[ur, phantom, "\circ"] & \mathbb{F}^n \arrow[u, "q_\alpha"'] \end{tikzcd}$$

Then since A,B are similar,

$$L=q_\alpha S^{-1}BSq_\alpha^{-1}$$

Let $q_\beta\equiv q_\alpha S^{-1}$. Then

$$Lq_\beta=q_\alpha S^{-1}BSq_\alpha^{-1}q_\alpha S^{-1}=q_\alpha S^{-1}B=q_\beta B$$

and so B is the matrix of L with respect to the basis β. ■

What if the linear transformation consists of multiplication by a matrix A and you want to find the matrix of this linear transformation with respect to another basis? Is there an easy way to do it? The answer is yes. It was illustrated in one of the above examples.

Proposition 18.5.10 *Let A be an $m\times n$ matrix and let L be the linear transformation which is defined by multiplication on the left by A. Then the matrix M of this linear transformation with respect to the bases $\{\mathbf{u}_1,\cdots,\mathbf{u}_n\}$ for $\mathbb{F}^n$ and $\{\mathbf{w}_1,\cdots,\mathbf{w}_m\}$ for $\mathbb{F}^m$ is given by*

$$M=\begin{pmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_m\end{pmatrix}^{-1}A\begin{pmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_n\end{pmatrix}$$

where $\begin{pmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_m\end{pmatrix}$ is the $m\times m$ matrix which has $\mathbf{w}_j$ as its j^{th} column.

Proof: Consider the following diagram. The situation is that

$$A\begin{pmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_n\end{pmatrix}=\begin{pmatrix}\mathbf{w}_1 & \cdots & \mathbf{w}_m\end{pmatrix}M$$

Hence the matrix M is given by the claimed formula. ■

Definition 18.5.11 *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.*

Which matrices are diagonalizable?

Theorem 18.5.12 *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 18.5.11 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 or the way we multiply matrices,

$$S^{-1}AS = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix} (A\mathbf{v}_1 \cdots A\mathbf{v}_n) = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_n^T \end{pmatrix} (\lambda_1 \mathbf{v}_1 \cdots \lambda_n \mathbf{v}_n)$$

$$= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} = D.$$

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 $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$. ■

It makes sense to speak of the determinant of a linear transformation as described in the following corollary.

Corollary 18.5.13 *Let $L \in \mathscr{L}(V,V)$ 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 18.5.9 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

$$\det(A) = \det(B) \quad ■$$

Definition 18.5.14 *Let $A \in \mathscr{L}(X,Y)$ 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 18.5.15 *Let $A \in \mathscr{L}(X,Y)$. 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.

$$\begin{tikzcd}
X \arrow[r, "A"] & Y \\
\mathbb{F}^n \arrow[u, "q_\alpha"] \arrow[r, "M"] \arrow[ur, phantom, "\circ"] & \mathbb{F}^m \arrow[u, "q_\beta"']
\end{tikzcd}$$

The maps $q_\alpha, a_\alpha^{-1}, q_\beta, q_\beta^{-1}$ are one to one and onto. They take independent sets to independent sets. Let $\{Ax_1, \cdots, Ax_r\}$ be a basis for $A(X)$. Thus $\{x_1, \cdots, x_r\}$ is independent. Let $x_i = q_\alpha \mathbf{u}_i$. Then the $\{\mathbf{u}_i\}$ are independent and from the definition of the matrix of the linear transformation A,

$$\begin{aligned} A(X) &= \text{span}(Aq_\alpha \mathbf{u}_1, \cdots, Aq_\alpha \mathbf{u}_r) = \text{span}\left(q_\beta M \mathbf{u}_1, \cdots, q_\beta M \mathbf{u}_r\right) \\ &= q_\beta \,\text{span}(M\mathbf{u}_1, \cdots, M\mathbf{u}_r) \end{aligned}$$

However, it also follows from the definition of M that $A(X) = q_\beta M(\mathbb{F}^n)$. Hence $M(\mathbb{F}^n) = \text{span}(M\mathbf{u}_1, \cdots, M\mathbf{u}_r)$ and so the span of the $M\mathbf{u}_j$ equals $M(\mathbb{F}^n)$. If $\sum_j c_j M\mathbf{u}_j = \mathbf{0}$, then

$$\sum_j c_j q_\beta M \mathbf{u}_j = \mathbf{0} = \sum_j c_j A q_\alpha \mathbf{u}_i = \sum_j c_j A x_j$$

and so each $c_j = 0$. Therefore, $\dim(M\mathbb{F}^n) = \dim\left(q_\beta M(\mathbb{F}^n)\right) = \dim(A(X))$ which shows that the rank of the matrix equals the rank of the transformation. ■

The following result is a summary of many concepts.

Theorem 18.5.16 *Let $L \in \mathscr{L}(V,V)$ 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 $L\mathbf{v} = \mathbf{0}$ then $\mathbf{v} = \mathbf{0}$.*

Proof: Suppose first L is one to one and let $\{\mathbf{v}_i\}_{i=1}^n$ be a basis. Then if $\sum_{i=1}^n c_i L\mathbf{v}_i = \mathbf{0}$ it follows $L(\sum_{i=1}^n c_i \mathbf{v}_i) = \mathbf{0}$ which means that since $L(\mathbf{0}) = \mathbf{0}$, and L is one to one, it must be the case that $\sum_{i=1}^n c_i \mathbf{v}_i = \mathbf{0}$. Since $\{\mathbf{v}_i\}$ is a basis, each $c_i = 0$ which shows $\{L\mathbf{v}_i\}$ is a linearly independent set. Since there are n of these, it must be that this is a basis.

Now suppose 2.). Then letting $\{\mathbf{v}_i\}$ be a basis, and $\mathbf{y} \in V$, it follows from part 2.) that there are constants, $\{c_i\}$ such that $\mathbf{y} = \sum_{i=1}^n c_i L\mathbf{v}_i = L(\sum_{i=1}^n c_i \mathbf{v}_i)$. Thus L is onto. This shows that 2.) implies 3.).

Now suppose 3.). Then the operation consisting of multiplication by the matrix of L, $[L]$, must be onto. However, the vectors in $\mathbb{F}^n$ so obtained, consist of linear combinations of the columns of $[L]$. Therefore, the column rank of $[L]$ is n. By Theorem 8.6.7 this equals the determinant rank and so $\det([L]) \equiv \det(L) \neq 0$.

Now assume 4.) If $L\mathbf{v} = \mathbf{0}$ for some $\mathbf{v} \neq \mathbf{0}$, it follows that $[L]\mathbf{x} = \mathbf{0}$ for some $\mathbf{x} \neq \mathbf{0}$. Therefore, the columns of $[L]$ are linearly dependent and so by Theorem 8.6.7, $\det([L]) = \det(L) = 0$ contrary to 4.). Therefore, 4.) implies 5.).

Now suppose 5.) and suppose $L\mathbf{v} = L\mathbf{w}$. Then $L(\mathbf{v}-\mathbf{w}) = \mathbf{0}$ and so by 5.), $\mathbf{v}-\mathbf{w} = \mathbf{0}$ showing that L is one to one. ■

Also it is important to note that composition of linear transformation corresponds tomultiplication of the matrices. Consider the following diagram.

$$\begin{tikzcd}
X \arrow[r, "A"] & Y \arrow[r, "B"] & Z \\
\mathbb{F}^n \arrow[u, "q_\alpha"] \arrow[r, "{[A]_{\beta\alpha}}"'] \arrow[ur, phantom, "\circ"] & \mathbb{F}^m \arrow[u, "q_\beta"'] \arrow[r, "{[B]_{\gamma\beta}}"'] \arrow[ur, phantom, "\circ"] & \mathbb{F}^p \arrow[u, "q_\gamma"']
\end{tikzcd}$$

where A and B are two linear transformations, $A \in \mathscr{L}(X,Y)$ and $B \in \mathscr{L}(Y,Z)$. Then $B \circ A \in \mathscr{L}(X,Z)$ and so it has a matrix with respect to bases given on X and Z, the coordinate maps for these bases being q_α and q_β respectively. Then

$$B \circ A = q_\gamma [B]_{\gamma\beta} q_\beta^{-1} q_\beta [A]_{\beta\alpha} q_\alpha^{-1} = q_\gamma [B]_{\gamma\beta} [A]_{\beta\alpha} q_\alpha^{-1}.$$

But this shows that $[B]_{\gamma\beta}[A]_{\beta\alpha}$ plays the role of $[B\circ A]_{\gamma\alpha}$, the matrix of $B\circ A$. Hence the matrix of $B\circ A$ equals the product $[B]_{\gamma\beta}[A]_{\beta\alpha}$. Of course it is interesting to note that although $[B\circ A]_{\gamma\alpha}$ must be unique, the matrices, $[B]_{\gamma\beta}$ and $[A]_{\beta\alpha}$ are not unique because they depend on the basis chosen for Y.

Theorem 18.5.17 *The matrix of the composition of linear transformations equals the product of the matrices of these linear transformations in the same order as the composition.*

18.5.1 Some Geometrically Defined Linear Transformations

This is a review of earlier material. 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

$$A\mathbf{x} = T\mathbf{x} \tag{18.12}$$

for all $\mathbf{x} \in \mathbb{F}^n$. How does this relate to what is discussed above? In terms of the above diagram,

$$\begin{tikzcd}
\{\mathbf{e}_1,\cdots,\mathbf{e}_n\} & \mathbb{F}^n \arrow[r, "T"] & \mathbb{F}^m & \{\mathbf{e}_1,\cdots,\mathbf{e}_n\} \\
 & \mathbb{F}^n \arrow[u, "q_{\mathbb{F}^n}"] \arrow[r, "M"'] \arrow[ur, phantom, "\circ"] & \mathbb{F}^m \arrow[u, "q_{\mathbb{F}^m}"'] &
\end{tikzcd}$$

where

$$q_{\mathbb{F}^n}(\mathbf{x}) \equiv \sum_{i=1}^n x_i \mathbf{e}_i = \mathbf{x}.$$

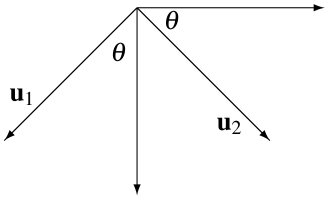
Thus those two maps are really just the identity map. Thus, to find the matrix of the linear transformation T with respect to the standard basis vectors,

$$T\mathbf{e}_k = M\mathbf{e}_k$$

In other words, the k^{th} column of M equals $T\mathbf{e}_k$ as noted earlier. All the earlier considerations apply. These considerations were just a specialization to the case of the standard basis vectors of this more general notion which was just presented.

18.5.2 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 unit vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$ and the problem of rotating 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. Thus the vector $\mathbf{u_3}$ is coming out of the page.



Let T denote the desired rotation. Then

$$T\left(a\mathbf{u}_1 + b\mathbf{u}_2 + c\mathbf{u}_3\right) = aT\mathbf{u}_1 + bT\mathbf{u}_2 + cT\mathbf{u}_3$$

$$= (a\cos\theta - b\sin\theta)\,\mathbf{u}_1 + (a\sin\theta + b\cos\theta)\,\mathbf{u}_2 + c\mathbf{u}_3.$$

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, $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$. From Proposition 18.5.10, if A is this matrix,

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix}^{-1} A \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{pmatrix}$$

and so you can solve for A if you know the $\mathbf{u}_i$.

Suppose the unit vector about which the counterclockwise rotation takes place is denoted as (a,b,c). 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 $\mathbf{u}_3 = (a,b,c)$ 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 otherwise 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 rotation about the positive z axis.) Then let $\mathbf{u}_3 = (a,b,c)$ and $\mathbf{u}_2 \equiv \frac{1}{\sqrt{a^2+b^2}}(b,-a,0)$. This one is perpendicular to $\mathbf{u}_3$. If $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is to be a right hand system it is necessary to have

$$\mathbf{u}_1 = \mathbf{u}_2 \times \mathbf{u}_3 = \frac{1}{\sqrt{(a^2+b^2)(a^2+b^2+c^2)}}\left(-ac, -bc, a^2+b^2\right)$$

Now recall that $\mathbf{u}_3$ is a unit vector and so the above equals

$$\frac{1}{\sqrt{(a^2+b^2)}}\left(-ac,-bc,a^2+b^2\right)$$

Then from the above, A is given by

$$\begin{pmatrix} \frac{-ac}{\sqrt{(a^2+b^2)}} & \frac{b}{\sqrt{a^2+b^2}} & a \\ \frac{-bc}{\sqrt{(a^2+b^2)}} & \frac{-a}{\sqrt{a^2+b^2}} & b \\ \sqrt{a^2+b^2} & 0 & c \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \frac{-ac}{\sqrt{(a^2+b^2)}} & \frac{b}{\sqrt{a^2+b^2}} & a \\ \frac{-bc}{\sqrt{(a^2+b^2)}} & \frac{-a}{\sqrt{a^2+b^2}} & b \\ \sqrt{a^2+b^2} & 0 & c \end{pmatrix}^{-1}$$

Of course the matrix is an orthogonal matrix so it is easy to take the inverse by simply taking the transpose. Then doing the computation and then some simplification yields

$$= \begin{pmatrix} a^2+\left(1-a^2\right)\cos\theta & ab\left(1-\cos\theta\right)-c\sin\theta & ac\left(1-\cos\theta\right)+b\sin\theta \\ ab\left(1-\cos\theta\right)+c\sin\theta & b^2+\left(1-b^2\right)\cos\theta & bc\left(1-\cos\theta\right)-a\sin\theta \\ ac\left(1-\cos\theta\right)-b\sin\theta & bc\left(1-\cos\theta\right)+a\sin\theta & c^2+\left(1-c^2\right)\cos\theta \end{pmatrix}. \tag{18.13}$$

With this, it is clear how to rotate clockwise about the unit vector (a,b,c). Just rotate counter clockwise through an angle of $-\theta$. Thus the matrix for this clockwise rotation 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 18.13 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 18.13 reduces to the correct matrix for rotation about the positive z axis.

18.6 The Matrix Exponential, Differential Equations *

You want to find a matrix valued function $\Phi(t)$ such that

$$\Phi'(t) = A\Phi(t),\ \Phi(0) = I,\ A \text{ is } p\times p \tag{18.14}$$

Such a matrix is called a fundamental matrix.

What is meant by the above symbols? The idea is that $\Phi(t)$ is a matrix whose entries are differentiable functions of t. The meaning of $\Phi'(t)$ is the matrix whose entries are the derivatives of the entries of $\Phi(t)$. For example, abusing notation slightly,

$$\begin{pmatrix} t & t^2 \\ \sin(t) & \tan(t) \end{pmatrix}' = \begin{pmatrix} 1 & 2t \\ \cos(t) & \sec^2(t) \end{pmatrix}.$$

What are some properties of this derivative? Does the product rule hold for example?

Lemma 18.6.1 *Suppose $\Phi(t)$ is $m\times n$ and $\Psi(t)$ is $n\times p$ and these are differentiable matrices. Then*

$$(\Phi(t)\Psi(t))' = \Phi'(t)\Psi(t) + \Phi(t)\Psi'(t)$$

Proof: By definition,

$$\begin{aligned}
\left((\Phi(t)\Psi(t))'\right)_{ij} &= \left((\Phi(t)\Psi(t))_{ij}\right)' = \left(\sum_k \Phi(t)_{ik}\Psi(t)_{kj}\right)' \\
&= \sum_k \Phi'(t)_{ik}\Psi(t)_{kj} + \sum_k \Phi(t)_{ik}\Psi'(t)_{kj} \\
&= \left(\Phi'(t)\Psi(t)\right)_{ij} + \left(\Phi(t)\Psi'(t)\right)_{ij}
\end{aligned}$$

and so the conclusion follows. ■

What do we mean when we say that for $\{B_n\}$ a sequence of matrices

$$\lim_{n\to\infty} B_n = B?$$

We mean the obvious thing. The ij^{th} entry of B_n converges to the ij^{th} entry of B. One convenient way to ensure that this happens is to give a measure of distance between matrices which will ensure that it happens.

Definition 18.6.2 *For A,B matrices of the same size, define $\|A-B\|_\infty$ to be*

$$\max\left\{\left|A_{ij}-B_{ij}\right|, \text{ all } ij\right\}$$

Thus

$$\|A\|_\infty = \max\left\{\left|A_{ij}\right|, \text{ all } ij\right\}$$

To say that $\lim_{n\to\infty} B_n = B$ *is the same as saying that* $\lim_{n\to\infty}\|B_n - B\|_\infty = 0$.

Here is a useful lemma.

Lemma 18.6.3 *If A,B_n,B are $p\times p$ matrices and* $\lim_{n\to\infty} B_n = B$, *then*

$$\begin{aligned}
\lim_{n\to\infty} AB_n &= AB, \\
\lim_{n\to\infty} B_nA &= BA,
\end{aligned} \tag{18.15}$$

Also

$$\|AB\|_\infty \leq p\,\|A\|_\infty\,\|B\|_\infty \tag{18.16}$$

$$\left\|A^k\right\|_\infty \leq p^{k-1}\,\|A\|_\infty^k \tag{18.17}$$

for any positive integer k and

$$\left\|\sum_{k=1}^m A_k\right\|_\infty \leq \sum_{k=1}^m \|A_k\|_\infty$$

For t a scalar,

$$\|tA\|_\infty = |t|\,\|A\|_\infty \tag{18.18}$$

Also

$$|A\mathbf{x}| \leq \sqrt{p}\,\|A\|_\infty\,|\mathbf{x}| \tag{18.19}$$

and

$$\|A+B\|_\infty \leq \|A\|_\infty + \|B\|_\infty$$

Proof: First consider the claim 18.16.

$$\begin{aligned}\|AB\|_\infty &\equiv \sup_{i,j}\left|\sum_k A_{ik}B_{kj}\right| \leq \sup_{i,j}\sum_k \|A\|_\infty \|B\|_\infty \\ &= \sup_{i,j} p\,\|A\|_\infty \|B\|_\infty \leq p\,\|A\|_\infty \|B\|_\infty\end{aligned}$$

Now consider 18.15. From what was just shown,

$$\|AB_n - AB\|_\infty = \|A(B_n - B)\|_\infty \leq p\,\|A\|_\infty \|B_n - B\|_\infty$$

which is assumed to converge to 0. Similarly $B_nA \to BA$. This establishes the first part of the lemma. Now 18.17 follows by induction. Indeed, the result holds for $k = 1$. Suppose true for $n-1$ for $n \geq 2$. Then

$$\left\|AA^{n-1}\right\|_\infty \leq p\,\|A\|_\infty \left\|A^{n-1}\right\|_\infty \leq p\,\|A\|_\infty p^{n-2}\,\|A\|_\infty^{n-1} = p^{n-1}\,\|A\|_\infty^n .$$

Consider the claim about the sum.

$$\left|\left(\sum_{k=1}^m A_k\right)_{ij}\right| = \left|\sum_{k=1}^m (A_k)_{ij}\right| \leq \sum_{k=1}^m \|A_k\|_\infty$$

Since this holds for arbitrary ij, it follows that

$$\left\|\sum_{k=1}^m A_k\right\|_\infty \leq \sum_{k=1}^m \|A_k\|_\infty$$

as claimed. The assertion 18.18 is obvious. Consider 18.19. Using the Cauchy Schwarz inequality as needed,

$$\begin{aligned}|A\mathbf{x}| &\equiv \left|\sum_{j=1}^p A_{ij}x_j\right| \leq \|A\|_\infty \sum_{j=1}^p |x_j| \\ &\leq \|A\|_\infty \left(\sum_{j=1}^p 1^2\right)^{1/2}\left(\sum_{j=1}^p |x_j|^2\right)^{1/2} \leq \sqrt{p}\,\|A\|_\infty |\mathbf{x}|\end{aligned}$$

Now consider the last claim.

$$\left|A_{ij} + B_{ij}\right| \leq \|A\|_\infty + \|B\|_\infty$$

and so,

$$\|A + B\|_\infty \leq \|A\|_\infty + \|B\|_\infty \blacksquare$$

Thus the convention for taking the derivative above could also be obtained by

$$A'(t) \equiv \lim_{h\to 0}\frac{A(t+h) - A(t)}{h}$$

because this corresponds to taking this limit for each $A_{ij}(t)$.

By analogy with the situation in calculus, consider the infinite sum

$$\sum_{k=0}^{\infty} \frac{A^k t^k}{k!} \equiv \lim_{n\to\infty} \sum_{k=0}^{n} \frac{A^k t^k}{k!}$$

where here A is a $p \times p$ matrix having real or complex entries. Then letting $m < n$, it follows from the above lemma that

$$\begin{aligned} \left\| \sum_{k=0}^{n} \frac{A^k t^k}{k!} - \sum_{k=0}^{m} \frac{A^k t^k}{k!} \right\|_{\infty} &= \left\| \sum_{k=m+1}^{n} \frac{A^k t^k}{k!} \right\|_{\infty} \leq \sum_{k=m+1}^{n} \frac{|t|^k}{k!} \left\| A^k \right\|_{\infty} \\ &\leq \sum_{k=m}^{\infty} \frac{|t|^k}{k!} \left\| A^k \right\|_{\infty} \leq \sum_{k=m}^{\infty} \frac{|t|^k p^k \|A\|_{\infty}^k}{k!} \end{aligned}$$

Now the series $\sum_{k=0}^{\infty} \frac{|t|^k p^k \|A\|_{\infty}^k}{k!}$ converges and in fact equals $\exp\left(|t|\, p \,\|A\|_{\infty}\right)$. It follows from calculus that

$$\lim_{m\to\infty} \sum_{k=m}^{\infty} \frac{|t|^k p^k \|A\|_{\infty}^k}{k!} = 0.$$

It follows that the ij^{th} entry of the partial sum $\sum_{k=0}^{n} \frac{A^k t^k}{k!}$ is a Cauchy sequence and hence by completeness of $\mathbb{C}$ or $\mathbb{R}$ it converges. Therefore, the above limit exists. This is stated as the essential part of the following theorem.

Theorem 18.6.4 *Let $t \in [a,b] \subseteq \mathbb{R}$ where $b - a < \infty$. Then for each $t \in [a,b]$,*

$$\lim_{n\to\infty} \sum_{k=0}^{n} \frac{A^k t^k}{k!} \equiv \sum_{k=0}^{\infty} \frac{A^k t^k}{k!} \equiv \Phi(t),\ A^0 \equiv I,$$

exists. Furthermore, there exists a single constant C such that for $t_k \in [a,b]$, the infinite sum

$$\sum_{k=0}^{\infty} \frac{A^k t_k^k}{k!}$$

converges and in fact

$$\left\| \sum_{k=0}^{\infty} \frac{A^k t_k^k}{k!} \right\|_{\infty} \leq C$$

Proof: The convergence for $\sum_{k=0}^{\infty} \frac{A^k t^k}{k!}$ was just established.
Consider the estimate. From the above lemma,

$$\begin{aligned} \left\| \sum_{k=0}^{n} \frac{A^k t_k^k}{k!} \right\|_{\infty} &\leq \sum_{k=0}^{n} \frac{p^k (|a|+|b|)^k \|A\|_{\infty}^k}{k!} \\ &\leq \sum_{k=0}^{\infty} \frac{p^k (|a|+|b|)^k \|A\|_{\infty}^k}{k!} \\ &= \exp\left(p\left(|a|+|b|\right) \|A\|_{\infty}\right) \end{aligned}$$

It follows that the ij^{th} entry of $\sum_{k=0}^{n} \frac{A^k t_k^k}{k!}$ has magnitude no larger than the right side of the above inequality. Also, a repeat of the above argument after Lemma 18.6.3 shows that the partial sums of the ij^{th} entry of $\sum_{k=0}^{\infty} \frac{A^k t_k^k}{k!}$ form a Cauchy sequence. Hence passing to the limit, it follows from calculus that

$$\left| \left(\sum_{k=0}^{\infty} \frac{A^k t^k}{k!} \right)_{ij} \right| \leq \exp\left(p\left(|a| + |b| \right) \|A\|_{\infty} \right)$$

Since ij is arbitrary, this establishes the inequality. ■

Next consider the derivative of $\Phi(t)$. Why is $\Phi(t)$ a solution to the above 18.14? By the mean value theorem,

$$\begin{aligned} \frac{\Phi(t+h) - \Phi(t)}{h} &= \frac{1}{h} \sum_{k=0}^{\infty} \frac{(t+h)^k - t^k}{k!} A^k = \frac{1}{h} \sum_{k=0}^{\infty} \frac{k(t+\theta_k h)^{k-1} h}{k!} A^k \\ &= A \sum_{k=1}^{\infty} \frac{(t+\theta_k h)^{k-1}}{(k-1)!} A^{k-1} = A \sum_{k=0}^{\infty} \frac{(t+\theta_k h)^k}{k!} A^k, \theta_k \in (0,1) \end{aligned}$$

Does this sum converge to $\sum_{k=0}^{\infty} \frac{t^k}{k!} A^k \equiv \Phi(t)$? If so, it will have been shown that $\Phi'(t) = A\Phi(t)$. By the mean value theorem again,

$$\begin{aligned} &\left\| \sum_{k=0}^{\infty} \frac{(t+\theta_k h)^k}{k!} A^k - \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k \right\|_{\infty} \\ = &\left\| h \sum_{k=0}^{\infty} \frac{k(t+\eta_k h)^{k-1} \theta_k}{k!} A^k \right\|_{\infty}, \ \eta_k \in (0,1) \\ \leq &\left\| h \sum_{k=0}^{\infty} \frac{k(t+\eta_k h)^{k-1}}{k!} A^k \right\|_{\infty} \end{aligned}$$

Now for $|h| \leq 1$, the expression $t_k \equiv t + \eta_k h \in [t-1, t+1]$ and so by Theorem 18.6.4,

$$\left\| \sum_{k=0}^{\infty} \frac{(t+\theta_k h)^k}{k!} A^k - \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k \right\|_{\infty} = \left\| h \sum_{k=0}^{\infty} \frac{k(t+\eta_k h)^{k-1}}{k!} A^k \right\|_{\infty} \leq C|h| \tag{18.20}$$

for some C. Then

$$\left\| \frac{\Phi(t+h) - \Phi(t)}{h} - A\Phi(t) \right\|_{\infty} = \left\| A \sum_{k=0}^{\infty} \frac{(t+\theta_k h)^k}{k!} A^k - A \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k \right\|_{\infty}$$

This converges to 0 as $h \to 0$ by Lemma 18.6.3 and 18.20. Hence the ij^{th} entry of the difference quotient

$$\frac{\Phi(t+h) - \Phi(t)}{h}$$

converges to the ij^{th} entry of the matrix $A \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k = A\Phi(t)$. In other words,

$$\Phi'(t) = A\Phi(t)$$

Now also it follows right away from the formula for the infinite sum that $\Phi(0) = I$. This proves the following theorem.

Theorem 18.6.5 *Let A be a real or complex $p \times p$ matrix. Then there exists a differentiable $p \times p$ matrix $\Phi(t)$ satisfying the following initial value problem.*

$$\Phi'(t) = A\Phi(t), \quad \Phi(0) = I. \tag{18.21}$$

This matrix is given by the infinite sum

$$\Phi(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k$$

with the usual convention that $t^0 = 1, A^0 = I, 0! = 1$. In addition to this, $A\Phi(t) = \Phi(t)A$.

Proof: Why is $A\Phi(t) = \Phi(t)A$? This follows from the observation that A obviously commutes with the partial sums,

$$A\sum_{k=0}^{n} \frac{t^k}{k!} A^k = \sum_{k=0}^{n} \frac{t^k}{k!} A^k A = \sum_{k=0}^{n} \frac{t^k}{k!} A^{k+1},$$

and Lemma 18.6.3,

$$A\Phi(t) = \lim_{n\to\infty} A\sum_{k=0}^{n} \frac{t^k}{k!} A^k = \lim_{n\to\infty} \left(\sum_{k=0}^{n} \frac{t^k}{k!} A^k\right) A = \Phi(t)A. \blacksquare$$

Now let

$$\Psi(t) = \sum_{k=0}^{\infty} \frac{(-A)^k t^k}{k!}$$

In the same way as above $\Psi'(t) = (-A)\Psi(t), \Psi(0) = I,$ and $A\Psi(t) = \Psi(t)A$.

Lemma 18.6.6 $\Phi(t)^{-1} = \Psi(t)$

Proof:

$$\begin{aligned}
(\Phi(t)\Psi(t))' &= \Phi'(t)\Psi(t) + \Phi(t)\Psi'(t) \\
&= A\Phi(t)\Psi(t) + \Phi(t)(-A)\Psi(t) \\
&= A\Phi(t)\Psi(t) - A\Phi(t)\Psi(t) = 0
\end{aligned}$$

Therefore, $\Phi(t)\Psi(t)$ is a constant matrix. Just use the usual calculus facts on the entries of the matrix. This matrix can only be I because $\Phi(0) = \Psi(0) = I$. ■

What follows contains all mathematically significant features of a typical undergraduate differential equations course without the plethora of loose ends which result when thoughtful students begin to wonder whether there exist enough generalized eigenvectors and eigenvectors to represent the solution. This seems to be never explained in these beginning differential equations courses. However, to see why there are enough of these, read the appendix on the Jordan form.

The formula in the following theorem is called the variation of constants formula. I have to admit that I don't see the mathematical significance for a typical undergraduate differential equations class when a substantial part of such a course is contained in the following theorem.

Theorem 18.6.7 *Let $\mathbf{f}(t)$ be continuous and let $\mathbf{x}_0 \in \mathbb{R}^p$. Then there exists a unique solution to the equation*

$$\mathbf{x}' = A\mathbf{x} + \mathbf{f} \text{ and initial condition } \mathbf{x}(0) = \mathbf{x}_0$$

This solution is given by the formula

$$\mathbf{x}(t) = \Phi(t)\mathbf{x}_0 + \Phi(t)\int_0^t \Psi(s)\mathbf{f}(s)\,ds$$

Where $\Phi(t)$ is the fundamental matrix described in Theorem 18.6.5 and $\Psi(t)$ is the fundamental matrix defined the same way from $-A$.

Proof: Suppose $\mathbf{x}' = A\mathbf{x} + \mathbf{f}$. Then $\mathbf{x}' - A\mathbf{x} = \mathbf{f}$. Multiply both sides by $\Psi(t)$. Then

$$\begin{aligned}(\Psi(t)\mathbf{x})' &= \Psi(t)\mathbf{x}'(t) + \Psi'(t)\mathbf{x}(t) \\ &= \Psi(t)\mathbf{x}'(t) - A\Psi(t)\mathbf{x}(t) \\ &= \Psi(t)\mathbf{x}'(t) - \Psi(t)A\mathbf{x}(t) \\ &= \Psi(t)\left(\mathbf{x}'(t) - A\mathbf{x}(t)\right)\end{aligned}$$

Therefore,

$$(\Psi(t)\mathbf{x})' = \Psi(t)\mathbf{f}(t)$$

Hence

$$\Psi(t)\mathbf{x}(t) - \mathbf{x}_0 = \int_0^t \Psi(s)\mathbf{f}(s)\,ds$$

Therefore, multiplying on the left by $\Phi(t)$,

$$\mathbf{x}(t) = \Phi(t)\mathbf{x}_0 + \Phi(t)\int_0^t \Psi(s)\mathbf{f}(s)\,ds$$

Thus, if there is a solution, this is it. Now if $\mathbf{x}(t)$ is given by the above formula, then

$$\mathbf{x}(0) = \Phi(0)\mathbf{x}_0 = \mathbf{x}_0$$

and also

$$\mathbf{x}'(t) = \Phi'(t)\mathbf{x}_0 + \Phi'(t)\int_0^t \Psi(s)\mathbf{f}(s)\,ds + \Phi(t)\Psi(t)\mathbf{f}(t)$$

$$= A\Phi(t)\mathbf{x}_0 + A\Phi(t)\int_0^t \Psi(s)\mathbf{f}(s)\,ds + \mathbf{f}(t) = A\mathbf{x}(t) + \mathbf{f}(t) \quad \blacksquare$$

Observation 18.6.8 *As a special case when A is a real or complex scalar, the above theorem shows that $\Phi(t)x_0 \equiv \sum_{k=0}^{\infty}\frac{t^k}{k!}A^k x_0$ is the solution of the differential equation*

$$x'(t) = Ax, \; x(0) = x_0.$$

Another solution to this is e^{At} where this is defined in Section 1.7. It follows that in this special case, the uniqueness provision of the above theorem shows that

$$\sum_{k=0}^{\infty}\frac{t^k}{k!}A^k = e^{At}$$

in the special case that A is a complex number.

18.6.1 Computing A Fundamental Matrix

In case that A is $p\times p$ and nondefective, you can easily compute $\Phi(t)$. Recall that in this case, there exists a matrix S such that

$$S^{-1}AS=D=\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_p\end{pmatrix}$$

where D is a diagonal matrix. Then $A=SDS^{-1}$. Now $\Phi(t)=$

$$\sum_{k=0}^{\infty}\frac{SD^kS^{-1}}{k!}t^k=S\sum_{k=0}^{\infty}\frac{D^k}{k!}t^kS^{-1}=S\begin{pmatrix}e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_p t}\end{pmatrix}S^{-1}$$

Thus you can easily compute $\Phi(t)$ in this case. For the case where a λ_k is complex, see the above observation for the meaning of $e^{\lambda t}$. Thus one can explicitly find the fundamental matrix in this case.

In fact you can find it whenever you can compute the eigenvalues exactly. Suppose that the matrix A is defective. A chain based on the eigenvector $\mathbf{v}_1$ is an ordered list of nonzero vectors

$$(\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_m)$$

where $(A-\lambda I)\mathbf{v}_{k+1}=\mathbf{v}_k$ and $(A-\lambda I)\mathbf{v}_1=\mathbf{0}$. Given such a chain, $m>1$, consider

$$\mathbf{x}(t)\equiv\sum_{k=1}^{m}\frac{t^{m-k}}{(m-k)!}\mathbf{v}_ke^{\lambda t}$$

Then

$$\begin{aligned}\mathbf{x}'(t)&=\lambda\sum_{k=1}^{m}\frac{t^{m-k}}{(m-k)!}\mathbf{v}_ke^{\lambda t}+\sum_{k=1}^{m-1}\frac{(m-k)\,t^{m-(k+1)}}{(m-k)!}\mathbf{v}_ke^{\lambda t}\\&=\lambda\sum_{k=1}^{m}\frac{t^{m-k}}{(m-k)!}\mathbf{v}_ke^{\lambda t}+\sum_{k=1}^{m-1}\frac{t^{m-(k+1)}}{(m-(k+1))!}\mathbf{v}_ke^{\lambda t}\\&=\lambda\sum_{k=1}^{m}\frac{t^{m-k}}{(m-k)!}\mathbf{v}_ke^{\lambda t}+\sum_{k=2}^{m}\frac{t^{m-k}}{(m-k)!}\mathbf{v}_{k-1}e^{\lambda t}\end{aligned}$$

Recalling that $A\mathbf{v}_k-\mathbf{v}_{k-1}=\lambda\mathbf{v}_k$ for $k>1$, and $A\mathbf{v}_1-\lambda\mathbf{v}_1=\mathbf{0}$,

$$=\sum_{k=1}^{m}\frac{t^{m-k}}{(m-k)!}A\mathbf{v}_ke^{\lambda t}-\sum_{k=2}^{m}\frac{t^{m-k}}{(m-k)!}\mathbf{v}_{k-1}e^{\lambda t}+\sum_{k=2}^{m}\frac{t^{m-k}}{(m-k)!}\mathbf{v}_{k-1}e^{\lambda t}$$

$$=A\sum_{k=1}^{m}\frac{t^{m-k}}{(m-k)!}\mathbf{v}_ke^{\lambda t}=A\mathbf{x}(t)$$

Thus each such chain results in a solution to the system of equations. Also, each such chain of length m results in m solutions to the differential equation. Just consider the chains

$$\mathbf{v}_1,(\mathbf{v}_1,\mathbf{v}_2),\cdots,(\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_m),$$

each determining a solution as described above. Letting $\mathbf{x}_k(t)$ denote the solution which comes from the chain $(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k)$ and the above formula involving a sum, it follows that $\mathbf{x}_k(0) = \mathbf{v}_k$. Thus if you consider the solutions coming from the chains

$$\mathbf{v}_1, (\mathbf{v}_1, \mathbf{v}_2), \cdots, (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m)$$

and consider the vectors obtained by letting $t = 0$, this results in the ordered list of vectors

$$\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m$$

This is a linearly independent set of vectors. Suppose for some $l \leq m$

$$\sum_{k=1}^{l} c_k \mathbf{v}_k = \mathbf{0}$$

where not all the $c_k = 0$ and l is as small as possible for this to occur. Then since $\mathbf{v}_1 \neq 0$, it must be that $l \geq 2$. Also, $c_l \neq 0$. Do $A - \lambda I$ to both sides. This gives

$$\mathbf{0} = \sum_{k=2}^{l} c_k \mathbf{v}_{k-1} = \sum_{k=1}^{l-1} c_{k+1} \mathbf{v}_k$$

and so l was not as small as possible after all. Thus the set must be linearly independent after all.

Note that for $\mathscr{C}(\lambda_k)$, a chain based on an eigenvector corresponding to λ_k,

$$A : \text{span}(\mathscr{C}(\lambda_k)) \to \text{span}(\mathscr{C}(\lambda_k))$$

Letting A_k be the linear transformation which is the restriction of A to $\text{span}(\mathscr{C}(\lambda_k))$, what is the matrix of A_k with respect to the ordered basis

$$(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m) = \mathscr{C}(\lambda_k)?$$

$$(A - \lambda_k I)\mathbf{v}_j = \mathbf{v}_{j-1},\ j > 1$$

while $(A - \lambda_k I)\mathbf{v}_1 = \mathbf{0}$. Then formally, the matrix of A_k is given by M where

$$\begin{pmatrix} \lambda_k \mathbf{v}_1 & \mathbf{v}_1 + \lambda_k \mathbf{v}_2 & \cdots & \mathbf{v}_{m-1} + \lambda_k \mathbf{v}_m \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_m \end{pmatrix} M$$

It follows that M is of the form

$$\begin{pmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{pmatrix},$$

a matrix which has all zeros except for λ_k down the main diagonal and 1 down the super diagonal. This is called a Jordan block corresponding to the eigenvalue λ_k.

It can be proved that there are chains $\{\mathscr{C}(\lambda_k)\}_{k=1}^{r}$ of such vectors associated with each eigenvalue λ_k such that the totality of these vectors form a basis for $\mathbb{C}^n$. Then you simply use these solutions as just described to obtain a matrix $\Phi(t)$ whose columns are each solutions to the differential equation

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

and since the vectors just mentioned form a basis, this yields a fundamental matrix.

With respect to this basis, the matrix A becomes similar to one in Jordan Canonical form and the existence of this basis is equivalent to obtaining the existence of the Jordan form. This very interesting theorem is discussed in an appendix if you are interested.

Example 18.6.9 *Find a fundamental matrix for the system*

$$\mathbf{x}' = \begin{pmatrix} 2 & 1 & -1 \\ -1 & 0 & 1 \\ -1 & -2 & 3 \end{pmatrix} \mathbf{x}$$

In this case the eigenvectors and eigenvalues are

$$\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \leftrightarrow 1, \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} \leftrightarrow 2$$

The characteristic polynomial is $(\lambda - 1)(\lambda - 2)^2$ and so 2 is an eigenvalue of algebraic multiplicity 2. Corresponding to this eigenvalue, there is the above eigenvector and a generalized eigenvector $\mathbf{v}_2$ which comes from solving the system

$$(A - 2I)\mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$$

This leads to the augmented matrix

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

a solution to this is $\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$. Therefore, there are three solutions to this differential equation

$$\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} e^t, \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} e^{2t}, t \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} e^{2t} + \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} e^{2t}$$

Then a fundamental matrix is

$$\Phi(t) = \begin{pmatrix} 0 & -e^{2t} & e^{2t} - te^{2t} \\ e^t & e^{2t} & te^{2t} - e^{2t} \\ e^t & e^{2t} & te^{2t} \end{pmatrix}$$

You can check and see that this works. Other situations are similar. If you can believe that there will be enough of these vectors to obtain a basis formed by chains, then this gives a way to obtain a formula for the fundamental matrix. However, to verify that this is the case, you will need to deal with the Jordan canonical form. From the above discussion involving a power series, the existence of a fundamental matrix is not an issue. This is about finding it in closed form using well known functions and it is only this which involves the necessity of considering the Jordan canonical form.

18.7 Exercises

1. Find the matrix with respect to the standard basis vectors for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/3$.

2. Find the matrix with respect to the standard basis vectors for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/4$.

3. Find the matrix with respect to the standard basis vectors for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/12$. **Hint:** Note that $\pi/12 = \pi/3 - \pi/4$.

4. Find the matrix with respect to the standard basis vectors 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.

5. Find the matrix with respect to the standard basis vectors for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $\pi/3$ and then reflects across the y axis.

6. Find the matrix with respect to the standard basis vectors for the linear transformation which rotates every vector in $\mathbb{R}^2$ through an angle of $5\pi/12$. **Hint:** Note that $5\pi/12 = 2\pi/3 - \pi/4$.

7. Let V be an inner product space and $\mathbf{u} \neq \mathbf{0}$. Show that the function $T_{\mathbf{u}}$ defined by $T_{\mathbf{u}}(\mathbf{v}) \equiv \mathbf{v} - \operatorname{proj}_{\mathbf{u}}(\mathbf{v})$ is also a linear transformation. Here

$$\operatorname{proj}_{\mathbf{u}}(\mathbf{v}) \equiv \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{|\mathbf{u}|^2} \mathbf{u}$$

   Now show directly from the axioms of the inner product that

$$\langle T_{\mathbf{u}} \mathbf{v}, \mathbf{u} \rangle = 0$$

8. Let V be a finite dimensional inner product space, the field of scalars equal to either $\mathbb{R}$ or $\mathbb{C}$. Verify that f given by $f\mathbf{v} \equiv \langle \mathbf{v}, \mathbf{z} \rangle$ is in $\mathscr{L}(V, \mathbb{F})$. Next suppose f is an arbitrary element of $\mathscr{L}(V, \mathbb{F})$. Show the following.

   (a) If $f = 0$, the zero mapping, then $f\mathbf{v} = \langle \mathbf{v}, \mathbf{0} \rangle$ for all $\mathbf{v} \in V$.

   (b) If $f \neq 0$ then there exists $\mathbf{z} \neq \mathbf{0}$ satisfying $\langle \mathbf{u}, \mathbf{z} \rangle = 0$ for all $\mathbf{u} \in \ker(f)$.

   (c) Explain why $f(\mathbf{y})\mathbf{z} - f(\mathbf{z})\mathbf{y} \in \ker(f)$.

(d) Use part b. to show that there exists $\mathbf{w}$ such that $f(\mathbf{y}) = \langle \mathbf{y}, \mathbf{w} \rangle$ for all $\mathbf{y} \in V$.

(e) Show there is at most one such $\mathbf{w}$.

You have now proved the Riesz representation theorem which states that every $f \in \mathscr{L}(V, \mathbb{F})$ is of the form

$$f(\mathbf{y}) = \langle \mathbf{y}, \mathbf{w} \rangle$$

for a unique $\mathbf{w} \in V$.

9. ↑Let $A \in \mathscr{L}(V, W)$ where V, W are two finite dimensional inner product spaces, both having field of scalars equal to $\mathbb{F}$ which is either $\mathbb{R}$ or $\mathbb{C}$. Let $f \in \mathscr{L}(V, \mathbb{F})$ be given by

$$f(\mathbf{y}) \equiv \langle A\mathbf{y}, \mathbf{z} \rangle$$

where $\langle \rangle$ now refers to the inner product in W. Use the above problem to verify that there exists a unique $\mathbf{w} \in V$ such that $f(\mathbf{y}) = \langle \mathbf{y}, \mathbf{w} \rangle$, the inner product here being the one on V. Let $A^*\mathbf{z} \equiv \mathbf{w}$. Show that $A^* \in \mathscr{L}(W, V)$ and by construction,

$$\langle A\mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{y}, A^*\mathbf{z} \rangle .$$

In the case that $V = \mathbb{F}^n$ and $W = \mathbb{F}^m$ and A consists of multiplication on the left by an $m \times n$ matrix, give a description of A^*.

10. Let A be the linear transformation defined on the vector space of smooth functions (Those which have all derivatives) given by $Af = D^2 + 2D + 1$. Find $\ker(A)$. **Hint:** First solve $(D+1)z = 0$. Then solve $(D+1)y = z$.

11. Let A be the linear transformation defined on the vector space of smooth functions (Those which have all derivatives) given by $Af = D^2 + 5D + 4$. Find $\ker(A)$. Note that you could first find $\ker(D+4)$ where D is the differentiation operator and then consider $\ker(D+1)(D+4) = \ker(A)$ and consider Sylvester's theorem.

12. Suppose $A\mathbf{x} = \mathbf{b}$ has a solution where A is a linear transformation. Explain why the solution is unique precisely when $A\mathbf{x} = \mathbf{0}$ has only the trivial (zero) solution.

13. Verify the linear transformation determined by the matrix

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

maps $\mathbb{R}^3$ onto $\mathbb{R}^2$ but the linear transformation determined by this matrix is not one to one.

14. Let L be the linear transformation taking polynomials of degree at most three to polynomials of degree at most three given by

$$D^2 + 2D + 1$$

where D is the differentiation operator. Find the matrix of this linear transformation relative to the basis $\left\{1, x, x^2, x^3\right\}$. Find the matrix directly and then find the matrix with respect to the differential operator $D+1$ and multiply this matrix by itself. You should get the same thing. Why?

15. Let L be the linear transformation taking polynomials of degree at most three to polynomials of degree at most three given by D^2+5D+4 where D is the differentiation operator. Find the matrix of this linear transformation relative to the bases $\left\{1,x,x^2,x^3\right\}$. Find the matrix directly and then find the matrices with respect to the differential operators $D+1, D+4$ and multiply these two matrices. You should get the same thing. Why?

16. Show that if $L \in \mathscr{L}(V,W)$ (linear transformation) where V and W are vector spaces, then if $L\mathbf{y}_p = \mathbf{f}$ for some $\mathbf{y}_p \in V$, then the general solution of $L\mathbf{y} = \mathbf{f}$ is of the form $\ker(L) + \mathbf{y}_p$.

17. Let $L \in \mathscr{L}(V,W)$ where V,W are vector spaces, finite or infinite dimensional, and define $\mathbf{x} \sim \mathbf{y}$ if $\mathbf{x}-\mathbf{y} \in \ker(L)$. Show that $\sim$ is an equivalence relation. ($\mathbf{x} \sim \mathbf{x}$, if $\mathbf{x} \sim \mathbf{y}$, then $\mathbf{y} \sim \mathbf{x}$, and $\mathbf{x} \sim \mathbf{y}$ and $\mathbf{y} \sim \mathbf{z}$ implies $\mathbf{x} \sim \mathbf{z}$.) Next define addition and scalar multiplication on the space of equivalence classes as follows. $[\mathbf{x}] \equiv \{\mathbf{y} : \mathbf{y} \sim \mathbf{x}\}$.

$$\begin{aligned} [\mathbf{x}] + [\mathbf{y}] &\equiv [\mathbf{x}+\mathbf{y}] \\ \alpha[\mathbf{x}] &= [\alpha\mathbf{x}] \end{aligned}$$

Show that these are well defined definitions and that the set of equivalence classes is a vector space with respect to these operations. The zero is $[\ker L]$. Denote the resulting vector space by $V/\ker(L)$. Now suppose L is onto W. Define a mapping $A : V/\ker(K) \mapsto W$ as follows.

$$A[\mathbf{x}] \equiv L\mathbf{x}$$

Show that A is well defined, one to one and onto.

18. If V is a finite dimensional vector space and $L \in \mathscr{L}(V,V)$, show that the minimal polynomial for L equals the minimal polynomial of A where A is the $n \times n$ matrix of L with respect to some basis.

19. Let A be an $n \times n$ matrix. Describe a fairly simple method based on row operations for computing the minimal polynomial of A. Recall, that this is a monic polynomial $p(\lambda)$ such that $p(A) = 0$ and it has smallest degree of all such monic polynomials. **Hint:** Consider $I, A^2, \cdots$. Regard each as a vector in $\mathbb{F}^{n^2}$ and consider taking the row reduced echelon form or something like this. You might also use the Cayley Hamilton theorem to note that you can stop the above sequence at A^n.

20. Let A be an $n \times n$ matrix which is non defective. That is, there exists a basis of eigenvectors. Show that if $p(\lambda)$ is the minimal polynomial, then $p(\lambda)$ has no repeated roots. **Hint:** First show that the minimal polynomial of A is the same as the minimal polynomial of the diagonal matrix

$$D = \begin{pmatrix} D(\lambda_1) & & \\ & \ddots & \\ & & D(\lambda_r) \end{pmatrix}$$

Where $D(\lambda)$ is a diagonal matrix having λ down the main diagonal and in the above, the λ_i are distinct. Show that the minimal polynomial is $\prod_{i=1}^{r}(\lambda - \lambda_i)$.

21. Show that if A is an $n\times n$ matrix and the minimal polynomial has no repeated roots, then A is non defective and there exists a basis of eigenvectors. Thus, from the above problem, a matrix may be diagonalized if and only if its minimal polynomial has no repeated roots. (It turns out this condition is something which is relatively easy to determine. You look at the polynomial and its derivative and ask whether these are relatively prime. The answer to this question can be determined using routine algorithms as discussed above in the section on polynomials and fields. Thus it is possible to determine whether an $n\times n$ matrix is defective.) **Hint:** You might want to use Theorem 18.3.1.

22. Recall the linearization of the Lotka Volterra equations used to model the interaction between predators and prey. It was shown earlier that if x,y are the deviations from the equilibrium point, then

$$\begin{aligned} x' &= -bxy-b\frac{c}{d}y \\ y' &= dxy+\frac{a}{b}dx \end{aligned}$$

If one is interested only in small values of x,y, that is, in behavior near the equilibrium point, these are approximately equal to the system

$$\begin{aligned} x' &= -b\frac{c}{d}y \\ y' &= \frac{a}{b}dx \end{aligned}$$

Written more simply, for $\alpha,\beta>0$,

$$\begin{aligned} x' &= -\alpha y \\ y' &= \beta x \end{aligned}$$

Find the solution to the initial value problem

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 0 & -\alpha \\ \beta & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix},\ \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

Also have a computer graph the vector fields $(-xy-\ y,xy+x)$ and $(-y,x)$ which result from setting all the parameters equal to 1.

23. This and the next problems will demonstrate how to solve any second order linear differential equation with constant coefficients. Suppose you want to find all smooth solutions y to

$$y''+(a+b)y'+aby=0$$

where a,b are complex numbers, $a\neq b$. Then in terms of the differentiation operator D, this can be written as

$$(D+a)(D+b)y=0$$

Thus you want to have $\ker((D+a)(D+b))$. Thus the two operators $(D+a)$ and $(D+b)$ commute. Show this carefully. It just comes down to standard calculus manipulations. Also show that $\ker(D+a)$ is of the form Ce^{-at} where C is a constant.

Use this to verify that $(D+a)$ is one to one on $\ker(D+b)$ and $(D+b)$ is one to one on $\ker(D+a)$. Then use Lemma 18.3.4 to verify that the solution to the above equation is of the form

$$\ker((D+a)(D+b)) = C_1 e^{-at} + C_2 e^{-bt}$$

Here we write e^{-at} to signify the function $t \to e^{-at}$. This shows how to solve any equation $y'' + \alpha y' + \beta y = 0$ where α, β are real or complex numbers. This is because you can always write such an equation in the form discussed above by simply factoring the quadratic polynomial $r^2 + \alpha r + \beta$ using the quadratic formula. Of course you might need to use DeMoivre's theorem to take square roots. This is called the general solution to the homogeneous equation. Once you have found the general solution, you can then determine the constants C_1, C_2 in such a way that the solution to a given initial condition where $y(0), y'(0)$ are given may be obtained. For example, if I wanted $y(0) = 1, y'(0) = 0$, this would be easy to find. I would just need to solve the equations

$$\begin{aligned} y(0) &= C_1 + C_2 = 1 \\ y'(0) &= (-a)C_1 + (-b)C_2 = 0 \end{aligned}$$

then solve these equations as in the first few chapters of the book.

24. What if the second order equation is of the form $(D+a)^2 y = 0$? Show that

$$\ker\left((D+a)^2\right) = C_1 e^{-at} + C_2 t e^{-at}$$

To show this, verify directly that the right side is in $\ker\left((D+a)^2\right)$. Next suppose $y \in \ker\left((D+a)^2\right)$. Thus

$$(D+a)((D+a)y) = 0$$

and so

$$(D+a)y = Ce^{-at}$$

because you know $\ker(D+a)$. Now simply solve the above equation as follows. You have

$$y' + ay = Ce^{-at}$$

Multiply both sides by e^{at} and verify that this yields

$$\frac{d}{dt}\left(e^{at}y\right) = C$$

Now finish the details to verify that everything in $\ker(D+a)^2$ is of the desired form.

25. The case of real coefficients is very important and in fact is the case of most interest. Show that if α, β are real, then if y is a solution to $y'' + \alpha y' + \beta y = 0$, then so is $\bar{y}$, the complex conjugate.

26. If you have the solution to $\ker\left(\left(D^2+(a+b)D+ab\right)\right)=C_1e^{at}+C_2e^{bt}$ and a,b are complex but $a+b,ab$ are real, show that in fact, they are complex conjugates so $a=\alpha+i\beta, b=\alpha-i\beta$. Then explain why you get exactly the same solution for $\ker\left(\left(D^2+(a+b)D+ab\right)\right)$ by simply writing in the form
$$e^{-\alpha t}\left(B_1\left(\cos\left(\beta t\right)\right)+B_2\sin\left(\beta t\right)\right)$$
The advantage in doing this is that everything is expressed in terms of real functions and typically you are looking for real solutions.

27. Find the solutions to the following initial value problems. Write in terms of real valued functions.

    (a) $y''+4y'+3y=0, y(0)=1, y'(0)=1$

    (b) $y''+2y'+2y=0, y(0)=0, y'(0)=1$

    (c) $y''+4y'+4y=0, y(0)=1, y'(0)=1$

    (d) $y''+5y'+4y=0, y(0)=0, y'(0)=1$

    (e) $y''+4y=0, y(0)=1, y'(0)=0$

    (f) $y''-2y'+5y=0, y(0)=2, y'(0)=1$

28. If you have an initial value problem of the form
$$y''+b(t)y'+c(t)y=0,\ y(t_0)=y_0, y'(t_0)=y_1$$
Explain why it can be written in the form
$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}'=\begin{pmatrix} 0 & 1 \\ -c(t) & -b(t) \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} x_1(t_0) \\ x_2(t_0) \end{pmatrix}=\begin{pmatrix} y_0 \\ y_1 \end{pmatrix}$$
Easy if you take $x_1=y$ and $x_1'=x_2$. Then you just plug in to the equation and get what is desired.

29. Suppose you have two solutions $y,\hat{y}$ to the "homogeneous" equation
$$y''+b(t)y'+c(t)y=0 \tag{18.22}$$
The Wronskian of these is defined as
$$W(y,\hat{y})(t)\equiv W(t)\equiv\det\begin{pmatrix} y(t) & \hat{y}(t) \\ y'(t) & \hat{y}'(t) \end{pmatrix}$$
Show that
$$W'(t)+b(t)W(t)=0$$
Now let $B'(t)=b(t)$. Multiply by $e^{B(t)}$ and verify that
$$\left(e^{B(t)}W(t)\right)'=0$$
Hence
$$W(t)=Ce^{-B(t)}$$
This is Abel's formula. Note that this shows that the Wronskian either vanishes for all t or for no t.

30. Suppose you have found two solutions $y,\hat{y}$ to 18.22. Consider their Wronskian. Show that the Wronskian is nonzero if and only if the ratio of these solutions is not a constant. **Hint:** This is real easy. Just use the quotient rule and observe that in the quotient rule the numerator is a multiple of the Wronskian.

31. $\uparrow$Show that if you have two solutions $y,\hat{y}$ of 18.22, whose Wronskian is nonzero, then for any choice of y_0, y_1, there exist constants C_1, C_2 such that there is exactly one solution to the initial value problem

$$y'' + b(t)y' + c(t)y = 0,\ y(t_0) = y_0, y'(t_0) = y_1$$

which is of the form $C_1 y + C_2 \hat{y}$. When this happens, we say that we have the general solution in the form $C_1 y + C_2 \hat{y}$. Explain why all solutions to the equation must be of this form.

32. $\uparrow$Suppose you have found a single solution to the equation

$$y'' + b(t)y' + c(t)y = 0$$

How can you go about finding the general solution to

$$y'' + b(t)y' + c(t)y = 0?$$

Hint: You know that all you have to do is to find another solution $\hat{y}$ such that it is not a scalar multiple of y. Also, you know Abel's formula. Letting $B'(t) = b(t)$,

$$\begin{vmatrix} \hat{y}(t) & y(t) \\ \hat{y}'(t) & y'(t) \end{vmatrix} = Ce^{-B(t)}$$

and so you have a differential equation for $\hat{y}$

$$-y(t)\hat{y}'(t) + \hat{y}(t)y'(t) = e^{-B(t)}$$

You don't care to find all examples. All you need is one which is not a multiple of y. That is why it suffices to let $C = 0$. Then

$$\hat{y}'(t) - \frac{y'(t)}{y(t)}\hat{y}(t) = -\frac{e^{-B(t)}}{y(t)}$$

Now solve this equation for $\hat{y}$. First multiply by $1/y$. Verify that

$$\frac{d}{dt}(\hat{y}/y) = -\frac{e^{-B(t)}}{y^2(t)}$$

Now from this describe how to find $\hat{y}$ which is not a constant multiple of y.

33. If you have the general solution for $y'' + b(t)y' + c(t)y = 0$, two solutions $y,\hat{y}$ having nonzero Wronskian, show that it is always possible to obtain all solutions to

$$y'' + b(t)y' + c(t)y = f(t)$$

and that such a solution will be of the form $A(t)\,y(t) + B(t)\,\hat{y}(t)$. This is called variation of parameters. **Hint:** You want it to work. Plug in and require it to do so.

$$\begin{aligned}
&c(t)\,(y = Ay + B\hat{y})\,,\ b(t)\left(y' = \overbrace{A'y + B'\hat{y}}^{\equiv 0} + Ay' + B\hat{y}'\right)\\
&1\left(y'' = A'y' + B'\hat{y}' + Ay'' + B\hat{y}''\right)
\end{aligned}$$

Then show we get a solution if $A'y + B'\hat{y} = 0,\ A'y' + B'\hat{y}' = f$. Now get a solution to this using that the Wronskian is not zero. Then if z is any solution and $\tilde{y}$ is the one you just found, $z - \tilde{y}$ solves $y'' + b(t)\,y' + c(t)\,y = 0$ and so is of the form $C_1 y + C_2 \hat{y}$.

34. Explain how to write the higher order differential equation

$$y^{(k)} + a_{k-1}y^{(k-1)} + \cdots + a_1 y' + a_0 y = f(t)$$

as a first order system of the form

$$\mathbf{x}' = A\mathbf{x} + \mathbf{F}$$

Thus the theory of this chapter includes **all** linear ordinary differential equations with constant coefficients.

Appendix A

The Jordan Canonical Form*

Recall Corollary 18.4.4. For convenience, this corollary is stated below.

Corollary A.0.1 *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

$$m_k = \dim \ker (A - \lambda_k I)^{r_k} .$$

where the minimal polynomial of A is

$$\prod_{k=1}^{p} (\lambda - \lambda_k)^{r_k}$$

The Jordan Canonical form involves a further reduction in which the upper triangular matrices, T_k assume a particularly revealing and simple form.

Definition A.0.2 *$J_k(\alpha)$ 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 18.4.4 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. It is assumed here that the field of scalars is $\mathbb{C}$ but everything which will be done below works just fine if the minimal polynomial can be completely factored into linear factors in the field of scalars.

Note that in the conclusion of Corollary 18.4.4 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 A.0.3 *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 if $N\mathbf{v}_{k+1} = \mathbf{v}_k$ for all $k = 1, 2, \cdots, r$ and $\mathbf{v}_1$ is an eigenvector so $N\mathbf{v}_1 = 0$. 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 $\mathbf{v}_1$,

$$N : \text{span}(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r) \mapsto \text{span}(\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r). \tag{1.1}$$

Also if $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r\}$ is a chain, then $r \leq n$.

Proof: First note that 1.1 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. Suppose then $\sum_{k=1}^{r} c_k\mathbf{v}_k = 0$. Do N^{r-1} to it to conclude $c_r = 0$. Next do N^{r-2} to it to conclude $c_{r-1} = 0$ and continue this way. Now it is obvious $r \leq n$ because the chain is independent. 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: The claim is true if $k=1$. This follows from Claim 1. Suppose it is true for $k-1,k\geq 2$. Then $\{B_1,\cdots,B_{k-1}\}$ is linearly independent. Suppose $\sum_{q=1}^p c_q\mathbf{w}_q=\mathbf{0},c_q\neq 0$ where the $\mathbf{w}_q$ come from $\{B_1,\cdots,B_{k-1},B_k\}$. By induction, some of these $\mathbf{w}_q$ must come from B_k. Let $\mathbf{v}_i^k$ be the one for which i is as large as possible. Then do N^{i-1} to both sides to obtain $\mathbf{v}_1^k$, the eigenvector upon which the chain B_k is based, is a linear combination of $\{B_1,\cdots,B_{k-1}\}$ contrary to the construction. Since $\{B_1,\cdots,B_k\}$ is linearly independent, the process terminates. This proves the claim.

Claim 3: Suppose $N\mathbf{w}=\mathbf{0}$. ($\mathbf{w}$ is an eigenvector) Then there exist 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, $\mathbf{w}\in\operatorname{span}(B_1,\cdots,B_s)$ since otherwise, it could serve as a base for another chain. 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. $\mathbf{0}=\sum_{i=1}^s\sum_{k=2}^{r_i}c_i^k\mathbf{v}_{k-1}^i$ and so by **Claim 2,** $c_i^k=0$ if $k\geq 2$. It follows that $\mathbf{w}=\sum_{i=1}^s c_i^1\mathbf{v}_1^i$ and this proves the claim.

If $N\mathbf{w}=\mathbf{0}$, then $\mathbf{w}\in\operatorname{span}(B_1,\cdots,B_s)$. In fact, it was a particular linear combination involving the bases of the chains. What if $N^k\mathbf{w}=\mathbf{0}$? Does it still follow that $\mathbf{w}\in\operatorname{span}(B_1,\cdots,B_s)$?

Claim 4: If $N^k\mathbf{w}=\mathbf{0}$, then $\mathbf{w}\in\operatorname{span}(B_1,\cdots,B_s)$.

Proof of Claim 4: Suppose this true that if $N^{k-1}\mathbf{w},k\geq 2$, it follows that

$$\mathbf{w}\in\operatorname{span}(B_1,\cdots,B_s).$$

Then suppose $N^k\mathbf{w}=\mathbf{0}$. If $N^{k-1}\mathbf{w}=\mathbf{0}$, then by induction, $\mathbf{w}\in\operatorname{span}(B_1,\cdots,B_s)$. The other case is that $N^{k-1}\mathbf{w}\neq\mathbf{0}$. Then you have the chain $N^{k-1}\mathbf{w},\cdots,\mathbf{w}$ where $N^{k-1}\mathbf{w}$ is an eigenvector. The chain has length no longer than the lengths of any of the B_i by construction. Then by **Claim 3,**

$$N^{k-1}\mathbf{w}=\sum_{i=1}^s c_i\mathbf{v}_1^i=\sum_{i=1}^s c_iN^{k-1}\mathbf{v}_k^i$$

And so, $N^{k-1}\left(\mathbf{w}-\sum_{i=1}^s c_i\mathbf{v}_k^i\right)=0$ which implies by induction that

$$\mathbf{w}-\sum_{i=1}^s c_i\mathbf{v}_k^i\in\operatorname{span}(B_1,\cdots,B_s).$$

This proves **Claim 4.**

Since every $\mathbf{w}$ satisfies $N^n\mathbf{w}=\mathbf{0}$ this shows that $\operatorname{span}(B_1,\cdots,B_s)=\mathbb{F}^n$. By **Claim 2,** $\{B_1,\cdots,B_s\}$ is independent. Therefore, this 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

$$S^{-1} = \begin{pmatrix} C_1 \\ \vdots \\ C_s \end{pmatrix}$$

From the construction, $N\mathbf{v}_j^k \in \operatorname{span}(B_k)$, and so $C_i N\mathbf{v}_j^k = \mathbf{0}$. It follows that $C_i B_i = I_{r_i \times r_i}$ and $C_i N B_j = 0$ if $i \neq j$. Let

$$C_k = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_{r_k}^T \end{pmatrix}.$$

Then noting that B_k is $n \times r_k$ and C_k is $r_k \times n$,

$$C_k N B_k = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_{r_k}^T \end{pmatrix} \begin{pmatrix} N\mathbf{v}_1^k & \cdots & N\mathbf{v}_{r_k}^k \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_{r_k}^T \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{v}_1^k & \cdots & \mathbf{v}_{r_k-1}^k \end{pmatrix}$$

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

$$S^{-1} N S = \begin{pmatrix} C_1 \\ \vdots \\ C_s \end{pmatrix} N \begin{pmatrix} B_1 & \cdots & B_s \end{pmatrix} = \begin{pmatrix} J_{r_1}(0) & & & 0 \\ & J_{r_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{r_s}(0) \end{pmatrix}$$

as claimed. ■

Now let the upper triangular matrices, T_k be given in the conclusion of Corollary 18.4.4. 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 A.0.3. Therefore, there exists S_k such that $S_k^{-1} N_k S_k$ is of the form given in Lemma A.0.3. 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 A.0.4 *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 gives the existence of the Jordan canonical form.

Theorem A.0.5 *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*

$$S^{-1}AS = \begin{pmatrix} J(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & J(\lambda_r) \end{pmatrix} \tag{1.2}$$

where $J(\lambda_k)$ is an $m_k \times m_k$ matrix of the form

$$\begin{pmatrix} J_{k_1}(\lambda_k) & & & 0 \\ & J_{k_2}(\lambda_k) & & \\ & & \ddots & \\ 0 & & & J_{k_r}(\lambda_k) \end{pmatrix} \tag{1.3}$$

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 18.4.4, 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 A.0.4 There exist matrices, S_k such that $S_k^{-1}T_kS_k = J(\lambda_k)$ where $J(\lambda_k)$ is described in 1.3. 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. ■

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 A.0.6 *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 $\{\text{rank}(A - \lambda I)^m\}$ and $\{\text{rank}(B - \lambda I)^m\}$. Since A and B are similar, these two sequences coincide. (Why?) Also, for the same reason, $\{\text{rank}(J_A - \lambda I)^m\}$ coincides with $\{\text{rank}(J_B - \lambda I)^m\}$. Now pick λ_k an eigenvalue and consider $\{\text{rank}(J_A - \lambda_k I)^m\}$ and $\{\text{rank}(J_B - \lambda_k I)^m\}$. 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,

$$\begin{aligned} \text{rank}(J_A - \lambda_k I)^m &= \sum_{i \neq k} m_i + \text{rank}(J_A(0)^m) \\ &= \sum_{i \neq k} m_i + \text{rank}(J_B(0)^m) \\ &= \text{rank}(J_B - \lambda_k I)^m, \end{aligned}$$

which shows $\text{rank}(J_A(0)^m) = \text{rank}(J_B(0)^m)$ 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 rank $(J_B(0)^m) = \sum_{i=1}^{p} \operatorname{rank}\left(J_{l_i}(0)^m\right)$, similar for rank $(J_A(0)^m)$. In going from m to $m+1$,

$$\operatorname{rank}\left(J_{l_i}(0)^m\right) - 1 = \operatorname{rank}\left(J_{l_i}(0)^{m+1}\right)$$

until $m = l_i$ 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 $m = k_{r-b}$,

$$\sum_{i=1}^{r} \operatorname{rank}\left(J_{l_i}(0)^m\right) = \sum_{i=1}^{r} \operatorname{rank}\left(J_{k_i}(0)^m\right)$$

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. ■

Appendix B

Directions For Computer Algebra Systems

B.1 Finding Inverses

►►

B.2 Finding Row Reduced Echelon Form

►►

B.3 Finding *PLU* Factorizations

► ►

B.4 Finding *QR* Factorizations

► ►

B.5 Finding Singular Value Decomposition

First here is how you do it using Scientific Notebook. It does it very easily.

►

Next, here is the rather elaborate procedure to get the singular value decomposition using maple. I am using maple 12. Maybe there is an easier way with a more up to date version. I am not sure.

►

B.6 Use Of Matrix Calculator On Web

There is a really nice service on the web which will do all of these things very easily. It is www.bluebit.gr/matrix-calculator/ or click on following link or google matrix calculator.

▶

You enter a matrix row by row, placing a space between each number. When you come to the end of a row, you press enter on the keyboard to enter the next row. When you have entered the matrix, you select what you want it to do.

▶

Bibliography

[1] **Apostol T.** *Calculus Volume II Second edition,* Wiley *1969.*

[2] **Baker, Roger**, *Linear Algebra*, Rinton Press 2001.

[3] **Davis H. and Snider A.,** *Vector Analysis* Wm. C. Brown 1995.

[4] **Edwards C.H.** *Advanced Calculus of several Variables,* Dover 1994.

[5] **Chahal J.S.,** *Historical Perspective of Mathematics 2000 B.C. - 2000 A.D. Kendrick Press, Inc. (2007)*

[6] **Golub, G. and Van Loan, C.,***Matrix Computations*, Johns Hopkins University Press, 1996.

[7] **Greenberg M.D.** *Advanced Engineering Mathematics* Prentice Hall 1998 Second edition.

[8] **Gurtin M.** *An introduction to continuum mechanics,* Academic press 1981.

[9] **Hardy G.** *A Course Of Pure Mathematics, Tenth edition,* Cambridge University Press 1992.

[10] **Horn R. and Johnson C.** *matrix Analysis,* Cambridge University Press, 1985.

[11] **Jacobsen N.** *Basic Algebra* Freeman 1974.

[12] **Karlin S. and Taylor H.** *A First Course in Stochastic Processes,* Academic Press, 1975.

[13] **Kuttler, K.** Linear Algebra and Analysis

[14] **Nobel B. and Daniel J.** *Applied Linear Algebra,* Prentice Hall, 1977.

[15] **Rudin W.** *Principles of Mathematical Analysis*, McGraw Hill, 1976.

[16] **Salas S. and Hille E.,** *Calculus One and Several Variables,* Wiley 1990.

[17] **Strang Gilbert**, *Linear Algebra and its Applications,* Harcourt Brace Jovanovich 1980.

[18] **Wilkinson, J.H.,** *The Algebraic Eigenvalue Problem, Clarendon Press Oxford 1965.*

[19] **Yosida K.,** *Functional Analysis, Springer Verlag, 1978.*

Index